UNIVERSITY OF CALIFORNIA

Los Angeles

Speech as Writing: Literary Dialect Orthography in the United States 1790-1930

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in English

by

Craig Messner

2021

ABSTRACT OF THE DISSERTATION

Speech as Writing:

Literary Dialect Orthography in the United States 1790-1930

by

Craig Messner

Doctor of English

University of California, Los Angeles, 2021

Professor Christopher J. Looby, Chair

The study and characterization of the literary uses of non-standard American English writing systems was once a topic of research central to the study of American literature. *Speech as Writing: Literary Dialect Orthography in the United States 1790-1930* argues that the emergence of new computational tools and theoretical insights enables a return to the general study of at least one common component of literary dialect – non-standard orthography. The use of non-standard orthographic systems in crafting dialect literature differs from the use of non-standard syntax or vocabulary in that it presents a full system of meaning independent from the encoding facet of orthography typically explored by linguistics or cognitive science. *Speech as Writing* employs these insights alongside a computational methodology drawn from corpus linguistics and information theory to explore how this novel understanding of

orthography can contribute to novel understandings of nineteenth and early twentieth-century United States literature.

The dissertation of Craig Messner is approved.

Michael A. North

Michael C. Cohen

Brian Kim Stefans

Christopher J. Looby, Chair

University of California, Los Angeles

2021

Table of Contents

List of Figures

List of Tables

Acknowledgments

I recognize the following institutions for financial support that aided in the completion of this dissertation: the English Department at UCLA and the UCLA Graduate Division.

I am honored and privileged to recognize my dissertation committee. Chris Looby is the very model of a scholar and mentor, and I am a better person for having known him. Brian Kim Stefans opened my eyes to a whole world of materials and perspectives central to this project I would have never discovered otherwise. Michael Cohen for offering practical and down to earth advice in order to help reign in a very broad project and Michael North for offering generous feedback and an excellent model of what it is to be a scholar with broad interests.

A hearty thanks to all who made my intellectual life at UCLA so much richer. The participants in the Americanist Research Colloquium, M/ELT and the 20/21 Working Group. The folks associated with UCLA Humanities Technology and the Digital Humanities program; Miriam Posner, Dave Shepherd, Ashley Sanders, John Lynch and Tom Garbelotti among so many others. Thank you also to Mark Seltzer and Allison Carruth for key intellectual interventions during my early graduate years.

Thank you to my graduate student peers, those compatriots of the TA offices. Kim Calder, Jen MacGregor, Crescent Rainwater, Alexandra Verini, Ellen Truxaw, Rebecca Hill, Kathryn Cai, Vanessa Febo, Grant Rosson, Sujin Youn, Jessica Cook, Lauren Dembowitz, Abraham Encinas, Stacey Shin, Mike Vignola, Cailey Hall, Lindsay Wilhelm, Sam Sommers. Greg Toy, Ben Beck, Tim Fosbury and Kiel Shaub for being especially patient when I had something interesting only to me to talk about. Jay Jin, Efren Lopez, Kirsten Lew, Will Clark, Jordan Wingate, Angelina Del Balzo, Jonathan Kincade, Gabe Mehlman and Martin Zirulnik for the above, and for helping me get out of the house now and then.

The UCLA Neuroscience cohort I happened to befriend. Shivan Bonanno, Nick Hardy and Ryan Guglietta were my first close friends at UCLA and provided support, stability and commiseration. Everyone else (Don Julian, Esther Nie, Eric Geschwing, Kathy Myers among others) for the years of conversation and companionship they provided.

Justine Murison guided me during my early intellectual development and gave me an excellent direction in which to set my course.

Thank you to my brother and intellectual companion, Mark Messner, as well as his beautiful family, Laura, Zillah and Wyatt. Despite our divergent disciplinary paths we have always been able to meet in the odd neutral ground of linear algebra and science fiction. Thanks also to our parents, Scott and Stacy, who instilled a voracious intellectual curiosity in us both.

Finally, the utmost thanks to my favorite UCLA neuroscientist, Catherine Schweppe, and our beloved cat Gnocchi for their unceasing support and care.

Vita

EDUCATION

2018	C. Phil, English, University of California, Los Angeles
2015	M.A., English, University of California, Los Angeles
2012	B.A. highest honors, English, University of Illinois Urbana-Champaign

PUBLICATIONS AND PRESENTATIONS

Forthcoming	"Syntax" in *Text Analytics for Literature: Tools & Methods from the Digital Humanities,* Bloomsbury Academic
2019	Panel moderator, UCLA EPIC Humanities Now: Transformative Teaching conference, Innovative Teaching with Technology Panel
2014	"Pym's Games" in *Poe Studies: History, Theory, Interpretation.* Volume 47
2014	"Poe, Gaming and the Remediation of Horror Tropes", UC Davis Contours of Algorithmic Life conference
2014	"Reading Algorithmic Translations" panel, UCLA Southland conference

SELECT AWARDS, GRANTS, AND FELLOWSHIPS

2019	UCLA Department of English Dissertation Fellowship
2013 and 2014	UCLA Graduate Summer Research Mentorship fellowship
2011	Robert D. Novak Scholarship

General Introduction: Orthography as Substructural Style

"What can be calculated by means of computerized mathematics is another subject, and a strategic one" — Kittler

The popularity of internet-based communication forms has helped to unsettle one common notion of the self. The humanist picture of the person as a rational individual actor has long been critiqued in the academy, but only the proliferation of digitally-mediated phenomena like "junk news," QAnon and "alternative facts" has reopened the question of the self in the broader public discourse. At its core this discussion centers on the formation of belief. Commentators both professional and avocational who choose to plunge into the depths of conspiratorial thought return to the same central question—what about these particular beliefs, as absurd as they may seem, make them possible to be believed in? Answers to this question vary, but all, even just implicitly, reconsider the relationship between the phenomenological experience of self and the environment it navigates, calling into question the self's ability to police the transit of notions from being exterior sources to becoming interior beliefs.

One theory that recurs in these discussions could be termed "linguistic contagion theory." For these commentators, exposure to exterior notions and their communities, rather than active selection, leads to a modification in beliefs. Something about the way these notions are expressed allows them to bypass humanistic checks on irrationality,

leading to neuronal, and thus behavioral, changes in those continually exposed. [1] Again, this theory echoes concerns already voiced in academia. Derrida's influence especially inspired many scholars to produce excellent scholarship that details the anxiety over such breakages and crossings in a variety of historical and national contexts in a manner more nuanced than current popular attempts.[2] Like these earlier academic efforts, much of the popular discourse focuses on the rhetoric used to embed these notions in text. However, they also place great emphasis on the low-level linguistically substructural elements of these utterances. In large part the focus falls on how missives from Q-believers or headlines from irreputable news sources actually hang together as linguistic units — their choices of diction, their grammatical structure, their punctuation.[3] A few major factors likely contribute to the switch to this specific focus.

---

[1] For example, a *Vox* article analyzing the demography and behavior of QAnon believers on the /r/greatawakening subreddit comes to the conclusion that believers in Qanon are "largely casual" and hold other interests typical of their demography, and typical of the white male segment of the US population at large. They consume Q content created by a smaller group of hardcore users, but do so, seemingly, uncritically. While these users may be primed to accept such content, it is the content itself that then is the causal shift in their belief.
https://www.vox.com/2018/8/8/17657800/qanon-reddit-conspiracy-data

[2] See for example Looby, *Voicing America: Language, Literary Form, and the Origins of the United States*.

[3] Analyzing the "drops" of "information" left by Q or the following posts of supporters using natural language toolkits has become a cottage industry. Understanding the mindset and efficacy of these postings

For one, such texts are more easily available and digestible now than at any previous point in history. Social media platforms make voluminous corpora of such utterances available at the click of a mouse, and ubiquitous computing makes them digestible, if only in the terms of syntax, diction and sentiment most amenable to current natural language processing technologies. The tool then determines the focus — since computer processing is much better at capturing substructural regularities than semantic sense this becomes the main thrust. Additionally, substructural forms become a way to distinguish "fake" utterances from "real" ones. Identifying concrete substructural differences between a poorly sourced Facebook group news post headline and the headline of a major newspaper insulates the latter from the former. Both may make the same claim to truth on a rhetorical level, but their substructural differences allow them to be distinguished. Finally, and most exotically, there lurks the suspicion that these substructural elements are being weaponized. Elements of syntax or orthography are suspect simply because they are processed unconsciously. If computers can reveal patterns in these elements that the phenomenological self cannot, and patterns carry meaning, who knows what kind of suggestions could be smuggled in through an errant apostrophe or inverted clause?

All of these aspects likely contribute to the current popular focus on what I have decided to term “substructural style”, but it is the last most outlandish claim that holds

---

is perceived as having national security implications to the degree that even NED-connected news and intelligence site *Bellingcat* has taken up the topic: https://www.bellingcat.com/news/americas/2021/01/29/the-qanon-timeline/.

the most interest. One need not subscribe to some substructural form of the discredited pseudo-science of neuro-linguistic programming and accept that repeated substructural patterns can cause unmonitored behavioral change to agree that they can still hold some form of stylistic meaning ingested more or less nonconsciously during reading. The emerging academic study of "junk news" makes use of this point. Although "junk news" often refers to content that straightforwardly misreports facts or invents fictional occurrences, scholars of the phenomenon also stress the unique stylistic profile junk news or "clickbait" articles often employ. Benjamin Horne and Sibel Adali use computational methods to unveil some of these features in their paper, "This Just In." Horne and Adali note that titles heavily contribute to the classification of junk news. Junk news "[attempts] to squeeze as much substance into titles as possible" by omitting "stop-words [common connector words] and nouns" in order to pack in as many "proper nouns and verb phrases" as possible.[4] This, along with junk news's use of stylistic elements consistent with satire, lead Horne and Adali to conclude that junk news attempts to "convince through heuristics" in contrast to "real" news, which convinces "through argument."[5] Horne and Adali ground this argument in the language of cognitive science, drawing on the concept of processing time to make their distinction between heuristic and argument. Even without the possibly contentious framework

[4] Horne and Adali, “This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News,” 5.

[5] Horne and Adali, "This Just In," 6

Horne and Adali provide, the effect of this heuristic form of persuasion remains. Attentive readers can no doubt distinguish between the content of real and junk news from stylistic clues alone, even without the context of author or publication.

The features that allow the distinction Horne and Adali draw remain largely obscure to the reading mind. While a practiced reader may be able to distinguish between junk and real news, it would be hard for them to make an account of the stylistic features that helped lead them to this discrimination. Even in retrospect these specific features may not become apparent — for most readers, the two styles will simply "feel" distinct based on the association of these elements with their respective genres. The verb and noun structures that at least partially subtend the phenomenological experience of discriminating between these texts only become fully apparent when viewed from a third-person perspective outside of the moment of reading itself. Understanding substructral meaning requires a hermenuetics of suspicion of a different stripe — one that does not take phenomenological understanding of a text as its starting point.

While the use of autonomous agents in creating and spreading junk news certainly deserves its own scrutiny — not only through the lenses of public policy and political science, but also as the most disturbingly effective application of computer-generated, -aided and/or -distributed literature to date — junk news also simply represents the most recent entry in a history of substructural style that includes

orthographic experimentation among its many chapters.[6] Orthography, the system of spelling utilized to encode semantically available linguistic units in a given language, has long been used as a tool for producing literary meaning in the substructural fashion. Both orthographic choice and the grammatical stylistics of junk news produce independent meaning in the same way — as these styles of writing enmesh with a consistent literary context over a suitable period of time they develop connotations independent of the content they directly encode.

Being a new phenomenon, a full history of the substructural style associated with "junk news" will remain unwritten until future scholars fully unpack the implications of its use. In contrast, orthographic style possesses an ancient history — one as old as writing itself. Though mainstream linguistic opinion has long held that humans possess a "language instinct," writing itself was an invention — one not initially intended to act as a permanent codification of speech sounds.[7] Archaeological evidence reveals that writing began as a form of accounting and inventorying that eventually transformed into

[6] For more on the substructural elements typical of junk news see Woolley and Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*.

[7] For the most definitive statement of the "language instinct" claim see Pinker, *The Language Instinct: How the Mind Creates Language*. It should be noted that even this thesis does not exist without controversy, and that one of the chief dissenters is himself a central figure in the study of orthography. See Sampson, *The 'Language Instinct' Debate*.

the system of communication we commonly consider it to be.[8] The discovery of writing's extra-linguistic origins calls into question many standard notions concerning writing, primary among them the assumption that writing is merely a crystallization of verbal speech.[9] As linguist and cognitive psychologist David R. Olson argues, this recasts orthography as a participant in a feedback loop between language and technology, acting as "a set of concepts that reveal, make explicit, what was actually there all along," in this case the potential to decompose the continuous sounds of spoken language into discrete phonetic units that are representable in forms originally designed for counting

[8] Schmandt-Besserat, *Before Writing, Vol. I: From Counting to Cuneiform*. Denise Schmandt-Besserat's groundbreaking work demonstrates how the Sumerian use of clay *bullae* tokens to indicate goods slowly became abstracted into a system of symbols before finally reaching the full syntacticization of cuneiform writing.

[9] See Sproat, *Language, Technology, and Society* for an apt example of this "standard" viewpoint. Though Sproat acknowledges the technological origin of writing he still views orthographic processing as largely a mapping of grapheme to phoneme, no doubt due to his professional focus on developing computer-based natural language processing systems. Yet even Sproat stresses the importance of context to orthographic understanding. In his discussion of the Mycenaean Greek Linear B writing system he notes that despite the system's numerous orthographic ambiguities all that matters is that "native speakers can figure out what is being said," a task achieved that documents written in Linear B focus almost solely on "agricultural products" where the graphemes "pe-ma" almost certainly represents "sperma" (seed) despite its divergence from the standard way Linear B should represent those sounds. Sproat, *Language, Technology, and Society*, 46.

and bookkeeping.[10] In this sense phoneticization itself (or syllabization, ideogramization, so forth and so on) constitutes an originary moment of stylistic decision. Translating the continuous domain of sound into a discrete form requires a process of selection that compresses the information originally contained in speech, leaving some elements un-rendered or rendered ambiguously. In turn, the choice of orthographic system employed for this translation feeds back into conceptions of auditory language, altering the understanding of speech itself by emphasizing certain features at the expense of others. Both writing and speech contribute to the other's implementation and elucidation in this reciprocal way with neither dominating the other's independent existence. As 'integrationist' linguist Roy Harris points out, the "invention of the sound spectrograph" revealed that no writing system "captures with any approximation to accuracy the facts of speech."[11] While the mission of writing may have been to capture speech sounds in media more permanent than air, its failure to do so accurately necessitates its status as a separate semiotic system with meaning influenced by, but not reducible to, oral language. In turn, this allows Harris to argue that writing develops a system of meaning that increasingly diverges from any initial

[10] Olson, *The Mind on Paper: Reading, Consciousness and Rationality*, 96. Olson's contention that phonetics inheres hidden in language until technology "unlocks" its potential might be more usefully replaced with a Deleuzian ontology. Such a viewpoint would consider linguistic ability a field of virtual possibilities actualized by concrete instantiations like phoneticization. This allows to avoid the smack of essentialism in Olson's original claim.

[11] Harris, *Rationality and the Literate Mind*, 139.

connection to oral speech, integrating strategies from diverse semiotic systems to generate "its own forms of expression" as exemplified in the styles of writers like Apollinaire, Mallarmé and Joyce.[12]

The conclusions reached by Olson and (especially) Harris remain controversial among linguists, leaving the exact relationship between orthography and speech an open question. Regardless, their interventions destabilize the notion of a tightly paired correspondence between speech and orthography enough to allow the latter room to develop associated sets of style and meaning of its own accord. While Harris pegs the aforementioned set of experimental writers as paragons of writing-specific meaning making, valorizing Apollinaire's and Mallarmé's use of page layout and the extreme orthographic manipulations of Joyce's *Finnegan's Wake*, he misses out on a historically earlier and much more pervasive form of written meaning making — dialect orthography. Harris's oversight is more than understandable. The eccentricity of his example writers' experimentations stand out conspicuously against the backdrop of almost any collection of comparison texts. In contrast, the relatively subtle orthographic modifications employed by dialect works only emerge as meaningful when understood as the choice to use a specific set of dialect conventions that differ from one or more established orthographic standards at the time of writing. Dialect orthography as a source of independent meaning distinct from the words said orthography encodes emerges historically within the life of a given language through a combination of the

[12] Harris, *Rethinking Writing*, 225.

visibility of the variations themselves and their consistent pairing with specific literary-political projects. Substructural styles of this form require standardized notions of orthographic encoding and decoding, notions that themselves shift and modify with the long march of historical time.

This project investigates one such historically-specific formation of orthographic substructural style, the "dialect literature" of the nineteenth century United States. The broad popularity of orthographic experimentation during this period occurred concomitantly with the development of a de facto standard American English orthography, a unique pairing of conditions that allows the period to serve as a paradigmatic example of the development of substructural style. The reading mind of the era found itself faced with an ever-expanding network of orthographic expressions offering independent literary meaning through their relationships to both other nonstandard orthographic choices and the emerging standard. The meaning of these orthographic choices both slipped "below" the reading consciousness by secreting itself in automatic linguistic processing and "above" by enmeshing themselves in a relational network of orthographic meaning too expansive for one subject to fully comprehend. By supplementing traditional literary methodology with a corpus-based computational approach to comparing orthographic sequences, this project will excavate individual moments of this form of meaning both as embodied in individual texts and in the general field of orthography that subtends them.

On a larger scale, this project seeks to further the state of scholarship on nineteenth century United States dialect literature. Work on dialect literature in this

period has evolved away from its roots in phonetics-centric studies. By and large, modern scholars understand that judging nonstandard works by their correspondence to a particular oral speech pattern underestimates the critical insights a more autonomous understanding of written dialect can provide. However, a small phonocentrism lurks even in these modern works. Current scholarship on dialect literature of the nineteenth century often treats the individual linguistic elements that contribute to the determination as a packaged unit. In this approach, choices of syntax, diction and orthography matter in as much as they allow for the determination of the use of dialect and the characteristics (racial, national or otherwise) they ascribe to a given subject. Even though this approach does not tie dialect literature to phonological accuracy, by leveling the differences between the linguistic elements that make up dialect literature it misses the ways each can independently function as a source of meaning. Treating dialect literature as an inseparable bundle does specific disservice to orthography, the most flexible of these traditional elements when used in a literary context by and for competent speakers of a language. This project will use this realization to re-evaluate the use of orthography in the dialect literature of the nineteenth century in a manner that builds on the specific form of semi-conscious meaning orthography provides.

Decoding Dialect Orthography — Orthographic Terms

Orthographic meaning proper — the correlation of systems of standard orthographic style and their exceptions with political-literary meaning — relies on orthography's ability to convey word and sound meaning while not being reducible to either. Central to the premise of this project is that orthography has a dual nature. Linguistically it provides a system of morphological or phonetic hints that allow a reader to decode a word. This means that most linguistic definitions center on its role as a system that define the spelling of words. Following Phillip Baker, the sociolinguist Mark Sebba offers a useful division between spelling, script and orthography. Scripts (such as the Roman script used by the Germanic and Romance languages) contain many orthographies — community-standardized versions of those scripts tailored to the representation of a specific language (e.g. English or Spanish). Spelling, then, is the "application of those conventions to write actual words" in a given language.[13] This project embraces this definition for both its clarity and its theoretical precision — Sebba's tripartite division illuminates both the distinction between use (spelling) and system (orthography) and the distinction between system (orthography) and the medium it inhabits (script). Orthography might also be defined through reference to its constituent unit, the grapheme. As with the term orthography itself, the actual definition of "grapheme" remains a point of scholarly controversy. One traditional definition, what

[13] Sebba, *Spelling and Society: The Culture and Politics of Orthography around the World*, 11.

Dimitrios Meletis terms the "referential view," pegs the grapheme to phonetics.[14] In this view graphemes are distinguished by their ability to represent distinct speech sounds, meaning that they do not necessarily coincide with what we typically call letters or characters. For example, this school would deem the "ch" in "chalk" to be a single English grapheme as it corresponds to the English phoneme [tʃ]. Meletis contrasts this definition with what he terms the "analogical view," where the grapheme is internally defined by its distinctiveness from other graphemes in a given graphematic system. This school adopts the minimal pair process from phonetics in order to argue that a particular character or set of characters that distinguishes one word from another should be considered a grapheme.[15] In this view, both 't' and 'p' would be considered grapehemes on the evidence that they can distinguish at least one pair of words (say "teat" and "peat") from one another. Neither of these possible definitions is wholly satisfactory. The former ties the grapheme too closely to spoken language, denying the possibility of it having independent existence. The latter assumes well-orderedness, that the system of graphematics is complete and unambiguous. These are fatal flaws when dealing with the literary usage of nonstandard orthography. Speakers are forever deferred, leaving open the possibility that a given grapheme might only refer to a private phonetic language ensconced the in the author's mind.

[14] Meletis, "The Grapheme as a Universal Basic Unit of Writing", 26.

[15] Meletis, 26.

Literary use of orthography also does not demand consistency or coherency — the threshold of "readable enough, in context" is a low bar to cross. Rather than wade too deeply into this theoretical debate, this project will take the agnostic route and use the term "grapheme" to refer to a single printed character. This luxury is afforded by turning the focus of this dissertation away from both speech and, to some degree, traditional semantics, a decision that will be justified in the discussions to come. It also offers a distinct advantage when it comes to the computational focus of this project in that it allows for a definition of orthography that supplements the description offered by Baker and Sebba. This notion of grapheme allows one to jettison the notion of spelling employed by Baker and Sebba (a context-heavy facet that inherently ties orthography to sound and linguistic semantics) and define orthography as "*a system, more or less consistent, that directs the ordering of graphemes*." Defining orthography this way gives it some breathing room. It allows for the possibility that orthographic choice might be meaningful on its own, and not just a vehicle for other systems of meaning to employ. Under this definition, Sebba's example of a flyer that reads "Free skool" conveys its specific orthographic meaning (that the author is associated with some sort of liberatory political project) simply due to the placement of the "k" in relation to the other graphemes. Even without resolving the semantic sense of the fragment by determining what "free" and "skool" actually mean, or resolving the fragment to even mental sound by corresponding each grapheme with a particular phoneme, this particular orthographic choice and the system it implies conveys meaning solely by its dissimilarities to other grapheme orderings known or imagined to be possible by its

reader. Without specific knowledge of the political history of the 'ch' to 'k' substitution this reader may be tempted to interpret the spelling of "skool" as an amusingly ironic typo. The orthographic meaning of the phrase arises from the history and politics associated with the particular choice of an orthographic system itself.

The Primacy of Orthography

The principle that orthography has the capability to mean even when devoid of reference outside the field of orthographic possibilities itself might be termed the "primacy of orthography." This is the second, more controversial life that orthography leads. In general, scholars of literature and language interested in orthographic choice tend to treat it as a phenomenon secondary to spoken language. The meaning of orthography is tied, either implicitly or explicitly, to the process of decoding. Computational linguists concern themselves with orthographies of best fit, and the resolution of orthographic sequences to orally linguistic ones. In this discourse orthographies are deemed "deep" or "shallow" depending on the amount of computation required to complete this transfer. The most shallow orthography would have a mapping of one grapheme to one phoneme (or similar unit, e.g. morpheme), while deeper orthographies are more ambiguous on a surface level, and thus require additional steps to achieve the conversion.[16] Sociolinguists, especially those interested in spoken dialect,

[16] The theory was made most familiar by Frost and Katz, "The Reading Process Is Different for Different Orthographies: The Orthographic Depth Hypothesis."

find use in orthography by actually realizing the decoding process. Ethnographic interviews, and the directly recorded speech stemming from them, have typically served as the gold standard for this type of scholarship.[17] However, a paucity of evidence either recorded on auditory medium or in the well-regulated orthography of the international phonetic alphabet (IPA) has historically led some scholars to peruse the written record for phonological insights. Modern scholars of this ken tend to understand the risks of drawing broad linguistic conclusions out from orthographic decoding; earlier scholars, however, were often more cavalier in the face of this risk.[18] The very foundation of the study of non-standard orthography in the literature of the United States emerges from this cavalier approach to text. In his influential short study, "A Theory of Literary Dialect," Sumner Ives takes the opportunity to push back against what he sees as the foundational sins of his field, originally codified in George Krapp's *The English Language in America*. Ives describes his approach as taking seriously "the validity and justice" authors attempt to do to their representations of dialect, in contrast to Krapp's general binary distinction between standard and nonstandard forms.[19] Ives centralizes

[17] Labov, *Language in the Inner City: Studies in the Black English Vernacular* is considered foundational in this sense.

[18] See Krapp, *The English Language in America*. Krapp recognizes that a writer will only record what they themselves view as the "distinguishing marks" of a particular dialect form, but still finds linguistic, and not purely literary, use in seeing "just what the details selected are." Krapp, 231.

[19] Ives, "A Theory of Literary Dialect", 151.

the importance of the decoding function to literary meaning. Indeed, when discussing the elements of dialect writing, Ives implies a tight relationship between a written element and its oral compatriate. Characters in works that utilize dialect "are made to speak" using written features, meaning that non-standard orthography becomes "unconventional features of pronunciation" mediated through "phonetic spelling."[20] Deviating from this plan is a negative trait, and condemns the author to the sin of eye dialect, linguistic variation without function. By binding meaning to decoding, Ives concomitantly binds the possibilities of orthographic expression, and denies the possibility that it may have another way to offer meaningful reference.

This is not to say that Ives is blind to the complexities of the actual process of decoding written utterances into spoken ones. He spends a large portion of "A Theory" qualifying his statements about authorial intent towards phonetic accuracy, and acknowledges that decoding is a complex process contingent upon the potentially quite varied knowledge of linguistic standards and deviations held by both reader and writer. This view accords with his contemporaries, who also understand the ambiguities of decoding while still binding orthography to its strictures. Richard Venezky's seven general principles of American orthography, as laid out in his foundational study *The American Way of Spelling*, represent a more general linguistics approach to this ambiguity. Venezky's rules remind the scholar of orthography that, at least in American English writing, "variation is tolerated," and that "regularity is based on more than

[20] Ives, 147.

phonology."[21] Venezky bakes ambiguity into his exposition of North American English orthography, even leaning on the "richness of both variability and irregularity" a feature that he sees as unique facet of North American orthography, but mainly does so in order to minimize its impact.[22] Being a linguist rather than a literary critic, especially in the 1970s, almost necessitates some sort of payoff in oral language. To this end Venezky's book largely focuses more on the sound system associated with his orthographic subject, focusing more on phonetics than graphematics. These historical acknowledgements of the ambiguity inherent in the decoding process lead to questions concerning decoding, especially when viewed through the lens of fidelity to a spoken dialect, falling justifiably out of favor with more modern critics of North American literary dialect.[23] Even more linguistically-inclined modern scholars of dialect like Lisa Minnick preserve accuracy to speech in only a qualified sense, as a feature that "raises important questions" about the use of dialect markers, even if it is just what the "significance of the level of accuracy" in written dialect actually signifies.[24]

---

[21] Venezky, *The American Way of Spelling*, 6.

[22] Venezky, 14.

[23] Roger W. Cole, "Literary Representation of Dialect: A Theoretical Approach to an Artistic Problem" serves as an exemplar of this view. Cole calls for the wholesale abandonment of linguistic principles in the analysis of dialect, instead arguing for a more internal approach to context.

[24] Minnick, *Dialect and Dichotomy*, 33. Minnick takes a middle road, where the "best practitioners of literary dialect create effects that are linguistically and artistically believable" but also where writing is unable to "reproduce spoken language exactly". What remains open here is what "linguistically and

What scholars of Ives and Venezky's era share with more modern researchers of literary dialect are the twin certainties that literary dialect consists of a unified bundle of textual features and that literary dialect features necessarily provide clues used in an individual reader's decoding even if the process itself is ambiguous or unimportant. Ives's dictate that literary dialect consists of "unconventional features of pronunciation, grammar and vocabulary," each represented by a corresponding feature in writing, remains largely uncontroversial in the modern era. A given piece of dialect literature may not necessarily use all of these features to produce its effect, but they all remain bundled together under the sign of "dialect." This attitude represents a small but impactful phonocentrism. Gathering these three elements into the fold of "dialect" flattens their differences and implies that they function in a unified fashion. This attitude does a particular disservice to orthography, the most writing-centric substructural element of the three. Even if one repudiates the importance of dialect accuracy, such an attitude remains too linguistic. Carefully examining the process of decoding orthography as well as its origins and general characteristics reveals that its process of decoding, whether the ultimate intention is to judge on the basis of accuracy or not, simply proceeds differently from the other two substructural elements. It also calls into question how much a literate reader relies on orthography to produce a more-or-less accurate decoding of speech sounds at all. Orthography as a specific

---

artistically believable" means, especially given further examinations of decoding. Minnick ultimately points out that the context provided by linguistic study of speech can be useful to the literary critic interested in dialect — context that can also be provided by a writing corpus alone.

element functions in a more deictic aspect than it appears on first glance, leaving it much expressive power to expend on producing a sense of meaning of its own.

The actual process of decoding graphematic sequences supports the argument for a much stronger notion of independent orthographic meaning. Literary usage of non-standard orthography implies the existence of an audience less interested in orthography for the clues it provides during the decoding process and more so for its potential aesthetic ends. Henry James, a writer not known for his use of non-standard orthography, provides an instructive example of how extreme the situation of decoding can become when the usage of a literate audience is pushed to its limit:

He came, in fact, from Mississippi, and he spoke very perceptibly with the accent of that country. It is not in my power to reproduce by any combination of characters this charming dialect; but the initiated reader will have no difficulty in evoking the sound… And yet the reader who likes a complete image, who desires to read with the sense as well as with the reason, is entreated not to forget that he prolonged his consonants and swallowed his vowels, that he was guilty of elisions and interpolations which were equally unexpected, and that his discourse was pervaded by something sultry and vast, something almost African in its rich basking tone…[25]

[25] James, *The Bostonians*, 5.

Had James used this passage from *The Bostonians* to depict, rather than describe, the African roots of Basil Ransom's southern speech, untold multitudes of dissertations would have benefited thereby. Had this (unlikely) counterfactual occurred, the orthography used to encode Ransom's voice could have become a locus of critical argumentation. The correspondence between James's claims to orthographic accuracy and his actual representations would become open to critical analysis in the same way that Mark Twain's boast that he "used a number of dialects" in the composition of *Adventures of Huckleberry Finn* has been ceaselessly examined for evidence that would support or undermine the diversity and veracity of his dialect orthographies. Decoding could have become a locus of potential meaning. Instead, James critiques the very mission of dialect literature by deferring the representation of accent to the reader. This critique implies that authors who perform elaborate orthographic acrobatics in order to 'better' represent dialect to their audiences misplace their efforts. Since the reader ultimately produces any given dialect using their own stock of experiences a simple deictic clue indicating which general phonology to employ will suffice. Though James's statement exudes strong overtones of satire, elitism, and racism it also smacks of a certain truth, an especially cutting one given that excerpts from *The Bostonians* were finding publication in *Century Magazine* alongside the dialect-peppered *Adventures of Huckleberry Finn*. Orthographic hints can certainly point towards 'correct' pronunciations but only when they are supported by a substructure of convention. The evidence of this can still be found in current debates over language instruction and

reform. The reader properly "initiated" will likely glean what sort of dialect to employ from context and ignore any further orthographic clues while the novice will lack the conventionalized substructure particular to dialect literature rendering them unable to recognize or "correctly" pronounce the intended dialect.[26]

Take, as an example, these renderings of the English word rendered in IPA as [wej]. Standard American English orthography encodes this word as "way." Yet "wey" or even "wa" seem to work just as well. Further confusion ensues upon the realization that some American English speakers pronounce the word with an audible "h"-like sound as [ʍej], even though they still write the word using the standard "way" sequence (these same speakers also likely pronounce the 'h' in words like "what" — thus making the dialect/voice transition inconsistent in both cases). Even if the addition of this sound tempts a change to the orthographic representation "whey," the standard orthography already uses this sequence to represent the cheese by-product, most often pronounced without an "h" sound and thus homonymously with [wej]. When these hypothetical readers find themselves confronted with an unfamiliar dialect orthography and must render it (whether mentally or audibly) into speech-sounds they can only rely on their disparate, community-learned rules of orthographic/phonological

[26] Evidence from the very apex example of non literary-competent readers, developing learners, even provides good evidence for non-phonological and context-based effects having more impact in decoding than previously thought. Theories integrating these insights include include dual-route and multiple pattern theories. See Treiman, "Learning to Spell Words: Findings, Theories, and Issues."

correspondence (eg. "way" -> [wej] vs. "way" -> [ʍej]) and the sounds of dialect speech they have previously encountered as guidepoints to interpretation. This both curtails the ability of writers who render dialect in orthography to control their readers' phonological interpretation of their writings, and de-centers the "expertise" of the dialect writer by showing the inherent provinciality of her own speech and orthographic representations — renderings are only "ideal" when relativized to the ortho-phonemic norms of the writer herself. James's passage seizes upon the Wittgensteinian humor of this situation. When the dialect/phonology interface is properly grasped as contingent upon community and experience, it seems almost easier to achieve a "dialect effect" by modifying the context in which a reader takes an utterance (e.g. "read Basil as though he is Southern") than by modifying the orthography itself. Seen from this angle, orthographic experimentalism no longer seems an exercise in fidelity. Instead it becomes its own end, a situation James cannot help but humorously undermine in a manner that calls into question the wisdom of over-investing in any particular theory of decoding at all.

Harris's chosen canon of "high literary" orthographic experimenters, Mallarme, Joyce and Apollinaire, are a nod to his own particular investment in orthography as its own end. While these three arguably do loom large in the history of orthographic substructural style, the technique also found a large amount of use in popular "dialect" literature. Dialect works span multiple genres, time periods, and authorial subject positions, but are typically unified in their usage of elements of substructural style to depict in writing some sort of particular regional-, racial-, or class- specific voice. Dialect

literature represents a historically earlier and ultimately more prolific incarnation of orthographic substructural style than Harris's exemplars, but perhaps escapes his notice for a few particular reasons. For one, dialect literature experiments with more substructural elements than orthography alone. Dialect works that use non-standard orthography also tend to use non-standard syntax and vocabulary to achieve their effect. These additional elements pull the reader's focus away from orthographic play, thus possibly diluting Harris's point. More importantly, the addition of non-standard syntax and vocabulary tempts the conclusion that dialect literature solely uses these elements as a means of representation of particular regional, racialized or classed spoken dialects rather than as a means of "free" expression. The addition of these elements signals an authorial desire to represent some spoken form *in toto*; their omission signals that the author is interested in the freedom of letter order choice to express some other further meaning not necessarily connected to the linguistic properties of an imagined speaker.[27] In effect, this view argues that syntax and vocabulary shackle orthography. Orthography is sufficiently distant from its role in the transmission of information to draw on non-linguistic forms of substructural meaning, while syntax and vocabulary choice must leave at least one foot in the linguistic domain. Combining these substructural utterances into written language then only serves to further subjugate orthography to its decoding function — a function James's example proves is redundant for an informed

[27] Apollinaire's *Calligrames* serve as the most clear enaction of this principle. The non-standard orthography of concrete poetry produces additional meaning through reference to visual art, something almost totally non-linguistic.

reader.

This project adopts the core tenet found in this extrapolation of Harris's canon — that orthography is uniquely suited to provide a form of meaning beyond decoding, or even linguistics — while discarding the implied conclusion that dialect literature's usage of other nonstandard substructural elements disqualifies it from making use of this capacity. When viewed in comparison to the syntactic or diction elements of writing, orthography does seem to live a less-linguistic life of its own. Take the following variations on Chomsky's famous example utterance "colorless green ideas sleep furiously":

kolerls grn eyedeyaz sleip phyoursly

sleep green ideas colorless furiously

Both of these examples mutate the typical usage of one of the substructural elements of written language to a roughly equally excessive degree. The first uses non-standard orthography, the second an unusual and disordered syntax. Both attempt to perform a decoding function — the process by which the recipient of the text translates it into a spoken or semantic utterance — but perhaps at varying levels of competency. Each also very conspicuously provokes the question of their further substructural meaning, simply by being so eccentric in their variation. Despite these similarities, the actual process of

reading each utterance varies greatly. The orthographic example reads in a somewhat bumpy form of the standard English left to right fashion. Certain words or orthographic choices may demand a conscious pause in the typically automatic decoding process, but by and large (and surprisingly) the utterance is legible. The same cannot be said for the syntactic example. The syntactic example possesses fewer deformations in at least one absolute sense. All four of the words have been rearranged, but this pales in comparison to the number of character replacements used in the orthographic example. Despite this, it still performs the decoding task much more poorly. Scanning the sentence requires frequent conscious intervention in order to reconstruct the higher order syntactic structures of the sentence and successfully decode its meaning. The smooth, "at-hand" nature of reading recedes in the face of such variation, an effect that remains surprisingly intact in the orthographic example. This effect is typically explained as a consequence of the redundancy of orthographic sequences. When related to *information entropy*, a concept that will play into our later methodological discussions, the notion of redundancy encapsulates the fact that each individual character in an orthographic sequence provides a large amount of information about what the following characters are likely to be. Typically this comes into play when dealing with the elision of characters. This redundancy, as well as higher level semantic and pragmatic context effects allow garbled or corrupted words to still be decoding-ready.[28] The power of

[28] The creation of encodings for telegraphic communications represent a pre-theoretical understanding of this concept. A more available modern example is the sort of abbreviations used for communication in textual messaging services, e.g. "pls" as a replacement for "please". Richard Bridgman's dictate that

redundancy also enables orthographic sequences to be substructurally creative. Authors can add, swap or even invent sequence elements with little fear of impacting their reader's ability to decode. These elements have the freedom to be substructurally meaningful, to gain surplus meaning over the function of decoding or the decoded sentence alone.

Syntax's relative lack of redundancy grants it less expressive flexibility. It is not entirely rigid — Horne and Adai's example of computational propaganda works syntactically, and minor word-order changes are an element of many dialect works. However, its origin as a function of language gives it less breathing room than orthography. The upper limit on syntax's ability to express meaning beyond decoding is at least partly due to it being tied more closely to oral language than orthography. As Schmadt-Bessarat's work shows, orthography is repurposed invention, drawn from a visual accounting system. Syntax, on the other hand, is linguistic instinct. Scholars of language even typically cite syntax as a key feature that distinguishes human speech from the sound systems of other animal life.[29] The two converge in writing, bringing with them the distinct markers of their original medium. Oral language, and the syntax

---

"uncompromising dialect is exasperating to read" marks itself very much as a statement contemporary to the 1960s, when such pervasive abbreviation and invention were not in play. Bridgman, *The Colloquial Style in America*, 50. Bridgman also misses out on an important general point — a literately competent reader likely won't doggedly decipher every dialect utterance phonetically, and instead rely on context clues to make decoding simply work.

[29] See Pinker, *The Language Instinct*.

that structures it, is natively ephemeral. Airborne sine waves are a volatile medium, and successful communication requires adherence to strict principles of order that allows for the reconstruction of an utterance's meaning even under imperfect conditions. The order of a sentence must largely accord with the sequencing expected by the instinctual apparatus used to decode it in order to guarantee the decoding function. Orthography's birthright is medium permanence, gifted to it by its origin in visual culture. As a consequence, literary orthography possesses a very different relationship to oral speech than written syntax. Orthography needs language less than language needs orthography (and vice versa) in a manner syntax does not mirror. The linguistic chicanery James employs in his description of Basil Ransom's speech simply cannot extend to syntax. Should he have wanted to Basil to employ grammatical features associated with a particular southern accent, simply stating that Basil uses them and then not representing them in writing would not have passed muster. Grammar's centrality to decoding the overall meaning of a sentence makes simply stating that Basil "replaced the indicative 'this' with the more typically southern 'this here'" rings false as opposed to actually instantiating it in text. When this strategy is applied to grammar instead of orthography it risks producing semantic confusions on the level of the sentence and constitutes an inaccuracy in a stricter sense than any similar orthographic variation. For a literate audience orthography's purpose in decoding is simply to hint — a function that can be replaced semantically. If grammar acts the rules to a particular game, orthography functions more as a metagame statement, an indication of what prior knowledge to bring to the game of decoding in order to play. In that it resembles a

metagame more than a game itself, orthography has more relation to semantics than syntax. Understanding the purpose and meaning behind a certain orthographic declaration (either through use, or in James's case, fiat) is just as historical and contextual as analyzing the meaningful aspects of any given sentence. These two elements of dialect, unlike syntax, are not *of* language — they are adapted to it.[30]

The original version of Chomsky's famous example — "colorless ideas dream furiously" — was designed to illustrate the extent of the syntax-semantics cleavage. Chomsky's sentence is perfectly grammatical, but semantically nonsensical. The two elements lead independent lives and belong to independent systems of meaning. Syntactically, the sentence indicates that some subject, described by some additional modifier, is performing an action in a certain mood. Semantically, the sentence means nothing, or at least serves as a poetic incitement to the production of meaning.[31] The same relationship holds for our orthographic version of the sentence. The modification of orthography does not the change syntactic or semantic meaning; it is free to allude

[30] Just as orthography emerges from a place other than language, modern researchers of cognition often argue that the basis of semantic meaning stems not from language itself but instead from a pre-linguistic understanding of the body's relationship to the environment. See, for example Clark, *Being There: Putting Brain, Body and World Together Again*, as well as the perhaps more familiar work of Maturana and Varela.

[31] Meaning is, however, historical, and so this sentence might now have a particular meaning associated with it — "a scholar of language is trying to be clever."

outwards as it wishes.[32] Despite their similarities these two elements also operate independently of one another. Modifying the orthography of the word "furiously" to be represented by the graphemes "frsly" is an act of meaning-making independent from the meaning of the word itself. Substructural modifiers of this kind belong to their own semantic realm, one that can be explored by its own appropriate set of tools.

The Nature of Substructural Style

Literary criticism typically concerns itself with the investigation of semantic phenomena. Close reading, whether backed by a hermeneutics of "depth" or "surface" ascribes meaning to portions of a text based on a process of contextualizing and exploring the meaningful possibilities they can bear.[33] The history of close reading is a story of expansion. Literary critics have extended the use of the technique to analyzing semantic phenomena of all stripes — systems of fashion, film, visual art, and so forth. Despite it being a semantic phenomenon akin to these examples, orthography has not,

[32] There are extreme cases that disprove this rule — an unreadable orthography, or an orthography that leaves word meanings fundamentally ambiguous due to making dissimilar wordforms too orthographically similar.

[33] Best and Marcus, "Surface Reading: An Introduction", offers a defense of surface against the typical notion of hermenuetic depth, perhaps best represented by Frederick Jameson. Both, however, tend towards a similar use of close reading — this is a controversy over close *qua* depth versus close *qua* surface.

in general, received such a treatment. The nature of substructural style, non-standard orthography included, frustrates approaches that work in a top-down or phenomenology-first manner. This resistance is anathema when one is attempting to use close reading as at least an entrance into a literary phenomenon. J. M. Coetzee, once more a linguist specializing in literary stylistics and less a novelist, eloquently stated the problem as early as 1969:

"The work is thus like a mountain, its lower slopes habitable by the positivist, its peak lost in the clouds. One reaches the peak by climbing from the lower slopes. To start at the peak and make the lower slopes one's objective, or to run up and down the mountain, is senseless."[34]

Coetzee's metaphor works as an appropriately stylized account of the difficulties the critic faces when attempting to square the "low-level" substructural stylistic elements of a work with the phenomenological experience of reading. Coetzee argues that an aporia exists in the process of reading. Lower level substructural elements produce certain experiences that operate phenomenologically, yet connecting the two remains problematic. Ultimately Coetzee questions the commensurability of low level and high level explanations. Commenting on his stylistic analysis of *Watt*, Coetzee

[34] Coetzee, "The English Fiction of Samuel Beckett: An Essay in Stylistic Analysis", 6.

writes that the “underlying rhythm of Watt’s thinking” is the “rhythm of doubt.” Though this rhythm “could not exist without the words, without these particular words” it is finally “the rhythm alone that we hear.”[35] The literary mountain, though clear at the heights of phenomenological experience (“we hear”) and the base of quantitative stylistics (“these particular words”) remains ineluctably clouded across its middle region. The traveler knows she can traverse from low to high and high to low, but the route she takes (and “outside” forces of context or history that assist her) remain unknown to even herself.

Coetzee's illustration plunges into the murky depths of the philosophy of mind, a necessary complication given the nature of substructural style. The situation he describes draws upon the literary version of a cluster of issues surrounding the transit between first and third person perspectives examined in that field. As it turns out, approaching the functioning and history of a cognitive process like orthographic processing as approached from the top down phenomenological point of view inherits a set of issues similar to those encountered when attempting to imagine the "something that it is like to be a bat."[36] Thomas Nagel’s famous thought experiment relies on the belief that bats have some form of phenomenal experience, an attribute we might not so

[35] Coetzee, 95.

[36] Nagel, "What Is It Like to Be a Bat?", 438.

readily ascribe to a cognitive subsystem.[37] Regardless, the issue of basis applies in both cases. Just as attempting to act in a bat-like fashion in order to experience bat-ness only reveals what it "would be like for [one] to behave as a bat behaves" attempting to experience orthographic processing phenomenologically only offers insight into what it is like for a higher-order cognitive self to decrypt orthography.[38] One could imitate the way one believes orthographic decoding might occur at the low level (say, by using a table of grapheme to phoneme conversions to iteratively decode the word) but the experience of this process would not be the "usual" experience of orthographic processing itself. Unlike sensations intended to provide actionable feedback (i.e. pain) the feeling of orthographic processing does not offer differentiable information that provides information about its cause and source. Barring the Jamesian "top-down" incitement to interpret non-standard orthography a particular way, the flow of information moves in an irreversibly bottom-up direction. The actual experience of decoding frustrates any attempt to treat the end result of orthographic interpretation as some sort of dream form by offering no information about the orthographic system and cognitive subsystems it supervenes upon.

---

[37] Nagel's point about the incommensurability of cognitive experiences is only a secondary aspect of the paper, with the primary being an argument against reductionist approaches to mind. This use of Nagel is then surely an abuse of academic technology, as we flirt with reductionist theories below. However, given that Nagel's desired "objective phenomenology" has not emerged in any meaningful form, the reappropriation of this sub-argument will hopefully be forgiven. Nagel, 449.

[38] Nagel, 439

The dogged insistence on the importance of history, politics and semantics dooms the literary critic to suffer most from this realization. Scientific naturalists simply invoke their birthright and employ tools like functional magnetic resonance imaging (fMRI) to cut the Gordian knot of consciousness and localize orthographic processing to known neural regions. Taken to the furthest extremes, this approach traverses into the domain of the eliminative materialism advocated for by Paul and Patricia Churchland. Eliminative materialism argues that "commonsense conceptions of psychological phenonmena" are not simply false but in fact "radically false" in that a completely comprehensive third-person perspective-based neuroscience will leave them "displaced" instead of "smoothly reduc[ing]" them.[39] In short, for the Churchlands, theories about the causes, ontology or classification of mental phenomena that stem from any sort of "folk psychological" first-person introspection on the character of a mental state have no purchase on the actuality of what happens on a lower neural level. These states don't exist in any real sense; they are phantasms available to higher order cognition produced for some beneficial reason (say, general status monitoring). The lower-level processes that produce these states do not leave their mark on them in any way, and thus cannot be distilled from the experience of the state itself. The dream form, in short, has no power in this regime. To borrow a term from Katherine Hayles, in the realm of the "cognitive nonconscious," a natively "third-person," non-phenomenologically based

[39] Churchland, "Elminative Materialism and the Propositional Attitudes", 67.

tool-set becomes a necessity.[40]

Rather than adopt the eliminative position wholesale (even though it can be a fruitful one, and has led to new interest in positions that are antihumanist or pessimistic), this project adopts a qualified version of this understanding of mind as a means of understanding non-standard orthography and substructural style in general.[41] The actual experience of reading orthography offers far less than the reality of the situation — the interaction between lower-level cognitive subsystems and a particular orthographic system they attempt to decode. Barring a Jamesian sort of higher-order intervention, this interaction is subject to the conditions Churchland describes. The phenomenal result of decoding non-standard orthography simply is not an account of the orthographic features that composed a textual sequence or the means by which they were processed. Despite this, one need not become a neuroscientist or resort to evolutionary psychology to win some form of third-person perspective on the decoding process. Regardless of one's general take on the nature of mind, the moment of orthographic processing is a neurology-centric situation on which phenomenology has little purchase. However, it is still first and foremost a historical situation, one

[40] Hayles draws on the concept of a cognitive nonconscious in a less eliminative way, using it mostly as a concept that encompasses assemblages of human and nonhuman thinkers. This perspective is actually somewhat compatible with the Churchlandian view, so long as one sees the brain itself as an assemblage rather than a singular entity. See Hayles, *Unthought: The Power of the Cognitive Nonconscious*.

[41] For example see Metzinger, *Being No One*. Metzinger seeks to deny the veracity of the phenomenal self in total, which is perhaps a bridge too far for this project.

conditioned on both the neural and orthographic fronts by the technological nature of writing. If orthography is anything, it is a medium carried within, shaped by the historical forces that govern writing and instantiated in individuals through vectors of schooling and culture. Literary criticism already plays host to a number of methodologies that have purchase on these phenomena from a third person perspective that deal less with phenomenally experienced content and more with the specific elements that produce these experiences. Purely historical approaches (e.g. book history), media studies and computational humanities approaches all fit this description, and all can contribute to the study of substructural style.

History serves as a source of comparative context. Comprehending the meaning of substructural elements requires both a historical understanding of the social conditions that produced a particular substructural style and a field of possibilities that style inhabits. Substructural elements only develop meaning in relation to this field of possibilities, a statement that likely holds true for literary meaning in general but becomes more complicated when the comprehension of these possibilities cannot occur purely phenomenally. Grasping the potential meaning of a close-read word or phrase necessitates a first-person understanding of the semantic possibilities associated with an utterance, a task analogous to but ultimately distinct from determining the import of the difference between the placement of elisions in two different orthographic systems. This project builds that necessary context by using a corpus approach to substructural style. A common methodology in linguistics, corpus stylistics has recently been making

inroads in literary criticism.[42] This technique argues for the significance of even substructural stylistic differences in particular by building a third-person image of the styles used in contemporaneous texts. While this process is often performed computationally, counting and performing statistical analysis by hand is still relatively common. Frequently, a distinction is drawn between "corpus-driven" and "corpus-based" approaches. The former treats the corpus as a realm of empirical discovery and encourages the investigator to approach it without prior assumptions about the possible regularities they might find, while the latter approaches the corpus with a hypothesis it seeks to prove or disprove.[43] Rather than adhere to one of these positions, this project ranges between them. Each pole represents a valuable aspect of exploring substructural style. The former allows for the gulf between noncognitive and cognitive experiences of style. Recurring stylistic elements can escape active cognitive notice. They remain stranded somewhere on the middle of Coetzee's stylistic mountain, yet potentially actively contribute to the experience of reading. The latter corpus-based approach allows for the possibility of bilateral transit between these two levels. The phenomenal experience of some stylistic element may well offer some hypothesis about lower level features that proves correct upon examination. Remaining agnostic on this

[42] A number of recently published volumes signal this inroad. See, for example, Fischer-Starke, *Corpus Linguistics in Literary Analysis* and Hoey, Mahlberg, Stubbs and Teubert *Text, Discourse and Corpora*.

[43] The distinction originates with Tognini-Bonelli, *Corpus Linguistics at Work*.

front respects the complexity of the interaction of these levels in a fruitful manner.

Regardless the choice of specific corpus-based approach, corpus methodologies in general answer a common critique of stylistics, perhaps most famously voiced by Stanley Fish. Fish argues for the incoherence of literary stylistics as commonly practiced by appealing to the ontology of text. For Fish, a text does not exist outside of its moment of reading in a particular context. He does not "deny any relationship between structure and sense" but does insist that if there is one "it is not to be explained by attributing an independent meaning to linguistic facts."[44] Surprisingly, corpus stylistics is not wholly incompatible with this statement. When used to analyze substructural style a corpus is a record of reception, of the general environment a work jockeys for position in. It too believes that linguistic units need context to have meaning, but denies that this context must come exactly at the conscious moment of reading. Fish's objection relies on a holistic form of mind, the very sort that the nature of substructural style occludes. While "top down" interpretation can certainly contribute to substructural meaning (again, drawing on the example of James and Basil Ransom), a large chunk of this meaning has been nonconsciously processed before any lucid discussion can even begin. Orthographic decoding, the meeting of a system of decoding and a system of orthography both produced by historically specific and ultimately contingent factors, already counts as form of reception well before the cognitive experience of the text takes hold.

[44] Fish, *Is There a Text in This Class?*, 77.

Taking this line re-ignites the risk of reduction to neuroscience. If at least part of substructural meaning stems from nonconscious interactions between a neural system and the elements it decodes making claims on the results of this process risks a sort of Chomskyean "functionalist guarantee." As a part of his project to put linguistics on scientific ground, Chomsky argued that a correct and minimal model of some mental phenomenon necessitates the reality of this process at a neural level, even if it is not instantiated in a direct fashion.[45] For Chomsky, discovering facts of language means discovering neural facts, actual information about how some process works *in neuro*. This project does not aim to make such claims. Regardless of one's attitude towards the accuracy of this guarantee, Chomsky's claim is couched in a linguistics that puts oral language, the instinctual portion of human communication, first. In theory, this is the realm of invariants, elements that persist regardless of language or era. Orthographic style does not share this form of being. Writing is a historically developed technology, not an innate capacity, and is subject to contingency in both encoding and decoding. There is simply no guarantee that any particular account of how decoding occurs or how meaning becomes associated with particular orthographic system captures the low-level neural process an individual uses to make such determinations — there was even a time when this did not happen at all. Insisting that each reader mentally iterates through a particular orthographic process one discrete step at a time would equate the

[45] See Chomsky, *Cartesian Linguistics*. The functionalist guarantee eventually fed into the functionalist theory of mind (perhaps most stridently advocated for by philosopher Jerry Fodor) and also gave rise to the modern discipline of cognitive science.

orthographic element of reading to long division, a move as sinful as it is inaccurate. The proper object of orthographic style is not its neural situation, but instead the unseen factors that allow the interaction of particular orthographic systems to occur and the resultant character these interactions produce. In literary scholarship the third-person investigation of these situations is often the domain of media studies.[46] The discipline itself was founded with writing at its center, whether this came in the form of Marshall McLuhan's exposition on the nature of "literate man" or Elizabeth Eisenstein's magisterial study of the impact of the printing press on cultural notions of the self.[47] As such it is a discourse accustomed to making claims on the development of a phenomenal sense of self from a third-person perspective. Though the primary thrust of Walter Ong's declaration that "features we take for granted in thought and expression" come about only due to the historical intervention of "the technology of writing" has been long the subject of controversy and challenge, the approach to meaning he and his followers provide still proves invaluable when scrutinizing substructural style.[48] Taking up this

[46] Though not exclusively. For example, Walter Benn Michaels flirts with similar territory in his discussions of the importance of intention to literary meaning. The moderately reductive position this project adopts makes his work in particular a less suitable theory of meaning than those typically promulgated in media studies; intention may well be important, but what if intention is not what we generally think it is? See Benn Michaels, *The Shape of the Signifier*.

[47] See McLuhan, *The Gutenberg Galaxy* and Eisenstein, *The Printing Press as an Agent of Change*.

[48] Ong, *Orality and Literacy*, 1.

mantle to some degree, Katherine Hayles argues in *How We Think: Digital Media and Contemporary Technogenesis* that thought extends beyond the neural to "conscious and unconscious perceptions," and the technological assemblages both utilize. As a result, the boundaries of the body are blurred "challeng[ing] our ability to say where or even if cognitive networks end."[49] Even if the processes involved in orthographic meaning lean more towards the unconscious and distributed ends of this spectrum, Hayles points out that it is the interaction of these historically bound systems that is responsible for producing meaning, making them perhaps more appropriate subjects for literature than neuroscience. The variations and distinctions inherent in a substructural style are both historical and not completely neural even if they are not directly phenomenologically available. Somewhat in spite of his Lacanianism, Friedrich Kittler's work on media might best reveal how the real (cognitive nonconscious) is best grasped through the third person perspective. For Kittler, what emerges as the "noise" or "nonsense" of various ways of integrating the real into the symbolic order (the conscious cognitive) across various media (orthography, recorded audio, etc.) points to the content of the real. In this view, the symbolic order is "simply an encoding of the real in cardinal numbers."[50] To derive meaning from these elements is not to delve into a secret or unveil what has been hidden but simply to process things differently.

[49] Hayles, 17.

[50] Kittler, *Discourse Networks 1800/1900*, 328.

Decoding Dialect Orthography in the United States

Historically speaking, dialect's emergence as a stylistic technique requires a contrasting notion of "standard" speech and orthography. Bereft of a centralized linguistic authority, the United States developed linguistic "norms" primarily through the homogenizing work of public schooling and a national media, processes by their nature work that unevenly, in fits and starts. Prior to the United States's acrimonious split with the British empire America had long played host to a variety of English dialects, both spoken and written.[51] As Jill Lepore demonstrates, the development of standardized orthography began in earnest during the early federal period as a way to "build American's fragile sense of national belonging."[52] Early attempts at creating a distinct "American language" relied on individual projectors like Daniel Webster, whose American Spelling Book became a defacto schoolroom standard during the early nineteenth century. Without the high-handed edicts of an American *les Immortals* similar projects of linguistic identity building have continued apace even to the current day. Scholars have comprehensively tracked how such standards of speech and writing have emerged and been contested across the course of United States history, highlighting the importance of factors including racialized and classist distinction,

[51] See Dillard, *A History of American English*.

[52] Lepore, *A Is For American*, 6.

mandatory education, and the prominence of newspapers, radio and television.[53] On a broad level what is most important for the purposes of investigating dialect orthography as a form of substructural style is the bare fact that identifiable standards did emerge out of this stew of factors. The contrast of an orthographic standard endowed with political and cultural power, even such a variously constituted one, endows the use of a non-standard orthography with meaning beyond personal style — the standardization arc of a given language generates the affordances that grant the individual choices of a particular piece of dialect literature meaning.

Although the flow of work on the subject has dwindled in recent years, the specifically literary (rather than linguistic) study of dialect in United States literature was once a mighty institution. The mid-twentieth century proved especially conducive to this form of criticism, perhaps in no small part due to the relative lack of institutional boundaries between philology, linguistics and literature prior to Chomsky's reformulation of linguistics as a mind science. The literary critical narrative of dialect

[53] Baron argues for some amount of continuity in the process of developing modern standard American English from what he deems "Federal English". Baron, *Grammar and Good Taste*, 12. The actual content of this continuous account takes many forms depending on the critic. For example, Bonfiglio draws on the racialized fear of Jewish and Italian-American others in the early twentieth century to provide an account of how the rural midwest accent, rather than the dialect associated with the cultural hubs of Boston and New York City, became the basis for the emerging standard. Bonfiglio, *Race and the Rise of Standard American*, 4. Regardless of the preferred narrative, debates over spoken dialect inevitably bleed into literary orthography, even if they do not determine it.

literature naturally stems from this origin point, and largely consists of scholars debating or modifying the basis provided by these first critics. A standard narrative of the changing attitudes towards literary dialect in the United States might start with the "vox populi" narrative largely preferred by the pre-war critics. These critics viewed literary dialect as primarily (though not entirely) a tool of folk humor, a populist literary expression too long ignored. In the introduction to their collection of early American humor writing *Mirth of a Nation* Walter Blair and Raven McDavid assert that American humor writing went long unnoticed at least in part due to the "earthy language" employed by these authors, leading to them being considered "inferior to more conventional writers."[54] For Blair and his contemporaries like Constance Rourke the writing of these American humorists had largely been ignored at least in part due to their usage of non-standard orthography, grammar and vocabulary.[55] To these critics the "down home" humor of Artemus Ward, Seba Smith and George Washington Harris did more than provide ample knee slappers and linguistic boners to chuckle at; they provided a satirically encoded view into the political beliefs of the nineteenth-century working class of the United States. In the following years, more "serious" works of later, regionally-specific literature that used non-standard linguistic features as partial

[54] Blair and McDavid, *The Mirth of a Nation*, ix. This particular publication comes from near the end of Blair's long career, but a similar sentiment was also expressed in his much earlier 1937 collection *Native American Humor*.

[55] See Rourke, *American Humor*.

elements also drew the notice of the field. Critics of this period focused especially on the perceived accuracy of a given writer's dialect system, an attitude already previous explored in Blair's exposition on Joel Chandler Harris's *Remus* stories and eye dialect. Twain too became subject to this form of scrutiny, with many critics evaluating the claims to linguistic prowess he makes in the introductory section of *Adventures of Huckleberry Finn*.[56]

The next wave of scholars largely objected to this original agenda. Modern researchers like Stephanie Foote and Alan Trachtenberg have almost completely inverted the "vox populi" hypothesis. These more modern studies show quite conclusively that dialect writers, especially in the post-bellum period, were mostly elites producing works for an elite readership of (similarly elite) literary journals.[57] Both Foote and Trachtenberg stress the importance of William Dean Howells and his editorship of *The Atlantic Monthly* to the dissemination of dialect literature within elite literary circles.[58] While Foote and Trachtenberg reach somewhat different conclusions about the

[56] Most notably Carkeet, "The Dialects in Huckleberry Finn."

[57] A conclusion almost fated by the centrality of *The Biglow Papers* in the American literary dialect corpus, a work written by poet, ambassador, *Atlantic Monthly* editor and general high-brow James Russell Lowell.

[58] Trachtenberg sees Howells's task (in his role as high-literature gatekeeper) as policing the bounds of taste. In the minds of such gatekeepers, the "artist of the real is the artist of America" attempts at a vulgar form of realism (such as dime novels) must be quashed. Trachtenberg, *The Incorporation of America*, 196. Foote sees regionalism and its dialects as performing a nation-building function, filling the

impact of dialect's high culture appeal, they, and most modern accounts, share in uncovering the very non-regional print history of seemingly regionally specific dialect works. Some amount of critical dissent from the Ives-ian "accuracy" approach also began to emerge. These critics didn't necessarily repudiate the notion of written accuracy to phonological features, but instead complicated the picture with more sophisticated notions of decoding and dissent stemming from the standpoint of race and the racialization of certain forms of dialect.[59] Recent critical history sees a decrease in general studies of literary dialect in the United States. Modern linguistics still shows ample interest in the topic, and numerous single author or text studies still enter the field, but literary criticism has to some degree left the notion of the general study of dialect behind.[60] The reason for this can be glimpsed in one of the two most impressive recent general studies of American dialect literature, Gavin Jones's *Strange Talk*.[61] Of

---

"imagined need" of its elite readers for a nostalgic version of American identity. As she aptly demonstrates this occurs during a period with "steadily increasing waves of immigration" making the gatekeeping function twofold. Foote, *Regional Fictions: Culture and Identity in Nineteenth-Century American Literature*, 5.

[59] See especially Holton, *Down Home and Uptown* who approaches the history of dialect literary criticism from both angles.

[60] For example see Minnick as well as Fishkin, *Was Huck Black?*.

[61] The other, Nadia Nurhussein's *Rhetorics of Literacy* breaks new ground by primarily focusing on non-standard poetry and its use in the production of a standard American English in nineteenth and early twentieth century schoolhouses.

dialect literature in general, Jones ultimately concludes the phenomenon was "an ambivalent literary genre in which both radical and conservative motivations were confused."[62] Jones (rightfully, given his parameters) shifts the focus of his study from the linguistic elements of literary works to history, and in doing so produces a conclusion that ultimately denies the notion of a general study of literary dialect whatsoever. Jones's act is a hard one to follow because it is almost undoubtedly correct. The advent of modern archival technologies alongside scholarly recoveries of once-popular writers has delivered unto us the truth — dialect literature simply *is* United States literature. In a nation so fraught with linguistic, political, national, and ethnic insecurities the shadow of dialect flits into almost every literary work of the nineteenth and early twentieth centuries.[63] From a purely historical viewpoint, characterizing the dialect corpus in general would be in some sense be tantamount to characterizing American literature in general, a task easy to shy away from.

Despite the above caveat, this project seeks to open up new space for a general study of dialect literature in the nineteenth and early twentieth century United States

---

[62] Jones, *Strange Talk* 15.

[63] Another way to approach this realization is demonstrated by North in *The Dialect of Modernism* and Baker in *Modernism and the Harlem Renaissance*. Both track African-American dialect's impact on the "high modernism" of the early 20th century. At the same time the popular presses were flooded by dialect-using works produced by authors such as Gene Stratton-Porter, Edna Ferber and Zane Grey. The inundation of the literary field by dialect was even then total, no matter whether its presence was felt directly or second-hand.

through the use of two critical and one methodological intervention. In spite of the very real twists and turns that characterize the history of literary critical work on dialect fiction, one particular assumption, the prevously-explored implicit phonocentrism of "bundling" orthography, grammar and vocabulary, unites them all. Even Jones, writing in the thoroughly post-Derrida world, ends up implicitly accepting this conflation.[64] Considering orthography as a distinct phenomenon invites the literary critic to return to the consideration of style, specifically orthographic substructural style, not in the service of arguing for accuracy to phonology but as a means into a form of literary cognition pervasive yet seemingly invisible. Well before the typewriter, Kittler's "discursive machine gun," cleaved literary expression from the romantic notion of the self, non-standard varieties of orthography and the literary subjects that either consciously or non-consciously produced, received and iterated on them were developing their own notion of the real.[65] The production of an orthographic subject, a form of self with non-conscious access to all the politics and history wrapped up in orthographic processing, glimpsable only in the third person analysis of the fruits of its expression, is what generally characterizes the field. Understanding what these subjects actually say requires a second critical intervention. While dialect in general relies on an opposing some sort of standard to produce meaning, the chaotic and decentralized nature of linguistic standardization in the United States led to individual non-standard

[64] For example, his references on p.p. 4 and 5 include all three elements considered in a bundle.

[65] Kittler, *Gramophone, Film, Typewriter*, 203.

orthographic systems becoming pseudo-standards in their own right. Beyond single authors, beyond single texts, beyond conscious orthographic planning, the precepts encoded in these individual systems battled for position throughout the nineteenth and early twentieth centuries, producing meaning in their conflicts and hybridizations. This form of meaning itself, more inherently creative than grammatical or vocabulary modifications, understood the whole of dialect literature as a field even if the subjects drawing upon it for their own ends did not. In order to comprehend this field this project intervenes by using one last third-person perspective, computation, to unify the historical and media studies approaches explored above. Delving in to the orthographic contours of the self from a from the position of the latter requires the use of massive amounts of context generated by the corpus-based investigations of the former. Generating a sufficient amount of context to be able to make strong claims on the meaning of a particular orthography demands a particularly large corpus, making individual quantification of each text impossible without algorithmic assistance. To that end, the corpus assembled for this project was produced and analyzed with the aid of a suite of self-produced natural language processing tools in the Python programming language. These tools, as well as some discussion on the theoretical issues inherent in exploring a corpus from a third-person algorithmic perspective, will be investigated in depth in the following methodological introduction. Comprehending this field opens up the possibilities of individual texts and authors, granting insight into what their orthographic self thought both they and their characters were. To this end this project concludes with focused case studies that cross from computation into close reading and

history in an attempt to unify understandings of orthography from both the first and third person perspectives.

Methodological Introduction

Substructual Style and Automatic Processing

Accepting that orthographic choice constitutes a sort of substructural style not fully available to introspective first-person investigation comes with an attendant set of methodological anxieties. An account of the orthographic meaning of a particular text that flows from the phenomenological experience of reading, no matter the critical adjunct it is paired with (history, theory) runs the risk of under-specifying, or at worst misrepresenting, its contribution to the meaning of the text at a whole. Minimizing the influence of one's own unconscious orthographic *a prioris* requires a perspective fully outside of first person experience.

Literary studies has historically called upon computation to fulfill this methodological role. Stretching from the concordance work of Fr. Roberto Busa to the modern disciplinary formation of digital humanities, scholars of literature have employed algorithmic means to gain non-anthropic perspectives on their texts of interest. Typically these efforts have been motivated by the desire to either surpass the temporal limits of human textual processing or to replicate it at a larger scale. Modern digital humanities has especially emphasized these advantages, allowing access to, in Franco Moretti's words, "a specific form of knowledge....Shapes, relations, structures. Forms. Models."[66] This "distant" approach to texts has been the most visible of the

[66] Moretti, *Graphs, Maps, Trees*, 1.

computational methodologies employed by recent studies, but it is far from the only one. Studies performed on single texts or small corpora by literary scholars have also proven fruitful, while utilizing an entirely different set of computational tools derived more from linguistics than computer science or statistics.[67] Termed stylometry or computational stylistics, works in this field tend towards a principle of minimal automation, preferring to use more straightforward algorithmic techniques in concert with human effort to draw conclusions over a smaller set of texts. These studies often employ the corpus driven or corpus based approaches discussed earlier and use their generally smaller corpora to generate statistical context.

The structured regularity of orthography makes it a natural phenomenon to investigate while utilizing either of these algorithmic perspectives. However, both also rely on theoretical notions that make neither completely apt for this project. The theoretical objections stemming from linguistics' general subscription to some sort of functionalist view of mind as well as its tendency towards phonocentrism were already expounded upon in the previous introductory section. This project adopts some linguistic methods, most notably the corpus approach, while rejecting these two disciplinary aspects. Despite this, this project resembles a corpus stylistics approach more than a "distant reading" approach. Rather than plunging into the "great unread" armed with a toolkit of relatively sophisticated computational approaches imported from computer science, this project adopts a minimal approach to computation. In part

[67] For example, see Hoover, *Language and Style in The Inheritors*.

this differing attitude stems from differences in the object of study. Distant reading tends to ramble in more traditional semantic domains, employing its means to quantify high-level phenomena like the changes in novelistic word choice and meaning during the course of the British 19th century.[68] Modelling semantic meaning is a more difficult task than analyzing orthographic sequences and requires larger datasets and more intricate computational methods.[69] Semantic-level studies also benefit from access to a cognitively available backdrop that throw their discoveries into relief. Individual words have known commonplace meanings and researchable historical etymologies. Having such background knowledge at hand makes the transit from model to significance relatively straightforward. Showing that the usage of a word changes over time just makes sense to an interlocutor who is familiar with the word on a phenomenal level, much more so than the change in the placement of certain graphemes across two orthographic systems. The benefit of computation in this case is an enhancement of scale and precision. In theory, a reader who read a large section of a nineteenth century

[68] See Heuser and Le-Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels."

[69] Latent Semantic Analysis (LSA) for example. LSA creates a semantic field by assuming that words closer in meaning co-occur in the same types of documents. Not only does this sort of analysis require a massive corpus in order to function properly, it is also theory-rich in its definition of semantics. Given the relative simplicity of sequence analysis, computational approaches to orthography can and should afford to be theory-poor. The information theory based approaches used in this project fit this description, as the notion of entropy they capture is impressively basic and intuitive. See Landauer, Foltz and Laham, "An Introduction to Latent Semantic Analysis".

novel corpus "by eye" could intuit at least some of these changes. The information needed to make this sort of determination is available as a part of readerly experience — if it seems like the word "condescension" is consistently used to mean one thing early in the corpus and a different thing in a later portion then that very may well be the case. A semantic distant reading can confirm or universalize this form of intuition (as not all readers will spot the same shifts) and automate the process that leads to the determination itself, but the type of claim it most naturally offers is different from traditional literary techniques in degree rather than kind.

Orthographic meaning, however, differs in kind. Reading a corpus of texts will not provide a complete account of its constituent orthographic variations and meanings on a granular level. The semi-automatic nature of orthographic processing complicates a reader's ability to develop a sense of meaning from a first-person perspective. An attentive reader of a corpus will certainly be able to provide an account of which orthographies differ from each other across texts and characters, but will likely be unable to provide an accurate account of why these distinctions came to mind. Unlike the semantic information typically studied in distant reading, developing a causal account of orthographic difference would require reading like a computer — keeping a numerical tally of graphemes and their positions, perhaps — a task that is not the traditional form of reading at all. Far from being distant, orthographic meaning is intimately linked to the individual subject in a cognitively holistic way, and prying open the black box that surrounds it requires a different set of techniques. Analyzing orthographic meaning requires developing an account of both the readerly, conscious

experience of the text and the nonconscious yet still historical processing that occurs in a more automatic fashion. The value of combining first and third person methodologies has already been articulated by scholars in the digital humanities. Ted Underwood, for example, encourages literary scholars to view these perspectives not as "competing epistemologies" but instead as "interlocking modes of interpretation that excel at different scales of analysis."[70] Studying orthographic meaning transforms this suggestion into a dictate. Understanding orthography fully requires a combination of techniques that can account for more than just a difference of scale. Orthographic meaning contains multiple components each belonging to its own currently impassable domain of processing. By containing components that differ in kind of processing rather than degree of scope orthographic meaning serves as a sterling example of a literary phenomenon that *requires* separate attention to both its conscious and nonconscious aspects from both the first and third person perspective.

To that end, this project utilizes two main forms of textual processing. The first is simply reading. A large percentage of the texts ingested into the corpus were also read, and any texts that proved interesting computationally were re-read and studied. This technique serves to account for the traditionally semantic first-person component of orthographic meaning. Computational techniques drawn from information theory provide the complementary third-person perspective. As will be detailed below, these approaches are time-tested, simple, and universally appropriate for comparing and

[70] Underwood, "Why Literary Time is Measured in Minutes," 363.

comprehending discrete sequences of data. Each contributes in kind to the readings that follow, and each provides its own insight into what the historical orthographic self of the nineteenth century contains.

Corpus Composition and Processing

The corpus of texts used for this project consists of 157 individual prose fiction works, ranging in publication date from 1794 to 1930. The majority of the works come from the middle and late portions of the nineteenth century. Availability is the primary reason for this particular historical spread. While the corpus would benefit from the addition of texts published earlier in the time range, there is a relatively meager amount of such works available in a machine-readable format. Materials were sourced from a variety of repositories, primarily Project Gutenberg, and were ingested into the corpus as plain text files. A full listing of the corpus texts, as well as their provenance and metadata, is available in appendix one of this work.

However, the natural unit of orthographic investigation is not the text. Any given work could potentially employ multiple non-standard orthographies. Certain texts (say, Marietta Holley's series of *Samantha* stories) have a single storyteller and as a consequence use a single orthographic system. Others (say Mark Twain's novels) employ a variety of orthographies. To account for these cases, it makes sense to separate dialogue potentially written in non-standard orthography by using some *a priori* principle. Ultimately, I elected to separate texts by character/speaker in order to fulfill

this requirement. If an author is using multiple orthographies in the same work it is likely that they are doing so in order to imbue individual speakers with regional, racial or class characteristics, all potential features this project seeks to investigate. Guided by this principle, the corpus in its initial processed form maps a list of speakers in a given text to the segments of dialogue they utter. The lists of characters contained in a given text were generated semi-manually. I created a small Python script that uses a standard named-entity recognition (NER) library to generate a list of candidate character names before combining and pruning them manually. I then tagged each character with one of three gender tags (m,f,n). This process generates tabular data in the form demonstrated by **Table 2-1.**

Even with such a list of characters at hand, manual attribution of dialogue for all the characters in a corpus of this size would be a mammoth task. I performed unautomated attribution of dialogue for 26 texts in the corpus. Based on this experience I estimate the average time for accurate hand annotation of one text to take roughly 4 to 5 hours of uninterrupted work. Given the size of the corpus taking on such a time commitment for every text would quickly prove untenable. To alleviate this burden, I decided to implement an automatic dialogue attribution system. The system was implemented using Python and the SpaCy and NLTK text processing libraries, and uses a deterministic sieve method to pair utterances with their most probable speaker as drawn from the manually generated character tables. Sieve approaches to dialogue attribution are a common solution to this task, and the specific algorithm used here was

drawn from Stephen Bradley's attribution system, named Entmoot.[71] When employing a sieve approach, each portion of dialogue, as demarcated by paired double quotation marks (""), is passed through a series of filters that adjusts the probability of that utterance belonging to a one of the pre-identified speakers based on a specific criterion.[72] These criteria range from the straightforward — explicitly parsing textual dialogue attributions of the form "[character] said" or tracking previous speakers in order to disambiguate the speakers engaged in unattributed dialogue chains — to the relatively complex — parsing previous sections of narration for expressive verbs, determining speaker gender so as to correlate it with the character list, disambiguating pronouns. In this case, the system is termed "deterministic" because it consumes input greedily. If the system recognizes a portion of dialogue in the input text it will inevitably pair it with a speaker from the character list. While this approach does increase the rate of false positive attributions, systems similar to the one employed still report high rates of

[71] See Bradley, "Quotation Parsing and Speaker Attribution in Narrative Texts." Off the shelf tooling for speaker attribution does exist, most notably the attribution system bundled in Stanford's NLP suite. However, at the time of implementation, the system was technically nonfunctional, leading me to simply reimplement Bradley's approach.

[72] Texts that do not use double quotes to indicate dialogue were rectified semi-manually with the aid of an additional Python script.

accuracy when compared to human attribution.[73] However, if a particular text seemed especially fruitful after both attribution and initial computational inspection, I generally elected to re-attribute said text by hand. This accounts for the 26 texts treated in this way. Ultimately, this whole process results in attributed texts, stored in tabulated (.csv) files in the format demonstrated by **Table 2-2**.

Computational Methodology — Stochastic Matrices

The attributed utterances can now be grouped by their assigned speakers, making them available for more specifically targeted computational transformation and inspection. Simply having all of a character's lines of dialogue grouped together is already useful if only because it makes analysis of a particular character's orthography by eye a much more straightforward task. However, a simple collection of utterances falls short of being an actual representation of a given character's orthography. The sum total of their utterances should in theory be a representation of the system of orthography a character employs, rather than a set of emissions from the system. It should represent not only the utterances that have (fictionally) occurred, but also the way a character might structure future utterances. Ultimately the per-character representation I employed is the stochastic, or Markov, matrix. A stochastic matrix

[73] For example, the foundational work of Elson and McKeown. The system detailed in Elson and McKeown, "Automatic Attribution of Quoted Speech in Literary Narrative" technically connects speaker mentions to segments of dialogue but still reports strong accuracy scores.

encodes a discrete series of symbols drawn from a restricted vocabulary of possibilities as a set of transition probabilities. It expresses the likelihood that any one symbol from this vocabulary will be followed by any other particular given symbol from the same vocabulary. In effect, it encodes the probability of a particular future symbol emission given a particular history of symbols. Take the simple example of **Table 2-3**, a matrix encoding the process output "aabacbbaa."

In the case of **Table 2-3** the vocabulary consists of three symbols, "a, b, c". The rows of the matrix indicate the present character, the columns the next potential character to be emitted. This means that the confluence cell of a given row and column provides the probability that the row character will be followed by the given column character. The transition probabilities for each row sum to unity (making this specifically a row stochastic matrix) — the history of the process completely determines the future emissions of the model. The transition probabilities are calculated using a history variable of one. This means that the actual algorithm producing the transition probabilities then only has to count the number of a times a given character follows from a given previous character and then divide this figure by the total amount of times that given character transitions into another. Using the simple example, "a" transitions into "a" two times, "b: one time and "c" one time. This leads to the transition row .5, .25, .25. The stochastic matrix approach necessitates a restricted vocabulary of symbols. A corpus of this size inevitably includes some number of works that employ rare graphemes unused by other works. For the purposes of this project, the grapheme vocabulary was restricted to the lowercase alphabetic characters (a-z), the single space ("

") and the apostrophe/single quote (').[74] In order to accommodate this restriction, usages of upper-case letters were regularized to their lower-case forms. The decision to make this particular restriction relies on the intuition that what matters most to non-standard orthography in this period is the ordering and omission of graphemes and the replacement of graphemes with the apostrophe or single quote. Indeed, a quick survey of the corpus texts reveals that these techniques alone account for much of the orthographic creativity their authors employ.

Applying this approach to the sum total of a particular character's utterances offers a convenient and computable way to summarize the characteristics of their particular orthography. Summary is an important term to emphasize in this case. Orthographies are not simple Markov processes and certainly have causal histories that cannot be captured with a single character of historical information.[75] Reducing character orthographies to this simple form is not in and of itself a form of analysis — it is merely a compression of the surface manifestation of a particular orthographic

[74] The apostrophe will prove useful as a marker of elision, a common element of nonstandard orthography, throughout and will play heavily into the chapter of orthographic extrema. Notably this is an inclusion Shannon does not use in his general calculations of the entropy of printed English -- his 27th character is the space. The elision apostrophe is relegated to the margins, but as will be seen in the "extrema" chapter provides a wealth of useful information. See Shannon, "Prediction and Entropy of Printed English," 54.

[75] For more on the computability of orthography, especially in regards to its ability to compress phonetics, see Sampson, *Writing Systems* and Sproat, *A Computational Theory of Writing Systems*.

system, an account of the regularities some deeper set of rules generates. The transition probabilities of one of these stochastic matrices are not these rules themselves, but rather a useful distortion of them into a simplified form. This distortion would be problematic if one wished to draw conclusions disciplinarily familiar to linguistics or cognitive science. The tight relationship between a model and its actual instantiation *in vivo* often drawn by these discourses renders the use-value of such a model secondary to its empirical explanatory power. Having scientific aspirations (in a modern sense) burdens the orthographic investigator with making some claim on the real — not necessarily that their model implements the orthographic processing of a given subject in a one-to-one fashion but that it has a strong functional similarity to the workings of said subject's neural processing.[76] Oddly enough, conceiving of the relationship between model and the real in such a fashion recapitulates in a more computational form the "top down" troubles of a literary approach to orthography. Some manifest phenomenon (a string of orthographic symbols) offers insight into a phenomenon unavailable to non-theoretical first person inspection, so long as the appropriate method of crossing the first and third person perspectives (with the latter in this case being the instantiation of a neuronal orthographic generator/decoder) is employed. Such a functionalist guarantee ends up producing something akin to an algorithmic version of a hermenuetics of suspicion.

Even beyond the previous theoretical objections raised against the functionalist

[76] This once again refers to the Chomskyean "functionalist guarantee."

approach, the logic of this connection between surface and depth falters when the interaction of fictional works and subjects replace subjects alone as the object of study. As previously explored, the actual workflow of decoding orthography becomes complicated when distinctions in types of cognition are factored in. While a particular nonstandard orthography associated with a particular character might draw on the lower level orthographic processing apparatus of the author in order to generate a readable system, there is enough top-down cognitive interference to frustrate the functionalist guarantee. The flow of encoding does not simply bubble up from the well of deep orthographic structure. Instead, when in the mouth of a fictional character, orthographic structure becomes subject to the diversions generated by the author's theoretical conceits about a given dialect. The resulting orthography has no guarantee of regularity — no guarantee that below the "surface" expression of ordered symbols there is any consistently expressible depth to discover. The minimum quotas of regularity required to achieve intelligibility fall well below what might be termed "consistent structure". In fiction, rules governing character placement can be heavily context reliant, specific, or simply inconsistent. The same character may well encode the same utterance as "hello" "hullo" or "'ullo" as a consequence of nothing more than authorial whim or attention span. No a priori predictors can truly anticipate these orthographic divergences. Even one or two deviations from a lower-level rule can neuter an otherwise accurate model of a system. Preliminary forays into context and history offer little as well. Even if a particular author has a critical reputation for employing "eye dialect," their system (even if it is judged phonologically inaccurate) could just as reasonably be

regular and systematic as it could arbitrary or chaotic. It may follow the "wrong" rules, but employ rules of a sort nonetheless. These practical considerations, as well as the theoretical ones considered in the general introduction, provide justification enough for the use of a surface-centric approach implied by the use of stochastic matrices as an analytical tool. In the particular case of literary non-standard orthography surface expression is, potentially, all there is be found.

Computational Methodology — Methods of Comparison

Abandoning latent structure in favor of surface expression threatens the very significance of orthographic meaning. Cleaving orthography from direct relationship with a deep structure instantiated in the physiology of an actual subject demands methodological change lest orthography be rendered meaningless. One solution, the one chosen by this project, is go to wide rather than deep. Fictional orthography may not be able tell an uncomplicated story about what is happening on a lower neural level, but it can shed light on a historical system of meaning distributed across many works and how one particular work relates to that system. These comparisons are also computationally achievable. The relationship of orthography and phonology might, per Sproat and Sampson, be a computationally "hard" problem, but the comparison of surface-level sequences of graphemes has any number of solutions.

Viewed from a distance, orthography is merely a sequence of distinct events. What counts is not whether this sequence performs any particular teleological function,

in this case encoding language, but that it is composed of a number of discrete elements drawn from a finite vocabulary expressed in an unambiguous order. Computer science and its related fields offer numerous ways to compare such sequences. Each comes with attendant advantages and disadvantages, both material — some can handle the comparison of sequences with greatly differing lengths and vocabularies, some trade accuracy for speed, some require large amounts of storage space — as well as theoretical. The primary theoretical consideration at play when calculating sequence similarity is the distinction between a distance comparison and a divergence comparison. Divergence is a mathematically weaker notion than difference, meaning that it has fewer governing axioms. Most importantly, divergence measures of similarity do not preserve the axiom of symmetry. Comparing the same two sequences could result in different similarity scores depending on which sequence is deemed the baseline and which the comparator. Distance measures do preserve symmetry—comparing any two given sequences will always return the same similarity score. On the face of things distance measurements seem like the most appropriate tool to use when comparing orthographies. Considering two different orthographies to be similarly different no matter their status as baseline or comparator accords with human intuition. If one is to be different from the other, surely it must be the same difference. However, as Kent Chang and Simon DeDeo demonstrate, divergence can capture important relationships between texts that are invisible to distance measures.[77] Chang and DeDeo point to enclosure as the principle relationship

[77] Chang and DeDeo, "Divergence and the Complexity of Difference in Text and Culture."

divergence can detect. When one text or orthographic sequence encloses another it is a superset of the comparator sequence, meaning that it contains all of the information the comparator sequence provides but also has additional regularities the comparator sequence does not. For example, take two orthographies, one that regularly uses an elision in place of a terminal 't' (e.g. bes') and one that applies an additional rule eliding 't' before a certain set of vowel characters in medial positions (e.g. tes'ed). A divergence measure of difference would capture this relationship by returning one value for the comparison of the two-ruled orthography with the one-ruled, and another with the positions switched, showing that one diverges more from their shared ground than the other.

In part, this discussion serves to introduce this project's approach to comparing orthographies. In order to capture different notions of similarity this project employed three different methods of comparison, each detailed below. The measures employed are relatively straightforward, and mostly couched in information theory. Non-information theory approaches also abound, but were discarded for mostly practical reasons.[78] These methods serve as important starting points, but they do not exhaust the complexities of the orthographies studied here. Indeed, each method compares sequences on an at least somewhat distinct theoretical basis, resulting in comparisons that prioritize a certain set of features over others. The differences these

[78] For example, an edit distance approach that compares sequences by how many steps it would require to transform the first sequence into the second may well be possible but would require a large amount of computation and some severe modification to fit this particular task.

algorithms have make a very real difference when it comes to which orthographies they judge as similar or dissimilar. Ultimately, they only become useful to the literary critic when contextualized under a specific theory of *why* two orthographies may be deemed the same. In this project, each serves the role as a guide to further investigation, rather than as a final conclusions themselves.

Principle Component Analysis (PCA) is a method that reduces high-dimensional data to a smaller dimensionality in order to make it more tractable. Such a reduction is commonly employed when feeding high-dimensional data into a further computational system or in order to visualize it on a standard two dimensional plot. Stochastic matrices are a paradigmatic example of high dimensional data. Each matrix can be thought of a series of n vectors of length n where n is the length of the vocabulary of output characters modeled by the matrix (in this case the lowercase alphabetic characters, the space, and the single quote). These vectors imply a concomitant n dimensional space. PCA uses some convenient properties of the eigenvalues and eigenvectors of these matrices to reduce this high-dimensional space to one of a more convenient dimensionality, in most cases the familiar cartesian 2D plane. In transforming the data, PCA attempts to preserve as much of the variance of the higher dimensional space as possible, with the practical result being that vectors that are close in distance in the higher dimensional space will also be close in the lower dimensional one.

Put more plainly, PCA reduces the complexity of numerical data while trying to retain its most defining characteristics. The algorithm takes the original matrix apart

into its component pieces in order to infer what it might look like as something with only 2 or 3 data points per entry instead of many. For the purposes of this project PCA was used as a way to tune and challenge intuitions about orthographic similarity between characters. Rather than reducing the dimensions of each single character's stochastic matrix in order to understand the similarity between the distribution of graphemes in that character's orthographic distribution, each character's matrix was transformed into a 1D vector and PCA was performed across the set of all characters. This method reduces the entirety of a character's grapheme distribution (of length n x n) to a Cartesian ordered pair. The final result is a data-set of points in the familiar (x,y) format that can be plotted on a normal 2d chart. Using PCA as a true measure of difference would require the additional step of analyzing the newly-inferred 2D points using some sort of distance measurement, perhaps the familiar Euclidean distance calculated using the Pythagorean theorem. However, the availability of difference measures that operate on the original higher-dimensioned data makes this step somewhat redundant. 2D data is also considered more appropriate as input to computational clustering algorithms that attempt to group data points by their relative similarity to the rest of the available points. These algorithms commonly use distance measures as part of the grouping process, partitioning data points into groups through an iterative process of clustering, recomparing distance, and regrouping. However, such algorithms often require the user to provide an *a priori* number of groups into which to gather the data. Given this project's investment in a corpus-driven approach to orthography, one that can reveal truths about orthographic similarity that might go

otherwise unseen, this approach was avoided in favor of simple visual inspection of PCA charts. The definitiveness of clustering algorithms is welcome when they are used as the first step in a further computational process, like training a neural network. However, since human intervention will eventually be needed to interpret the meaning of any orthographic similarities in any case, it makes more sense for the purposes of this project to respect the potential fuzziness of any group delineations by simply arguing for groupings individually when they prove useful or interesting.

Reading such a chart for the evidence of similarity is intuitive. **Figure 2-1** is an example of such a chart, where each point is the stochastic matrix of one character's orthography, reduced to a 2D point. Characters that are most similar in orthographic style reside in the same spaces of the chart. Already this chart reveals that narrators tend to gather at the middle-right hand side of the chart. Weaker clusterings also occur on a text or author basis — their characters tend to talk with some similarity. These results are not surprising, but they are useful. The emergence of these regularities makes the outlying elements, ones that will be discussed in the further chapters of this project, truly significant. Clearly, this straightforward visual approach is capturing some notion of orthographic closeness, making unexpected dissimilarities or similarities worthy of critical investigation. What this PCA chart does not reveal is *why* these orthographies have clustered the way they have. Since PCA's sole aim in reducing the dimensions of data is to try and retain a sense of the greatest variance between the individual data-points of a given set the actual X and Y axes of the graph mean comparatively little. Location in an unexpected cluster on a PCA chart provides enough reason to start

investigating a particular character in a particular text, but it is up to finer-grained computational methods and human interpretation to determine why such a clustering occurs.

Information theory provides a wealth of sequence comparison methodologies useful for the investigation of text. Information theory as promulgated by Claude Shannon in the early twentieth century deals primarily with characterizing the probability distribution of a given source of sequential information. Orthographic data, even just its surface expression, fits this description nicely. Some underlying process produces a stream of graphemes that can be characterized as a probability distribution over past information from the stream. This, as previously discussed, is the composition of the stochastic matrices used as the basis for the computational comparisons to follow. Information theory is especially interested in a measure Shannon termed (somewhat confusingly) entropy. Entropy is often described as a unit of "surprisal." It quantifies, in some sense, how predictable a random variable is given the various probabilities of its possible outcomes. If the variable tends heavily towards one outcome it is predictable, and considered low entropy. Resolving the event provides little new information — it reinforces what we knew about the probability distribution. If it has many nearly equally likely possible outcomes it is less predictable, and thus high entropy and more information-rich. Applied to orthography, the question becomes "how predictable is the next character in this sequence?". Some orthographies might have a few equally likely characters that follow, for example, an "r", while others might almost always follow an "r" with one particular character. Differences such as these characterize what makes one

orthography distinct from another.

Table 2-4 is a constructed example giving the entropy measures of three different probability distributions. For the purposes of this example, we can consider the probability distributions to each be a row from the stochastic matrix of three different different orthographies, each drawing from the same ten grapheme vocabulary. As with the rows of this project's empirically derived stochastic matrices, these sample distributions give that probability that one particular character from the grapheme vocabulary, perhaps 'a', will predict each of the other ten. In the first orthography, 'a' leads to only four other graphemes, each as likely to occur as the other. The second orthography has two possible successors, one quite likely and the other far less so. Finally, the last orthography is the most diverse, distributing the potential graphematic outcomes widely, if unevenly. Applying Shannon's entropy formula to each returns results of 2.0, .47 and 2.72, respectively. The first and last examples are relatively high entropy. Though the first is more uniform than the last, they are each less predictable in their own specific way. The first distribution has only four possibilities, but each is equally likely, making guessing the next grapheme a relatively long 25% shot. The last has more possibilities, but the likelihood of one appearing instead of any other is separated by relatively few percentage points. The second example, on the other hand, is quite predictable. There are only two outcomes, and one has a dominant presence. When parsing an orthography that uses this distribution, one would expect to see the 90% successor grapheme most of the time, making the distribution low entropy.

True to his nature as an engineer, Shannon originally conceived of entropy for

practical purposes. In his original interpretation entropy characterizes the amount of information, on average, that will be needed to represent the potential emanations of a given source — the perfect measure for designing efficient and redundant electronic signaling and coding systems for use on telegraph networks.[79] If the entropy of a distribution is calculated using a base two logarithm (as this project employs) the resultant entropy score is how many bits, binary yes/no digits, on average, one would need to encode an element of a sequence generated by one of these sources. This is especially intuitive in the case of the first example. Four equally likely outcomes could be encoded in a simple 2 bit-based scheme — 00 for the first outcome, 10 for the second, 01 for the third and 11 for the fourth.

Later mathematicians realized that entropy could take on other interpretations and began expounding further measures that have practical uses in a number of fields. These include the measure at hand, the Kullback-Leibler divergence (KLD). The KLD calculates the *relative entropy* of two probability distributions. Distributions with a relative entropy approaching 0 are more similar, with 0 indicating that the distributions are functionally the same. Put intuitively, the KLD interprets the entropy of a given probability distribution as a key characterization of the sorts of events that distribution might emit. Two distributions that are more likely to predict similar events are thus more similar than two distributions that are more likely to predict dissimilar events. This intuition crosses nicely into the orthographic domain. If the probability distributions (as encoded in a stochastic matrix) of two orthographies tend to predict

[79] See Shannon, "A Mathematical Theory of Communication."

similar graphematic emissions given the same prior grapheme it makes sense to call them similar. This entropic correspondence holds for a number of other measures less complex than the KLD. For example, it is not hard to imagine ranking orthographies solely by their average entropy, the average amount of uncertainty in predicting the next character of a sequence given a history of a particular grapheme. A more chaotic orthography where a given grapheme could be followed by any number of possible successor graphemes would have high entropy and one where particular graphemes tend to consistently predict a smaller set of successors would have low entropy. This approach has uses, but also an in-built limitation. In order for an orthography, be it considered standard or nonstandard, to function at all, it must fall within a certain entropic range. Even more eccentric orthographies will only vary slightly from the entropy Shannon himself calculated for standard American English orthography.[80] One could make an argument about particular orthographies being more or less random, but the very minute differences themselves could not tell the whole story. Understanding these less complicated measures does, however, justify the use of the KLD. The KLD builds on the intuition that a given orthography can be more or less entropic, but specifies to some degree the actual *ways* an orthography is more or less entropic through the use of comparison. Determining that two orthographies are similar using the KLD not only means that they tend towards similar patterns of characters, it also provides a natural context for understanding how both operate. This effect magnifies

[80] See Shannon "Prediction and Entropy of Printed English."

when multiple orthographies are compared in concert. As with the PCA approach, clusters containing the orthogrpahies that compare similarly to the rest of the field (i.e. they generate high relative entropy scores when compared to one set of other orthographies and low when compared to another) emerge. Unlike PCA, these comparisons are straightforwardly meaningful. Orthographic clusters calculated by KLD tend to produce the same sequences — a starting point easily transposable to more traditional literary criticism.

Justifying the use of the KLD for linguistic comparison and characterization need not rely on faith (and theoretical argument) alone. Numerous scholars, primarily in linguistics, have employed relative entropy produced through KLD for these ends. For example, Yuri Bizzoni, Peter Fankhouser, Stefania Degaetano-Ortlieb and Elke Teich use KLD paired with a clever "bin" style sampling system to characterize linguistic change over long periods of time.[81] They also provide an extensive list of projects that have used similar information methods, applying KLD or related measures to corpora ranging from Google Books to collections of scientific writing.[82] Almost all of these projects add computational steps to make the KLD more appropriate for their particular task. These steps implicitly answer the questions "what orthographies are being compared?" and "why?". For Bizzoni et. al., this means finding a way both to implement

[81] See Bizzoni, Degaetano-Ortlieb, Fankhauser and Teich, "Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach."

[82] Bizzoni, Degaetano-Ortlieb, Fankhauser and Teich, 2.

and to justify a system that places texts into "bins" based on predetermined chronological periods. They then compare the grammatical features of the texts in each bin with the other chronologically distinct bins in order to generate a theory of linguistic change. Using KLD for the specific purpose of comparing the orthographies of individual characters requires similar modification and justification. As previously noted, the decision to make the base orthographic unit the character is itself a theoretical decision that requires some justification as well as its own technical implementation. Somewhat at odds with the functioning of KLD, which compares two probability distributions, the stochastic matrices that result from this initial process are not a sole probability distribution, but many, one for each possible grapheme in the orthography. Comparing two stochastic matrices thus requires some sort of compound measure. The compound measure used in this project is quite simple. When comparing two stochastic matrices with KLD each grapheme probability distribution was compared with the corresponding grapheme probability distribution of the comparator matrix. The distribution for 'a' was compared to the other matrices distribution for 'a', 'b' for 'b', and so on. The resultant set of KLD generated relative entropies where then averaged to get an average similarity score for the two matrices. KLD also proves useful as a more finer-grained measure of similarity. Rather than solely being a hindrance, KLD's one-distribution domain allows for the further specification of orthographic similarities. After using one of the broader methods described above (PCA, average relative entropy via KLD) to identify a broad similarity of two orthographies KLD can be employed to compare their per-grapheme distributions in order to find where the similarity

specifically lies. KLD is also a divergence rather than a distance, meaning that it can reveal encapsulation relations as well as absolute difference (using a modified symmetrical version). Ultimately this fine-grained fashion is how this project mostly employed KLD. Having such a specific tool is quite useful when inspecting orthography. The usage of particular graphemes, especially the apostrophe, directs a surprising amount of the meaning orthography provides. KLD provides access to these detailed orthographic aspects leaving the task of determining broader similarities to other measures.

A different information theory measure, termed perplexity, proved more useful for finding broad similarities and distinction between orthographies. Perplexity measures how aptly a probabilistic model fits a given sequence. Unlike KLD, it is less used for the comparison of two probabilistic models, and more as a way of determining how likely it is that a model could produce a particular sample of data. In this aspect, perplexity is often employed to test how a machine learning model has inferred the regularities of a targeted textual feature. The usage of perplexity in this project extends this common usage to comparison. Rather than testing whether an inferred model fits the data it was trained on, perplexity was used to compare the stochastic matrix generated from the orthography of one character with the actual grapheme sequences attributed to the other characters in the corpus. In effect this tests how well the model of a particular character fits the sequences produced by all of the other characters, making the resulting perplexity score a measure of similarity. Typically this approach is used on models that use words as their unit of meaning, in this case the base unit is the

individual bigram (two grapheme sequence). Perplexity stands on the same foundational ground as the KLD — information entropy. However, in practical testing it performed in a different fashion than the KLD based ad-hoc average relative entropy measure described above. It also preserves the advantages of KLD over visual inspection of charts generated by PCA, in that it retains meaning during comparison. Also like KLD, perplexity holds a solid pedigree as a tool for the use of studying language. Beyond the model testing function it is often employed for, perplexity has been used for diachronic linguistic investigations, much in the vein of the KLD-based study cited earlier.[83]

Ultimately each of these measures only serves as a point of entry to the orthographic meaning of a text. Establishing the significance of any of these measures of difference requires additional critical intervention. When used in a computational or statistical context, the simple term "significance" is loaded with a very specific valence of meaning. In these fields the term almost invariably refers to "statistical significance," a process or test that establishes the non-randomness of a particular result. There are many such tests, and most function in fashion strikingly similar to the comparison measures described above. These particular tests use a known statistical distribution as a point of comparison in order to establish that the results of a computational analysis were not likely to have been generated by it. A purely random distribution, say the binomial distribution modeling the 50/50 flips of a fair coin, is a common comparator

[83] For example see Gamallo, Pichel and Alegria, "From Language Identification to Language Distance."

distribution. These methods argue that by showing that it was unlikely that such a random distribution could have produced the results of a computational or statistical test the phenomenon under investigation is itself unlikely to be an artifact of the process itself or random noise. Tests such as these are the genesis of the "p-values" often cited by scientific studies. A small p-value indicates more confidence that the process under computational description is not purely random (as represented by the null hypothesis, the comparator random distribution), implying that said process or phenomenon has some sort of real-world causal "oomph" of its own. This technical sense of significance needs no defense, and it is not the intent of this project to debate its general usefulness. However, for the purposes of literary criticism, and especially this project, this definition of significance proves both too strong and too weak. Literary scholarship establishes the significance of a textual feature in a variety of ways — close reading, historical analysis, theory, to name just three. All locate causality on a different plane from the text itself. A feature occurs in a text not because of its internal organization, but due to some exterior context that itself needs clarificatory attention. Statistical significance does not speak to this level of meaning. For literary studies the significance-establishing comparator of relevance is not a null hypothesis of total randomness, but some condition outside of the process itself made knowable through methodologically appropriate means. A literary feature may appear to be an outlier, or simply random, but also posses justification for causal significance through historical research or theoretical consideration. Establishing statistical significance alone does not entail literary significance, and proving statistical insignificance does not supersede the possibility of literary significance.

At the same time, care must be taken when traveling the computational world. Taking this contentious line towards traditional measures of statistical significance does not absolve oneself from other pitfalls associated with such measures. Most importantly, one must remain attuned to the particular sensitivities of the algorithmic methods they employ. In this case, the stochastic matrix models were generated from orthographic sequences with potentially very different lengths. Some characters are simply more loquacious than others and have more dialogue across the course of a given text. Less dialogue means these characters have less chance to use a variety of unique words and thus utter a given orthographic bigram, lessening the power of their model. This is an especially important consideration when dealing with the perplexity measure, but has impact on both. When dealing with broad strokes, this factor is a relatively minor issue. For example, **Table 2-5** collects the closest character matches (by perplexity difference) to Topsy from Harriet Beecher Stowe's *Uncle Tom's Cabin*.

Despite Topsy's relatively small model length of 3866, the effect of being compared to a model built from a large amount of graphemes (say a Narrator model, which can be built out of hundreds of thousands or even millions of graphemes) does not eclipse the strong amount of similarity she has when compared to a particular character with a smaller model length (say, Dinah with her length of 3717). While having more textual space allows a given character more possibility to present a given grapheme bigram, the nature of orthography means that some of these characters may still never or very seldomly utter the sort of bigram a different character uses quite frequently, making them quite dissimilar in one sense. From this example, as well as

empirical inspection of many others, it can be seen that this form of modelling captures this notion well.

The effect of differing lengths is felt more when attempting fine-grained comparisons. In this event, it can be mitigated a number of ways. The first is to use multiple algorithmic models when comparing speakers. Each can capture subtly different information, and analyzing their results through comparison can help control for this factor. The second is to seek outlier comparisons that occur among characters with relatively similar model input lengths. If a model with a shorter length returns as very similar to a particular character and is surrounded by models with longer lengths, there is likely something about that model that makes it a particularly good fit. This might again be seen in the example of Topsy's comparison to Dinah, where despite the relative brevity of Dinah's model compared to the other top matches she still ranks second overall. Finally, and most importantly, these computational results can be manually inspected and simply read, allowing for more traditionally literary arguments for significance.

Orthographic literary meaning does not deviate far from the typical functioning of such literary phenomena. The regularities and restrictions inherent in orthographic patterns tempt the conclusion that they should be analyzed on the surface level alone, or at least seen as solely the emanation of some discoverable lower-level cognitive function. However, as previously demonstrated, the actual situation of orthographic meaning is far more complex, and far more similar to the semantic textual features literary scholars typically investigate. In this case, as with literary studies in general, justification must

emerge from contextualized argument instead of statistical tests. This project, somewhat polemically, thus makes no use of statistical tests, instead relying on algorithmic approaches as a means of discovery and more traditional critical argumentation as a means of justification. The methodology I employed when investigating particular texts or characters reflects this approach. Algorithmically interesting results are treated not as an end in themselves but as an invitation to a process of re-discovery. If a particular text or character returned interesting algorithmic hallmarks it was re-read and the dialogue re-attributed manually. The orthographic specifics were reexamined, both computationally and manually, and contextualized within the field of the other texts found in the corpus as well as the broader realm of literary history. This approach does justice to the complexities of orthographic meaning, respecting its status as something both automatic and historical.

Figures

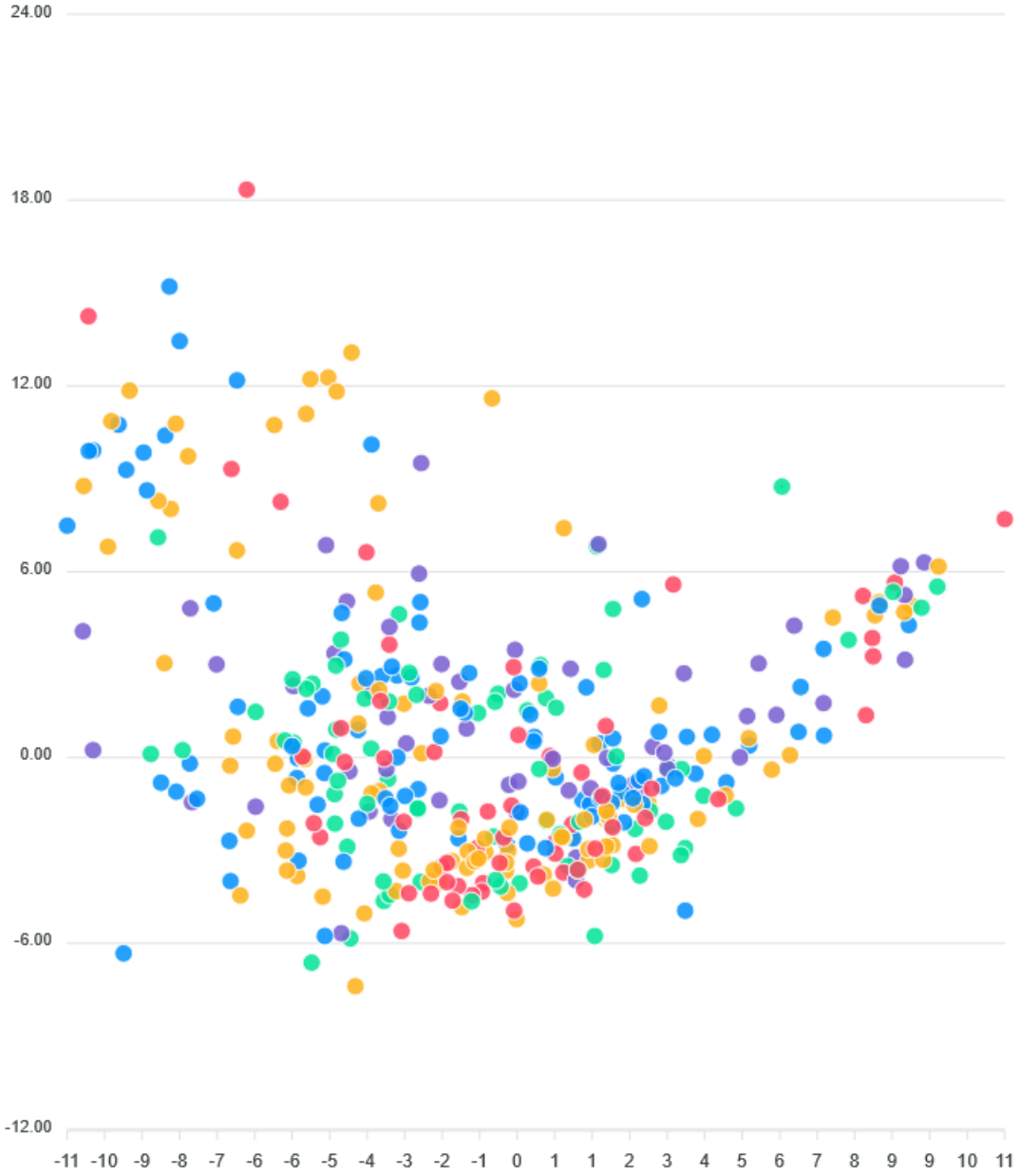

Figure 2-1. A scatter plot demonstrating the clustering properties of principle component analysis (PCA). Each point represents the stochastic matrix model of a particular character transformed into a two dimensional euclidean ordered pair by PCA. The plot is filtered to only include texts (and thus characters) from 1880-1890.

Tables

Character	Gender	Aliases
Grace Blackiston	f	Grace, Blackiston, Grace Blackiston
Matty Scamper	f	Matty, Scamper, Matty Scamper

Table 2-1. Two example entries of the tabular format used to disambiguate character entities. Each table is associated with a unique corpus text, and each entry records the primary name, perceived gender, and name aliases associated with a particular unique character.

Line	Attribution
No, dear, him I speak of could never think of me,	Todd

Table 2-2. An example entry of attributed text. Quotation-delimited utterances in the first column are paired with their speaker's main alias in the second. Attributions were performed both automatically and by hand. This sample is drawn from *The Country of the Pointed Firs* by Sarah Orne Jewett.

	a	**b**	**c**
a	.5	.25	.25
b	.66	.33	0
c	0	1	0

Table 2-3. An example stochastic matrix for the process output "aabacbbaa.” Each row is an individual probability distribution representing the likelihood that a particular given grapheme drawn from the vocabulary “a, b ,c” will be succeeded by another particular grapheme from the same vocabulary. Therefore, each row sums to unity (1).

0	.25	.25	0	0	.25	0	0	.25	0
0	.9	0	0	0	0	.1	0	0	0
.12	.03	.33	.12	.09	.13	0	0	.11	.07

Table 2-4. Rows drawn from 3 hypothetical stochastic matrices that encode 3 different orthographies that use the same 10 grapheme vocabulary.

Model 1 Text	Model 1 Character	Model 2 Text	Model 2 Character	Perplexity Difference	Model 1 Length	Model 2 Length
Uncle Tom's Cabin	Topsy	Uncle Tom's Cabin	Tom	12.22	3866	20029
Uncle Tom's Cabin	Topsy	Uncle Tom's Cabin	Chloe	12.31	3866	14824
Uncle Tom's Cabin	Topsy	Uncle Tom's Cabin	OldDinah	12.37	3866	3717
Uncle Tom's Cabin	Topsy	Jerome, A Poor Man	Jerome Edwards	12.47	3866	48367

Table 2-5. The closest character matches (by perplexity difference) to Topsy from Harriet Beecher Stowe's *Uncle Tom's Cabin*.

Twain's Orthography in Context

Despite the national referent of its title, Mark Twain's *The American Claimant* opens on a British scene. Before leaping off to the indicated claimant's own American shores the narrator takes time to introduce us to the lead of the novel's B plot, the "Honourable Kirkcudbright Llanover Marjoribanks Sellers Viscount-Berkeley, of Cholmondeley Castle, Warwickshire" son and heir of the Earl of Rossmore.[84] Lest his reader neglect the subtleties of British received pronunciation, the narrator quickly glosses this mouthful of a title with a pseudo-phonetic pronunciation guide, admonishing his readers to voice these honorifics as "K'koobry Thlanover Marshbanks Sellers Vycount Barkly, of Chumly Castle, Warrikshr." On first glance this joke on orthography seems to hinge largely on the amusing juxtaposition of what an American readership would likely recognize as the standard pronunciation of words like "Kirkcudbright" and the much altered British version of "K'koobry." In this reading, the narrator invites the reader to laugh at the illogical pairings of phoneme and grapheme that run rampant through British English, in turn implying that the reader's own (again, likely "standard" American) pairings of sound and letter offer a more sensible approach. Twain's joke is reminiscent of the Jamesian pronouncement of dialect by fiat examined in the general introduction, but lands the punchline after an additional beat.[85] When

[84] Twain, *The American Claimant*, 465

[85] To jog the memory: James simply declares that Basil Ransom uses phonetic elements consistent with southern speech rather than rendering them. The recurrence of this sort of joke throughout this project's

read through a minute orthographic lens, the target of the joke shifts — compared to some sort of idealized orthography that pairs each letter, or grapheme, with a sound, neither of the offered renderings makes perfect sense. The pronunciation guide offers no additional gloss for "Sellers," a word now suddenly under question given the divergence of "Kirkcudbright," and "Viscount" merely becomes "vycount,"" the fairly sensible orthographic reform of "y" for long "i," a substitution long considered by the spelling reform movement in the United States.[86] Even in what seems like a phonetic rendering, Twain provides little systematicity.

Twain uses this little joke to draw attention to not only the constructed nature of orthographic interpretation, but also the explicitly textbound life these conventional pairings of sound and letter lead. No system, British or American, seems adequate to describe the sound patterns presented, and even the ad hoc phonetic orthography he presents contains encodings drawn from other non-phonetic systems, such as American spelling reform, in a varied and not quite consistent fashion. Twain's aside demonstrates the playful aspect of orthographic uncertainty, part of what Sebba terms the independent meaning system of orthography and what this project refers to as

---

corpus tempts a title change (perhaps "Orthography: A Linguistic Joke") but is also revealing in its own way. For one, these jokes do not involve changes of syntax. If humor relies in part on the flexibility of meaning, this is further proof of syntax's relative inability to sustain its own form of meaning. Secondly, they indicate some amount of understanding of how orthography functions. This understanding itself could be only semi-conscious — it is easy to "feel" the limits of orthographic creativity.

[86] See Lepore, *A Is for American*, 37.

substructural meaning.[87] Though orthographies can help guide pronunciation, they do so in a manner that allows for novel invention and play while still attempting to fill their communicative role. As previously explored in the general introduction, the exact way these meanings express and the systems that undergird their interpretation are at least in part nonconscious and thus potentially obscure.

Twain could have rendered his pseudo-phonetic burlesque on British received pronunciation in any number of orthographic ways simply because he explicitly marks the section as a moment of high-class British speech through the surrounding setting. Even though the additional orthographic may provide additional assistance, Twain, like James, still shifts the burden of actual phonological interpretation to the reader, essentially instructing them to "read this however you imagine a British lord would." Also akin to James, this not-so-simple joke hides an incisive critical insight. Twain exploits the flexible meaning-making of orthography to direct two very different graphematic sequences towards the same end. Both say the same thing (again, read this like a British lord) without actually saying the same thing. Their orthographic meanings, as Twain seems to know or intuit, don't actually rely fully on the composition of the sequences themselves. Composing such a non-standard orthography (or just an orthographic joke wrapped around "standard" orthography) necessitates relying on one's own orthographic notions. The combination of top down phenomonological interrogation of one's own lower centers of orthographic processing and an priori

[87] Sebba, *Spelling and Society* 30.

rationalized invention of grapheme systems that ensues enters the world without an explicit history. Their meanings are commingled, an amalgamation of the nonconscious decoding process and some particular phenomenally-composed aim. Such cases necessarily require a set of techniques different from the sociological analyses provided by researchers like Sebba, as the context they reveal has been forged privately in the crucible of an individual mind. They are, in a word, literary.

This chapter adopts the above view of orthographic meaning in order to investigate Mark Twain's use of dialect in his novelistic fiction. Twain's prolific employment of non-standard orthographies across numerous works, characters and genres has long made his particular implementation of these deviations a critical target. Partly this is due to Twain's prominent position in the canon. He remains one of the most famous writers to ever use non-standard orthography, and the interpretation of what he means by doing so holds a large amount of potential import for the literary canon at large. This is augmented by a sense that Twain knows *something* about dialect and non-standard orthography. His famous claim in an introductory note to *Adventures of Huckleberry Finn* stating that he employed numerous dialects and "modified varieties" of dialects to compose the novel implies that he has spent no little amount of time composing the systems he then deploys.[88] While these systems may, as he claims, rely in part on "personal familiarity" with several forms of speech, it is hard to believe he

[88] See the explanatory note to Twain, *Adventures of Huckleberry Finn*. Twain also takes a shot at linguistic standardization, claiming that the note exists to keep readers from thinking the characters were in fact meant to talk alike. He seems clearly in the corner of orthographic creativity.

is being entirely serious. This, in combination with his *Claimant* joke indicates that his knowledge may extend beyond expertise. Twain hints at an understanding of orthographic substructural style — its creative potential and its technological nature — even as he makes face-level claims of accuracy.

Twain's lifelong obsession with print which began with his youthful occupation as a printer's devil and continued through his financing of novel print apparatuses granted him the perfect vantage point to glimpse orthography's technological nature.[89] His interests in print technology and literary composition most apparently commingle in, as might be expected, another joke. In a letter to the Remington company, Twain complains about the downsides of his early adoption of the newly-invented typewriter. He claims to have "entirely stopped" using the device as any letter composed by its hammer inevitably elicits a response demanding to know more about the machine itself.[90] The letter itself is, naturally, typewritten. Again, a simple joke, but also the understanding of a more subtle theoretical point. Twain once again uses orthography to impart his humor, but this time he uses the very shape of the graphemes to convey the punchline. In order for the joke to work, the letter must recognizably bear the markings of typescript from a typewriter rather than handwriting or print. This joke is only possible after the invention of the typewriter itself. Its new way of technologizing language, subtly distinct from those that precede it, creates a new form of meaningful

[89] See Michelson, *Printer's Devil: Mark Twain and the American Publishing Revolution*.

[90] Cited in Kittler, *Gramophone, Film, Typewriter*, 192.

orthographic expression simply due to its distinctiveness. Simply in the shape and texture of its graphemes, it "means" novelty and technology, and thus the associated attitudes (interest for interest's sake) his readers have towards such innovations. This orthographic meaning then conflicts with the semantic meaning contained in Twain's words. Twain claims to have stopped using the typewriter; but by using it not only indicates his approval of the machine, but also incisively jokes about his own American passion for technological novelty — he simply can't help himself but use it. These two planes of meaning contradict each other without full resolution, playing in the space opened up by this being a letter to the manufacturers themselves and thus possibly an exception to his general rule of typographical abstinence. Dissecting this joke may, as the saying goes, also kill its humor. However, doing so a crack in the black box of orthographic meaning, demonstrating its constructed, social meaning in a way that fundamentally impacts the possibilities it presents even when writing in a less novel medium.[91] The *Claimant* joke, despite relying on orthographic sequence rather than

[91] Type technology acts on Twain in the way Peter Galison aruges time-standardization and clock-making acts on Alfred Einstein during his development of the theory of relativity. Galison points out that Einstein's predecessor Henri Poincare had all of the mathematics required to abandon the theory of the lumineferous ether in favor of one (like relativity) that does not require a pervasive, invisible medium. Poincare, however, belonged to a different technological context where spatial notions dominated understanding, rendering him unable to make the jump. The de-naturalization of the axioms provided by one technological regime leads to both a different form of understanding and the meta-level

grapheme appearance, operates in the same fashion. Both cleave the pairing of linguistic and orthographic meaning through technological intervention in manner that exposes how the two discourses can be produce as much or more meaning in disharmony as they do when united.

As I will detail below, previous studies, though useful and insightful, have typically underestimated the impact of this sort of specifically orthographic realization. Twain's dialect usage is typically approached from a holistic perspective. Orthography, syntax and vocabulary are blended in a fashion that, as argued in the general introduction, does a disservice to the primacy of orthographic meaning. This leads some studies to fall into the trap of phonological accuracy. Too much focus on this one aspect of orthography renders Twain's specific use of its meaning-making potential invisible. This chapter approaches Twain's writings with the computational corpus approach used throughout the project in order to individuate and contextualize Twain's orthographic meaning. Combining corpus and literary techniques does the complex reality of Twain's orthographic compositions justice. Previous scholars who have already taken the step of moving past accuracy and into meaning provide abundant historical and critical insight into Twain's dialect work, but frequently do so at the expense of grappling with the complicated situation of composition described in the general introduction. Twain's

---

understanding of the historical nature of the process itself — something Twain seems to grasp at least in part when it comes to print. See Galison, *Einstein's Clocks and Poincare's Maps: Empires of Time*.

reputation for care in crafting his dialects does not exempt him from the general situation of orthographic decoding any writer must face.

Figure 3-1 is a PCA chart of the models generated from the Twain texts included in the corpus. All of these texts were attributed by hand, and consist of the majority of Twain's novelistic output. The chart reveals the sorts of high level consistencies one would hope for both from Twain's orthographic practice and the technique itself. The far right area of the chart contains mostly narrators, excluding those who speak in non-standard tongues (most notably, of course, Huck). Clusters of characters with largely standard orthographies as well as those that employ archaic forms of English sit to the left of these, roughly in the center of the diagram. Finally characters that employ Twain's version of African-American speech group in the upper left, with those that utilize more "backwoods" dialects below them.

Twain is not just consistent — he is remarkably so. The four points marked on the chart above are the four orthographic instantiations of Tom Sawyer. These four versions of Tom appear in four separate texts, the earliest being *The Adventures of Tom Sawyer*, published 1876, and the latest being *Tom Sawyer, Detective*, published 1896. Despite the twenty year range (though Twain was often composing these texts well before their publication dates) Tom's orthography remains remarkably consistent, with only the *Adventures of Tom Sawyer* version of Tom deviating to any real degree. For context, **Figure 3-2** highlights the orthographies of Natty Bumppo (in his various

guises) a character with a very similar distribution of appearances across the run of James Fenimore Cooper's *Leatherstocking Tale*.

If orthographic consistency is a desirable trait in a multi-narrative protagonist, Tom edges Natty out by a reasonable margin. Again, this is evidence that Twain seems to know *something* about the nature of orthographic meaning, something that need not be the best way to convert sounds to graphemes to form his effect. This something is grasped in the exceptions to Twain's relative consistency, as well as the context a larger view of the orthographic field provides. Taking this line is not straightforward. Discovering the fine-grained orthographic differences within Twain's overall consistency requires use of the more precise non-PCA computational tools introduced in the methodological introduction. However, even without enhancing the resolution of our comparisons, simply visually inspecting the first chart allows the precision to conclude analyzing Twain's *Claimant* joke. As the novel progresses Twain's narrator presents no further glossings of Rossmore or Berkeley's speech. Yet even after traveling to the United States, young Berkeley's newly-met landlady Mrs. Marsh notes that he has retained his British accent, at least in that he "mispronounce[s] the words that's got a's in them, you know; such as saying loff when you mean laff" (Twain, 528). Presumably both Berkeley and Rossmore hold on to their phonological tenets throughout the whole of the novel, yet Twain never represents them with the sort of dialect orthography he famously peppers throughout the rest of his writings. In fact, as demonstrated by the chart, both Rossmore pere and Rossmore fils own relatively standard orthographic

profiles. Twain avoids rendering modern British accents as dialect orthography across almost the entirety of his corpus. Even his 1888 story "Concerning the American Language," a short fictional dialogue featuring an American conversing with a Brit on the subject of language variety, only dips into alternate orthography for explanations of specific national differences like "nao and kaow for 'know' and 'cow'."[92] Otherwise, the American and British interlocutors share the same orthographic conventions. Twain only consistently employs orthographic variation to indicate British dialect in *The Prince and the Pauper* and *A Connecticut Yankee in King Arthur's Court*, works set in past eras of England. Oddly, this choice aligns these characters who are assigned variant orthography more with Twain's speakers of American dialect than the British characters who sometimes share their textual space. Even if these eccentricities, as it will be shown later, vary in substance, Twain unites these two sets of characters through their shared commitment to linguistic singularity. This shared eccentricity marks them as groups "othered" in the eyes of Twain's contemporaries. Twain literalizes this effect in *A Connecticut Yankee in King Arthur's Court*. Time-travelling protagonist Hank Morgan only fully (and wrongly) determines that he has found himself in an asylum when a local knight challenges him to "a passage of arms for land or lady" in an Arthurian dialect.[93] For Morgan at least, linguistic evidence is enough to condemn one to marginality, a threat that is similarly real for Twain's dialect-speaking characters in works like

[92] Twain, "Concerning the American Language," 407.

[93] Twain, *A Connecticut Yankee in King Arthur's Court*, 213.

Adventures of Huckleberry Finn. At the same time, the association of orthographic alterity with the speech of these historical periods imbues even Twain's modern dialect speakers with the nation-defining romantic power of a King Arthur or Prince Edward. Twain uses orthography as the literary technology it is, producing the notion of a type of subjectivity through its modification.

Past Critical Work

Twain's use of orthography has earned some amount of highly systematic scrutiny from literary critics and linguists. His famous claim to linguistic prowess made in the aforementioned explanatory note to *Adventures of Huckleberry Finn* has been much evaluated, and scholars such as Shelley Fisher Fishkin and Lisa Minnick Cohen have produced works that investigate portions of his dialogue corpus in order to specifically characterize some of his orthographic tendencies, especially those connected to race. In *Was Huck Black?* Fishkin analyzes grammatical and prosodic elements of Huckleberry Finn's speech in order to argue that Twain had a special connection to "black voice" that expressed itself in in the utterances of Huck, a non-black character. Similarly, Cohen carefully analyzes the implied phonology of Jim's speech in *Huckleberry Finn* in order to lend further texture to long-held debates over the effect of

Jim's racialized characterization.[94] Both of these studies contribute much to this debate, but both have their own critical blind spots related to the two key literary aspects of orthography presented above. Fishkin focuses largely on grammatical and other higher level structural units, mostly eschewing orthographic analysis and thus not delving into the broader range of literary meaning orthography affords. In turn, Cohen (perhaps appropriately for a linguist) sets her sights on phonology at the expense of orthography qua orthography, ignoring the independent system of meaning orthographic choice represents. Moreover, both of these major analyses (and near all of the analyses of Twain's orthography that have come before) focus exclusively on *Adventures of Huckleberry Finn*. Although Fishkin grounds her findings in earlier Twain short stories, critics have mostly eschewed the potential contextualization that Twain's other novelistic works — including those featuring the characters of *Huckleberry Finn* — can provide.

Other shorter studies offer a mixed analytical bag. Older linguistic analyses of Twain's dialect such as those conducted by Lee Pederson and David Carkeet approach their topics systematically, but both attempt to correlate the orthographies they analyze with spoken regional dialects. In addition, both scholars restrict their conclusions to a

[94] Lott, North, Baker and Jones all offer foundational positions in this debate. Fundamentally, it revolves around what degree Jim serves as a simple racial caricature, and how that caricature/non-caricature is being used.

single work, *Huckleberry Finn*, and single characters.[95] Of previous work on Twain's dialect Susan Tamasi's "Huck Doesn't Sound Like Himself: Consistency in the Literary Dialect of Mark Twain" most approaches the methodology advocated by this dissertation. Tamasi explicitly rejects phonological accuracy as a metric for dialect orthography, instead choosing to evaluate the consistency of Huck's dialect features across *The Adventures of Tom Sawyer* and *Huckleberry Finn*. In doing so, Tamasi discovers that Huck's orthography does change over the course of the two novels, becoming "significantly more standard" in *Huck Finn*.[96] Tamasi hesitates to draw conclusions from this result, instead offering a variety of possible explanations for this latter deviation. Thus, while Tamasi's use of contextual comparison and multiple texts approaches the methodology espoused by this study, her approach still falls slightly short of our desired approach. Without the additional context created by comparisons across broader sets of works and characters and the integration of literary analysis into her study, Tamasi finds herself in a situation where she cannot choose between potential interpretations of Twain's orthographic variations. The conclusions Tamasi tentatively offers — that the orthographic shift represents a change in "role from secondary character to protagonist" or that it acts to "portray multiple linguistic varieties" that make up Huck's speech — could be adjudicated given additional comparison or close

[95] See Pederson, "Negro Speech in the Adventures of Huckleberry Finn" and Carkeet, "The Dialects in Huckleberry Finn".

[96] Tamasi, "Huck Doesn't Sound like Himself", 141.

reading.[97] For example, if similar dialect shifts occur in other Twain works as characters move from background to foreground, the first assertion could be convincingly argued.

Motivated by these previous studies, this chapter will undertake broad computational comparisons of the orthographies Twain utilizes across multiple novels on a character by character basis. The results of these comparisons will be used to reframe the conclusions of these studies and demonstrate the potential of orthographic comparison for literary studies. After a discussion of computational methodology, the first section will examine potential re-evaluations of Fishkin's conclusions in the light of additional contextualizing orthographic information. In turn, the second will set its sights on the character Chambers from *Pudd'nhead Wilson*, a figure whose orthographic profile proves quite unique when placed in the context of a comparative corpus.

Corpus level comparison

Finer grained comparison of the Twain corpus using perplexity difference measures reveals a picture of general consistency in Twain's use of orthography. For example, the set of narrator comparisons found in **Table 3-1**, using the narrator of *The American Claimant* as a basis of comparison, consistently recorded the lowest scores, and thus the highest similarities.

[97] Tamasi, 142.

For the purposes of this comparison character-narrators utterances were split into two sets, one consisting of narration speech and the other of dialogue. Perhaps unsurprisingly, Twain's narrators, even his diegetic narrators, speak very similar "standard" dialects. Even in the works where Huck takes up the mantle of narration, his speech remains relatively closely aligned to the other narrators, although it does deviate from the scores shared by the main group by earning some additional resemblance to Tom (see **Table 3-2**).

And indeed, Twain employs a relatively consistent orthography for British characters of eras past. Twain's King Arthur in *A Connecticut Yankee in King Arthur's Court* speaks more like the Tudor era denizens of *The Prince and the Pauper* than characters with other "standard," regional, or racial dialects (see **Table 3-3**).

These results should not necessarily surprise. They primarily serve to establish the efficacy of the methodology described above, as well as to provide a contextual backdrop for results to come. Twain's general consistency makes his variation all the more meaningful. Yet even these possibly expected similarities still shed additional light on Twain's use of orthography. The impressive orthographic similarity of Twain's cast of pre-modern British characters demonstrates at least one of the ends of his dialect usage — generating a sense of alterity without necessarily distinguishing between the subtle (or perhaps even evident) orthographic distinctions of two different time periods. Twain does not even achieve this similarity through shared vocabularies — a choice that would make these connections more readily apparent. Arthur, for example, shares about

78% percent of his vocabulary with Miles Canty and 80% with Edward Tudor from *The Prince and the Pauper*, numbers buffered by similar common interstitial words and ultimately smaller than those produced from comparisons with other characters with more modern orthographies. The main difference inheres in orthography itself, specifically in the way Twain exploits its systematic, textbound nature to reframe the meaning of his choice of orthographic conventions. The uniformity across these distinct texts and character sets amalgamates their putative phonological referents — the speech patterns of the Tudor era and a fantastic Arthurian Britain — into a unified indicator of alterity, mediated by orthographic conventions not necessarily apparent to the reading eye.

Orthography and Was Huck Black?

Orthographic similarities can act as vehicles of literary meaning, but differences or singularities often provide more immediate material for analysis. In *Was Huck Black?* Shelley Fisher Fishkin takes aim at the unique features embedded in Huck Finn's linguistic tendencies, offering an almost poetic reading of his cadence and word use. From this reading, Fishkin determines that Twain inflected Huck's dialect with the speech patterns of African-American interlocutors, voices materially recorded in pseudo-fictional works like his dialogic short story "Sociable Jimmy." For Fishkin, this inflection represents a desire to inject the African-American speech patterns Twain enjoyed and admired into mainstream American literary discourse, an act of

"appreciation, rather than appropriation."[98] Rather than take issue with Fishkin's privileging of the prosodic over the orthographic, I argue that the addition of contextualized analysis of Huck's orthography across multiple novels in which he appears modifies Fishkin's conclusion. By presenting a character with prosodic elements of African-American speech and an orthography more consistent with other dialects, Twain grants Huck a hybridized voice. In doing so, he, perhaps unwittingly, demonstrates that such appreciations of African American voice have already existed — they have just escaped recognition.

Fishkin bases her analysis of Huck's dialect primarily on evidence from two Twain short stories, focusing most closely on "Sociable Jimmy." Orthographic comparison of Jimmy, Jim and Huck's respective dialects across multiple novels (*The Adventures of Tom Sawyer*, *Huckleberry Finn*, *Tom Sawyer Abroad* and *Tom Sawyer, Detective*) as collected in **Table 3-4** paints an intriguing picture.

Appropriate to Fishkin's conclusions, Huck's orthography most closely coincides with Jimmy's in *Adventures of Tom Sawyer*, deviating most from that baseline in *Abroad*. In an almost opposite occurrence, Jim approaches Jimmy most closely in *Abroad* and diverges in the earlier *Huck Finn*. Part of the key to these divergences lies in that most orthographically fertile grapheme, the apostrophe. From both the KL comparison scores of the apostrophe distribution for each character and the heatmaps of their orthographic models it becomes evident that a large portion of the variance

[98] Fishkin, *Was Huck Black?*, 108.

between versions of Jim, Jimmy and Huck stems from the differing use of this particular feature (See **Figures 3-3** and **3-4**).

Jim's orthography most fully departs from Jimmy's in *Huck Finn*, before returning to relative similarity in *Tom Sawyer Abroad*. Similarly, the comparison between Huck and Jim himself varies across novels, with their closest mark occurring in *Tom Sawyer* (see **Table 3-5**).

These comparisons act as a sort of sliding scale that roughly correlates with the chronology of the corpus, and ultimately, the overall chronology of Twain's novelistic output. In the earlier works both Huck and Jim speak in relatively "Jimmy-esque" fashion. As a result their own pairwise comparisons draw closer together. By the later novels Jim has more wholeheartedly adopted Jimmy's character of speech, while Huck has pulled away from both. On first blush, this wholeheartedly supports Fishkin's conclusion. Huck draws closest to Jimmy in *Adventures of Tom Sawyer*, while Jim in turn drops away. Yet despite this, Huck still most closely aligns orthographically with "standard" speaking characters in a way that Jimmy (and Jim) do not. Throughout the novels in which he appears, Huck's use of orthography ranks as more similar to the speech of Twain's narrators than to Jim or Jimmy's, and consistently measures up as closest to Tom's (See **Tables 3-6** and **3-7**).

Orthographically, at least, Huck remains closest to the dialect of his fellow white Missourian. Manually inspecting the stochastic matrix models of Tom and Huck from *Huckleberry Finn* next to Jimmy's own reveals the contours of their differences.

Compared to both Tom and Huck, Jimmy much more frequently uses the apostrophe as the final grapheme of a word. Additionally, Jimmy's vowel-consonant distribution differs from Huck and Tom. Consonants that more likely follow vowels in the speech of the two boys more often precede vowels in Jimmy's speech, with 'd' serving as the most notable example. In contrast, both Huck and Tom's apostrophe and vowel-consonant distributions much more closely resemble those of Twain's narrators. These differences are palpable in the heatmaps collected in **Figures 3-5, 3-6** and **3-7**.

By offering additional orthographic context to Fishkin's initial assertion a slightly different picture emerges. In anticipation of arguments over the origins of southern speech that would begin a few decades later, Twain presents Huck as exemplary evidence that such speech has always been mixed.[99] By juxtaposing prosodic elements of African American speech with orthographic conventions more in line with his narrators and white identified characters, Twain produces a linguistic tension only truly legible on the written page, where the conflation of two different dialect features may be seen. This tension serves to warn readers away from a powerful mistake — letting their orthographic expectations fall into regional or racialized silos. This conclusion also requires qualification. The fact remains that Huck's speech diverges much more extremely from Jimmy in the later novels — *Abroad* and *Detective*. Huck does take up

[99] Early commentators on dialect in the United States frequently attributed Scottish influence as the source of southern dialect, a claim that has been heartily challenged by evidence of African American influence. See Bonfiglio, *Race and the Rise of Standard American*, 225.

the mantle of narration in these works, tempting the conclusion that the difference in orthography stems from the downward effect of this changed role. However, the attributed corpus separates Huck's moments of narration from his dialogue, rendering this theory impossible. Twain's early slanting of Huck's orthography towards the one he employs for African American characters is very real — but in the light of this evidence from later texts, questionably conscious. In this view, the later novels serve as earmarks of orthographic backsliding. No longer interested in hybridity, Twain reverts to a different form of nonstandard orthography that deviates from his initial point. Alternatively, this shift demonstrates the complexity of using a "constructed" orthography — perhaps Twain simply forgot some of his precepts. However, examining the content of *Abroad*, one of these two later novels, offers a more satisfying possibility.

In theory, *Abroad* serves as Twain's version of a Verne-esque travelogue. Huck, Tom and Jim commandeer a mildly futuristic balloon vessel from a slighty mad scientist and use it to travel the globe — even though they spend most of their time in the deserts of northern Africa. While they find their fair share of misadventure along the way, including fighting off a pride of lions and buzzing some unaware Bedouins, the dialogue of the novel largely concerns questions of standardization and knowledge.[100] The novel steps between incidents of socratic trialogue. Unerringly, these moments feature Tom pitting his often largely accurate but incomplete pedagogical explanations of various phenomena, ranging from metaphor and mirage to maps and timezones, against the

[100] So much so that one critic deems the tale an exercise in exploring "man's epistemological limitations." Briden, "Twainian Epistemology and the Satiric Design of 'Tom Sawyer Abroad'," 43.

recalcitrantly skeptical questioning of Jim and Huck. During the exchange that occurs over the interpretation of a map, Huck attempts to apply the information it provides in a too-literal sense. Seeing "meridians of longitude" printed on the cloth leads Huck to believe that they must also occur on the earth itself.[101] Tom's initial attempt to correct him is met with a rebuff that stems from Huck's preference for intuitive, phenomenal knowledge. Huck points out that "you can see for yourself" that the map has the meridian lines as occuring on the ground, and Tom's assurance that no such lines actually exist on the globe lead Huck to deem it "a liar" (ibid). Huck is both right and wrong in deeming it so. As a piece of technological apparatus, the map combines two forms of knowledge in a manner that allows access to a level of useful information unavailable to one form alone. As is his wont, Huck focuses primarily on the intuitive, phenomenal element of the map, the aspect whereby it is simply presenting itself as "a representation of this other thing, but smaller." It also, however, pulls a third person form of knowledge into this regime by rendering the meridian lines as a piece of visual information in the same representational plane. Making the unseen seen, in this case, requires hybridization. The invisible meridian lines are pulled into the realm of phenomenal knowledge in order to make them cognitively available to this form of perception. They already exist in some sense — mathematically perhaps — and rendering them in pictoral form should, in theory, act as a pointer to this meaning rather than as a phenomenal representation. This is precisely what Huck misses.

[101] Twain, *Tom Sawyer Abroad*, 683.

Writing too is a tool, and in its mixture of "instinctual" syntax and vocabulary with the technology of orthography it is also, like the map, a hybrid one. Just like with the map, attempting to understand it as one thing completely available to one form of knowledge comes with risks. By tilting Huck's orthography away from Jim and Jimmy and towards Tom, Twain provokes a question about the coherency of this approach in linguistics as well as cartography. Even though the prosody of Huck's speech in the later novels may well still resemble Jim and Jimmy's own tones more than it does Tom's, this orthographic information not so readily available to phenomenological experience makes concluding that "Huck speaks like Jimmy" a somewhat harder pill to swallow. Whether intentionally or not, Twain invites us to consider the coherency of asking a tool, in this case writing, to present a unified truth. Making such a demand risks making a means an end, a proposition potentially dangerous when the subject said tool is meant to apprehend is a human self rather than the physical surface of a planet. This divergence recapitulates Twain's previously examined orthographic jokes in a slightly different light. Beyond just being a more serious application, this change is less clearly intentional (in the traditional sense) than the tension his jokes capitalize on. Returning to the typewriter joke helps us avoid drawing a too firm conclusion on whether Twain "intended" for Huck's cartographic confusion to serve as a key to his orthographic shift. One of the "poles" of this joke is the possibility that Twain is drawn to the typewriter's allure despite himself. The typewriter diagnoses and makes material some impulse generated by his subject positions — a technology-fond American, a writer with an interest in printing. The impulse, or whatever notion of subjectivity produces it, may not

be graspable from the top-down phenomenal position of the one impelled by its appearance. Yet, as Twain shows, the meaning it produces through the tool it employs, in this case typewritten text, can be imposed upon and used for very conscious purpose. The typewriter joke serves as a diagnostic, just as a shift in orthographic style can as well. Neither offers up a stable point of reference, neither truly fixes a subject. Both are tools and, as Twain seems to realize, subject to multiple use. However, the impulses that manifest themselves in the desire to use a certain tool in a certain fashion, to typewrite instead of scribble, to use one set of orthographic standards instead of another, reveals that the depths of the self may always escape the attempt to rationalize their contents from the top-down. Tools such as these provide novel ways to act "despite ones self," and thus in turn provide novel ways to de-calcify structures of subjectivity.

Pudd'nhead Wilson

Twain perhaps more explicitly continues this orthographic exploration in his 1894 novel *Pudd'nhead Wilson*. *Wilson* tracks the fate of two babies, one partially black and one white, born in the same household at the same time. Fearful of the vicissitudes of slavery, Roxy, the mother of Valet de Chambre (or Chambers), the child identified as black, switches him in the cradle with his future master Tom Driscoll. Due to the similarity of their complexions the entirety of the town, including Driscoll and Chambers themselves, remain ignorant of the swap well into their adulthood. The novel comes to a close with a general revelation of the swap and Tom (born Chambers) being

convicted of a series of robberies before being remanded into slavery. All of this is facilitated by the title character David "Pudd'nhead" Wilson's habit of collecting fingerprints. Appropriately, the child born as Tom (henceforth referred to as “adult Chambers”) has the most eccentric orthographic fingerprint of all of Twain’s characters investigated in this study. Adult Chambers registers high perplexity scores when compared to any of Twain’s other characters. This holds regardless of a character’s putative race, region or even family affiliation, with a similar amount of divergence even when compared to his biological mother, Roxy (see **Table 3-8**).

Chambers still compares most closely to Roxy and Jim, but only at an arm's length. For comparison, Roxy registers much closer scores with her nearest matches (see **Table 3-9**).

When compared to the overall corpus that includes more than just Twain's works, Chambers still retains Roxy as one of his top matches, qualified again by a relatively wide perplexity difference measure. Indeed, Chambers diverges from every character in the overall corpus by an unusually large margin. Roxy's orthography, on the other hand, compares to the rest of the corpus in a more standard fashion, to such a degree that even though she is second on Chambers's overall similarity list he is 1007th on hers.

Chambers has a singular orthography both when compared to Twain's other characters and to the overall corpus. Such unusual singularity demands further investigation of both the "hows" and the "whys" of Chambers's orthography. The set of computational measures performed corpus-wide already contains more granular insight

concerning the former question. In the methodological introduction, the characteristic mathematical differences between a difference and divergence measure was discussed. A difference measure is symmetrical — two entities, when compared, will always return the same difference measure. Divergence, on the other hand, is asymmetrical. This, as also previously noted, has the capacity to capture relationships of enclosure a difference measure does not register. The perplexity divergence specifically captures the liklihood for the model of one of the characters to generate the orthographic sequence associated with a different character in the corpus. One character's model may be quite adept at generating the orthography of its comparator, but the comparator's model does not necessarily hold the same power to explain the first character's actual grapheme sequences. The former orthographic model in some sense contains the latter. It produces all of the features of the comparator, but also contains additional regularities the comparator's model does not.

Quickly examining the perplexity divergence columns on **Tables 3-8** and **3-9** reveals the potential of such measures. Roxy's closest matches (by perplexity difference) also have relatively two-sided perplexity divergence scores. Her model tends to explain the orthography associated with her closest matches quite well, and vice versa. Chambers, on the other hand, has unusually one-sided scores. In the context of the Twain-only corpus, his closest matches have orthographic models that explain his orthography reasonably well (the div 2_1 column). In turn, he struggles to explain their models to any significance at all (the div 1_2 column). These models capture Chambers's orthographic features but also contain a signficant amount of features that differ from

what Chambers's model can explain, either by number or type.

It is tempting to dismiss this as a form of computational mirage, one of the constituent members of the third term of Twain's triad of "lies, damned lies, and statistics." The most plausible explanation along these lines comes from the relative lack of dialogue granted to Chambers across the course of *Pudd'nhead*. As the models used in this project derive simply from the empirical sequences of graphemes associated with a literary speaker, perhaps his relative lack of speech causes his model to be under-powered. Other characters may have more orthographic features simply because they spend more time in dialogue.

Again, it is reference to the broader corpus that quells this fear. Chambers's model does in fact predict the orthographies of certain characters in the broader corpus quite well — it is just that none of these characters happen to be Twain's.[102] **Table 3-10** contains the five characters from the full corpus whose orthographies Chambers's model best explains (as measured by "Perp. Divergence Base Chambers" column).

Two further oddities emerge with the piling on of this additional granularity. The first is an intriguing aside from our current line. Chambers's model best predicts the orthographic sequences used by a specific set of authors, namely Ellen Glasgow, George Washington Cable, and Charles Chesnutt. Many of the characters that rank immediately

[102] Empirical testing shows that this too is not necessarily a function of the length of model input. A sample model with a similar length (in this case Sheriff Plunkett from Edward Eggleston's *The Graysons*) is more distant from the overall corpus by perplexity difference, but does not have the same massive gap between perplexity divergence scores that Chambers evinces.

after these five belong to Glasgow's series of Virginia-centric historical novels, Cable's *Dr. Sevier* and *The Grandissimes*, or Chesnutt's *The Conjure-Woman* and *The Marrow of Tradition*. This grouping in and of itself has potential implications. More towards the current investigation, the orthographies that Chambers's model best predicts anticipate his own orthographic sequences quite poorly. On an orthographic level, relating to Chambers is a one-sided affair. Some examples from vocabulary comparison help provide texture to this realization. For example, in general, Chambers utilizes more interstitial apostrophe than a character like Jim's *Huckleberry Finn* incarnation. Where Jim says "Agin" Chambers says "ag'in." Similarly, when Jim would utter "busted," Chambers offers "bu'sted."[103] Chambers also adds additional graphemes, largely vowels, to the terminal ends of his wordforms. This is visible in his stochastic matrix model, and exemplified in his wordform "marse," the complement to Jim's usual "mars." As expected from the relatively close perplexity difference, Chambers's vocabulary closely resembles Roxy's. However even here there are some notable exceptions. Where Roxy says "foun'" Chambers utters "found"; where Roxy tends to use "noth'n'" or "nothin'" he uses "noth'n." In short, the orthography Twain grants the adult Chambers is something of a singular hybrid. He draws features used by Twain's other characters in to his own orthography without consistently resembling any particular one. While these differences are small given Chambers's relative lack of speech, the impact of his orthographic meaning is disproportionately large. Twain inserts the linguistically unique Chambers in

[103] Jim does employ "bu'sted" once in *Tom Sawyer Abroad*.

to a text that continually calls for the destabilization of norms of race and hereditary destiny, where identifications of white and black suddenly shift and family affiliations seem perpetually in doubt. The friction afforded by Chambers' unique voice, a difference perceptible, but not necessarily explicable, during the moment of reading, certainly contributes to this destabilization. It is hard not to notice that Twain grants this form of orthographic dialect to a character the townspeople eventually identify as white, a mocking stab at the tenets of racial purity. However, Twain's use of dialect orthography to this end also renders the distinction between page and sound palpable, if not quite visible, offering an inroad to deeper understanding of the textbound systematicity of orthography itself.

By the end of the novel Chambers's singular existence also becomes narrative. Upon becoming a free white man in the eyes of the law Chambers finds himself ostracized by the black community that once sheltered him. The white community of the town also cannot offer him refuge. He ends the novel in "an embarrassing situation," a person possessing what the white community views as the "basest dialect" and "vulgar and uncouth" bearing associated with the black community.[104] Ultimately, Twain simply writes him out of the novel. Now possessing a sense of self that makes him feel "at home and peace" only in the relatively nonsocial setting of the kitchen, Twain removes him with the simple statement that following his "curious fate" further would be a tale too long to tell (ibid). Importantly, Chambers himself discovers this new sense of self. His

[104] Twain, *Pudd'nhead Wilson*, 225

speech and mannerisms, notions once habitual and unthinking, become estranged, question marks that don't point to any particular sense of who he might be. Not quite black and not quite white, his identity and the sense of self attributed to him thereby, become as singular as his orthography. Chambers's new sense of identity, his reinterpretation of self, only emerges due to a perspective shift attributable to technological intervention. David "Pudd'nhead" Wilson uses his collection of fingerprints to both indict Tom (Roxy's son, so the "real" Chambers) on a murder charge and to destabilize Chambers's identity. Twain takes loving care to describe the actual scene of Wilson's intervention.[105] After enlarging copies of the prints with the aid of a "pantagraph," he reinforces the whorls and patterns of the print with ink (Twain, 211). This process allows him to "read" the prints, to interpret the bodily information that was not just unseen, but in fact un-conceptualized, prior to the utilization of this technological approach. Doing so more than discovers what Chambers has been all along. Rather, it produces a novel racial subject, one incomprehensible to the categories employed by the general community, and who defies phenomenal interpretation of his outward signifiers even when it is the subject himself doing the interpretation.

In combination with Chambers's unique orthography, this produces a final

[105] This is a controversial and much-discussed passage in Twain scholarship. Wilson is often accused of using pseudoscience to prop up existing notions of race, for example in Sundquist, *To Wake the Nations*. In contrast Leigh, in his article "Literary forensics," chooses to somewhat defend Wilson's actions in this scene. This project can afford to hold a neutral attitude towards Wilson's techniques; when viewed orthographically Twain is demonstrating the technological basis of racialized subjects in general.

Twain-ian orthographic joke, albeit one with a far more serious edge. Twain uses these two factors to rewrite a common understanding of the relationship between orthography and racialized subjectivity. Elements of dialect are "supposed" to point to a subject's nature, to reveal information about their sense of self. Even if the knowledge they produce is mistaken, syntax and vocabulary can be bent to this end. Like Chambers's physical appearance and mannerisms they are easily available on the phenomenological level — no apparatus required. Written orthography, on the other hand, follows fingerprints as a latent technology of the self that has the capability to actually produce new types of subjects. Both of these forms of knowledge are intimately close, literally inscribed on body or brain, yet they remain invisible unless some sort of paradigmic shift in view occurs. Regardless of whether the technology that produces this shift is itself pseudoscientific or even inaccurate in any sense, Twain shows that the end result is not discovery or re-inscription of extant norms, but creation. At its base what is created is still just another distinction — a new way to divide into categories. However, for it to serve this purpose room must be made for it in the prior-held and phenomenologically understood web of distinctions that precede it. The town is unsure what to do with a subject who presents as black yet, due to information gleaned through novel technological means, cannot be. It is similarly difficult to understand how Chambers fits in Twain's generally consistent field of orthographic meaning. This, however, is a situation for the reader, and specifically a reader properly equipped with a third-person way to render the differences visible in the first place. Simply reading Chambers's lines would not reveal his singularity, in the same way that the townsfolk

reading his physiognomy does not. Twain is, in effect, distinguishing between distinctions. Because they are produced by technological intervention, orthography and fingerprinting are forms of distinction that only make an impact on the phenomenal understanding of some form of self with difficulty. They can be read from a first-person perspective, but doing so risks radically misunderstanding their import. Once elements such as these become visible, they clash with prior-held intuitive and phenomenal senses of what counts as meaningful. They provoke questions without "natural" resolutions — how can Chambers's fingerprints possibly indicate that he is white when the structures of meaning cognitively available to the town make him so clearly black? How could an orthographic understanding of his dialect show Chambers to be a singular figure when he reads in a fashion not dissimilar to Roxy? Yet, in a different regime of knowledge, these conclusions are both to some degree wrong.

Twain lets his townsfolk off the hook by writing Chambers out. They do not have to take the time to adjudicate between the two levels of their new experience of Chambers. The critic of orthography is not so lucky. Twain leaves the interested reader with a conundrum. If both the phenomenal understanding of Chambers's orthography as decently consistent with his other black identified characters and the third-person understanding of his orthography as relatively singular are true, why not allow these two truths to exist disharmoniously? The new figure that emerges need not be brought under some broader phenomenal umbrella. When technology is involved man, and especially man's phenomenal understanding, is not the measure of all things. These two

understandings of the situation can be left to pass quietly in the night—perhaps there is no "complete picture" that could integrate or adjudicate between both at all.

Figures

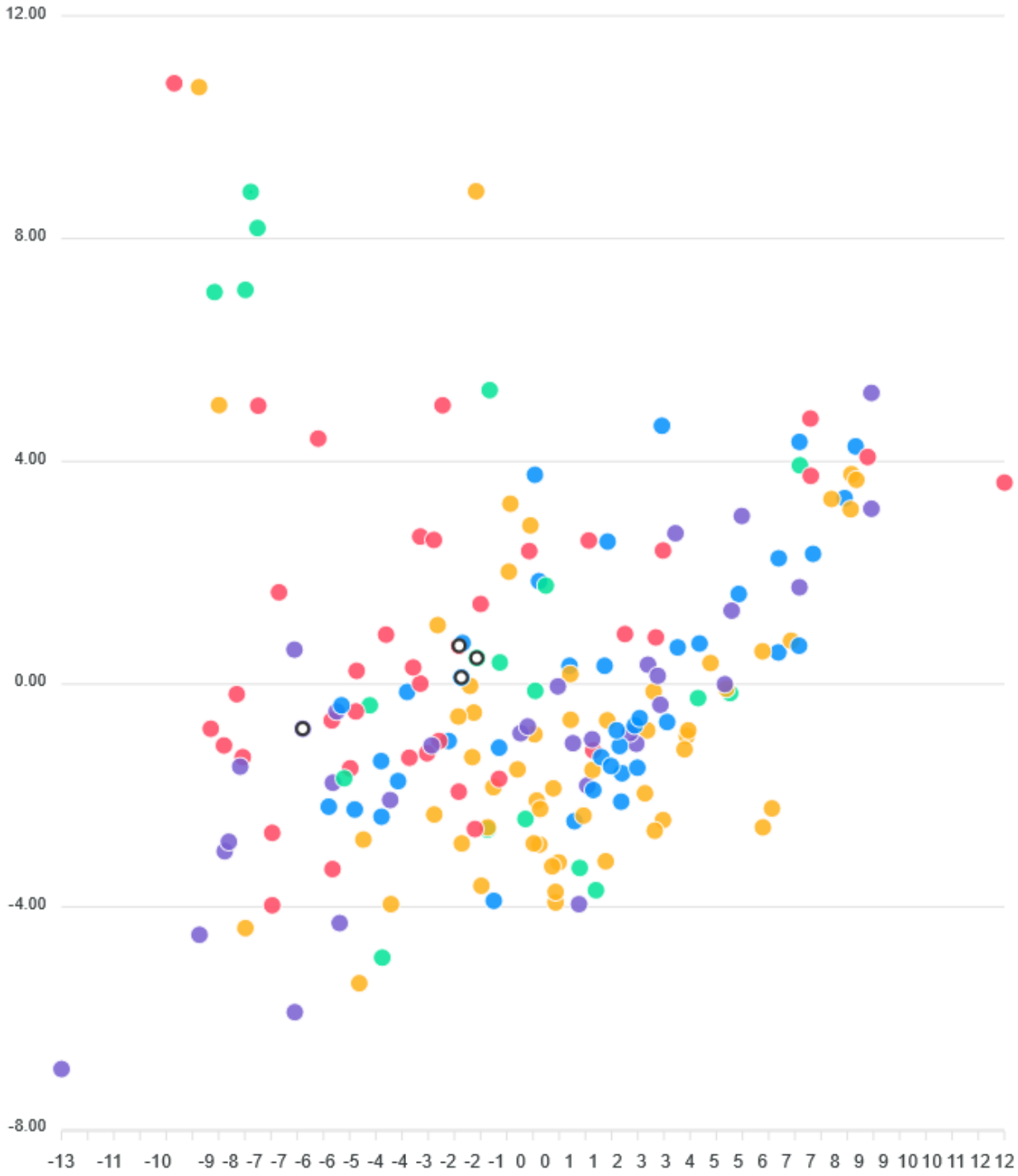

Figure 3-1. A scatter plot charting the PCA reductions of the stochastic matrix models generated from Mark Twain's characters. The highlighted points are the stochastic matrix

models generated from the four versions of Tom Sawyer collected in the corpus, one from each of *Huckleberry Finn, Tom Sawyer, Tom Sawyer Abroad and Tom Sawyer, Detective*.

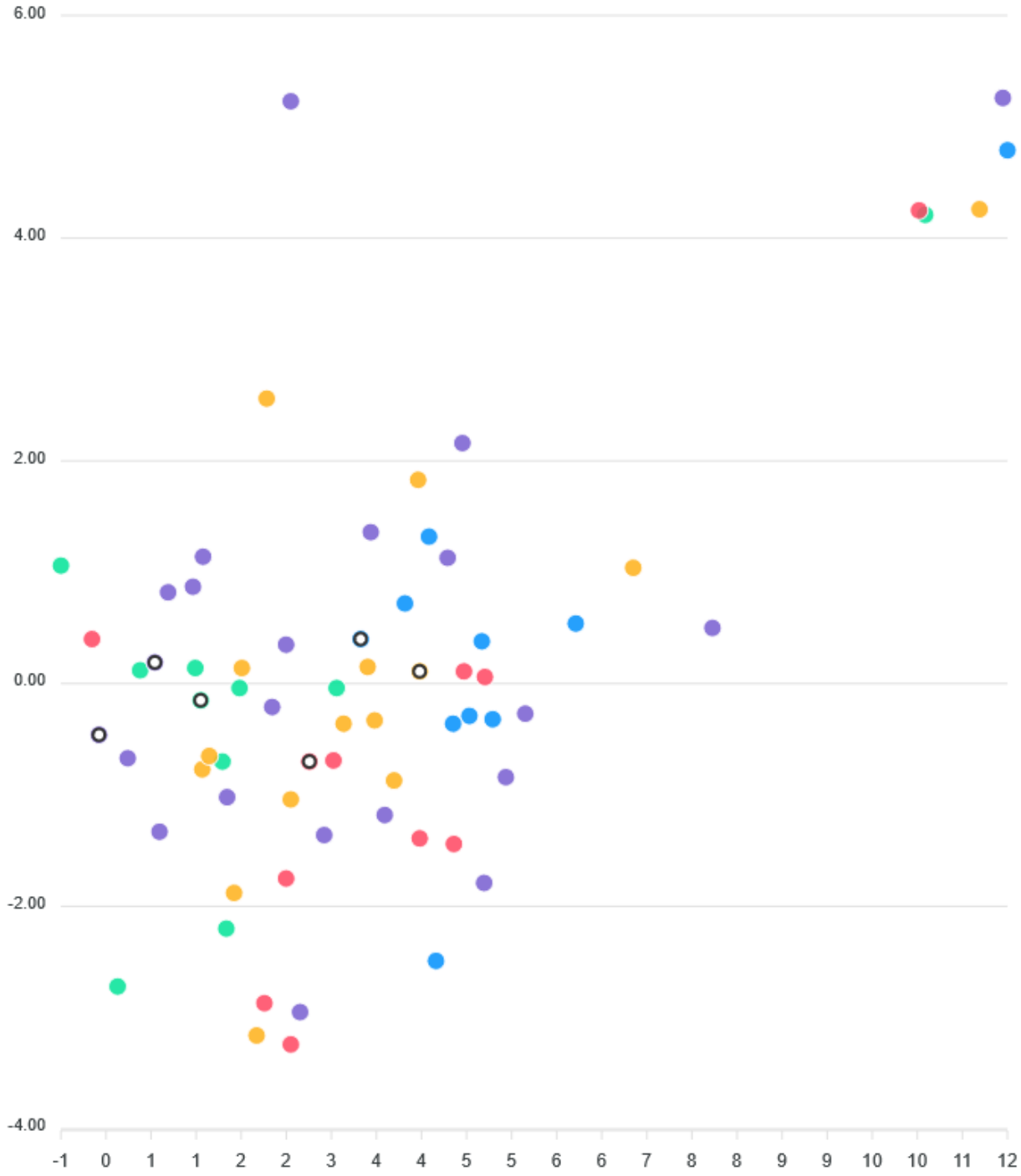

Figure 3-2. A scatter plot charting the PCA transformation of the stochastic matrices of characters found in James Fenimore Cooper's *Leatherstocking Tales*. The highlighted points are the instantiations of Natty Bumppo in his various guises (Leatherstocking, Pathfinder, etc.).

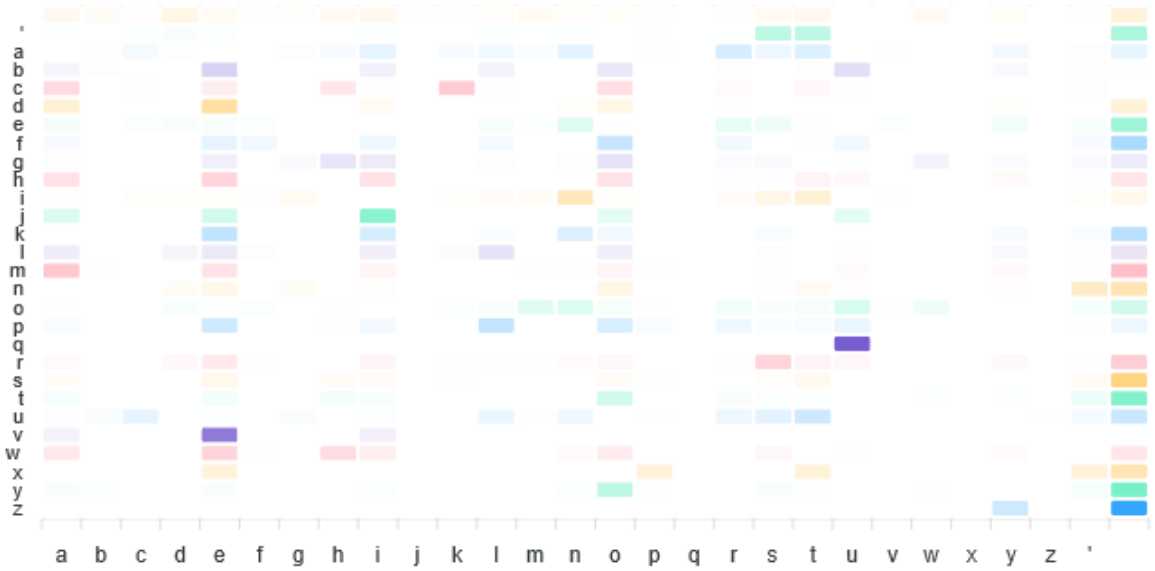

Figure 3-3. A heatmap that visualizes the stochastic matrix model associated with the *Tom Sawyer Abroad* version of Jim. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme. The broader variety of colored intersections associated with both the row and column (so predictor and predicted) instances of the apostrophe indicates it is used quite broadly.

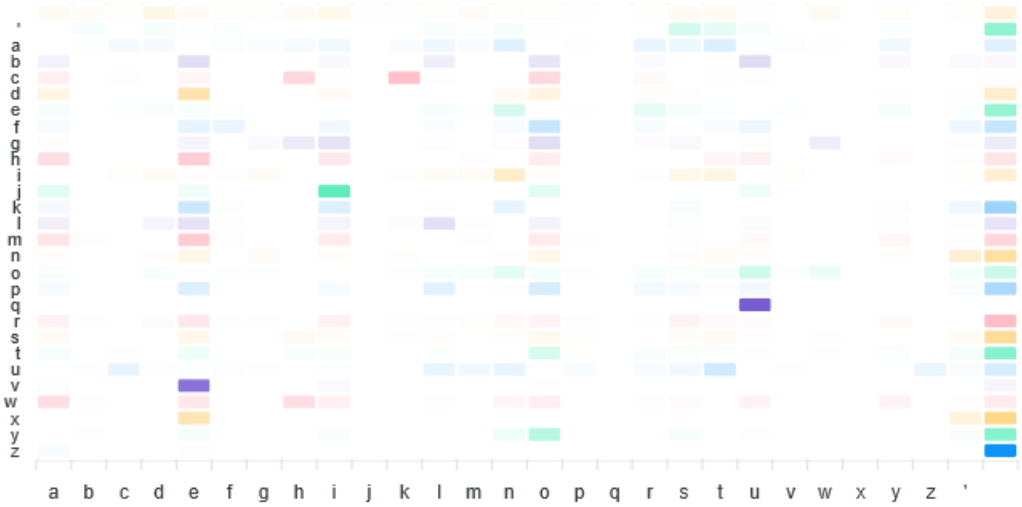

Figure 3-4. A heatmap that visualizes the stochastic matrix model associated with the version of Jim found in *Adventures of Huckleberry Finn*. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme. Here, the apostrophe is generally more controlled.

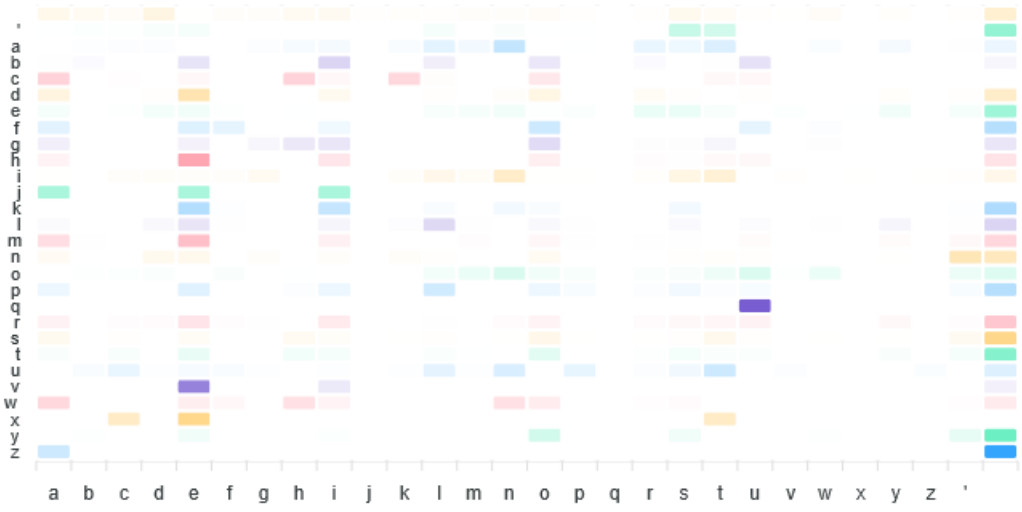

Figure 3-5. A heatmap that visualizes the stochastic matrix model associated with the character Jimmy from "Sociable Jimmy." Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme. The broad distribution of graphemes certain consonants (for example, 'd') predict helps distinguish Jimmy's orthography from the one used by Huck. This is indicated by there being multiple possibilities with roughly equal likelihood, as shown by the relatively similar darkness of their boxes.

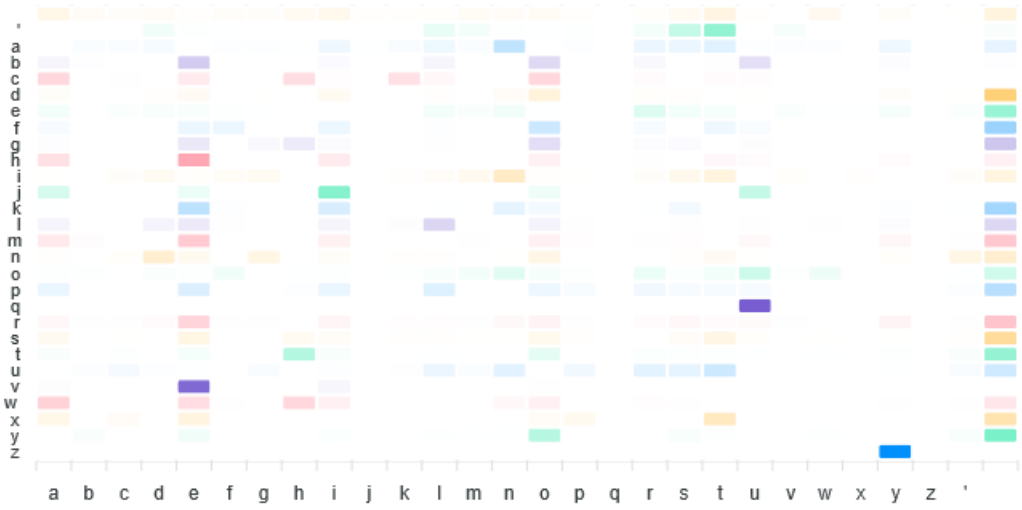

Figure 3-6. A heatmap that visualizes the stochastic matrix model associated with the *Adventures of Huckleberry Finn* version of Huck. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme. Certain consonants (for example, 'd') are tightly controlled by Huck's orthographic system. They have relatively few likely successors, and thus are characterized by a smaller number of darker-hued boxes.

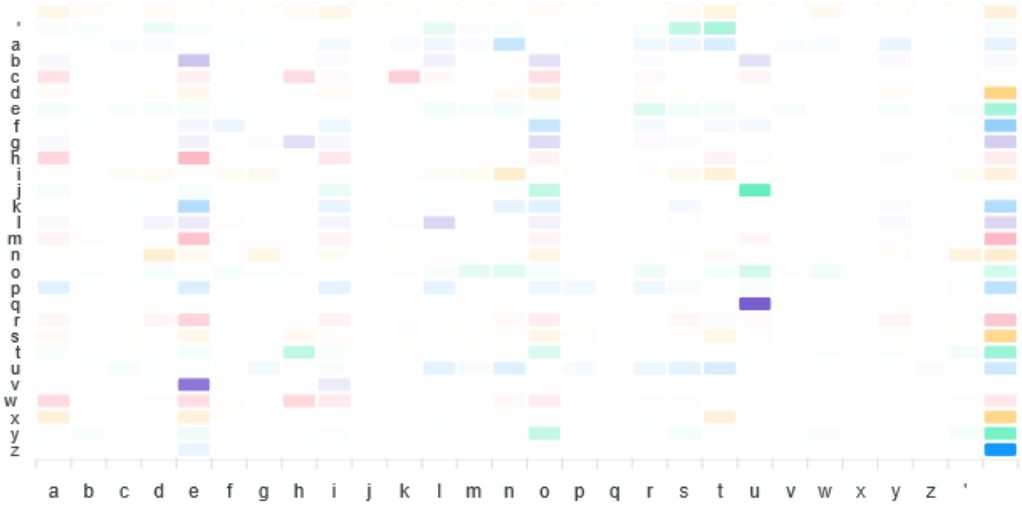

Figure 3-7. A heatmap that visualizes the stochastic matrix model associated with the *The Adventures of Tom Sawyer* version of Huck. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme. Certain consonants (for example, 'd') are tightly controlled by Huck's orthographic system. They have relatively few likely successors, and thus are characterized by a smaller number of darker-hued boxes.

Tables

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
The Prince and the Pauper	Narrator	10.1	9.99	10.2	0.23	2.97
Pudd'nhead Wilson	Narrator	10.13	10.16	10.11	0.34	4.12
A Connecticut Yankee In King Arthur’s Court	MorganNarrator	10.14	10.12	10.15	0.11	0.41
Roughing It	Narrator	10.24	10.16	10.32	0.19	2.09
Tom Sawyer Abroad	HuckN	10.29	10.72	9.85	0.54	2.65
The Innocents Abroad	Narrator	10.3	10.39	10.22	0.21	1.79
Adventures of Huckleberry Finn	HuckN	10.37	10.73	10.01	0.63	3.23

Tom Sawyer, Detective	HuckN	10.44	11.09	9.79	0.95	2.83
A Connecticut Yankee In King Arthur's Court	Clarence	10.52	10.38	10.66	0.66	1.04

Table 3-1. A table comparing the stochastic matrix generated from Twain's narrator in *The American Claimant* to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to the *Claimant* narrator. HuckN is the tag used for the model generated from Huck's instances of narration, and HuckI the tag used for the model generated from Huck's interstitial narrative moments. KL is the compound Kullback-Leibler based measure, and KL on apostrophe is the comparison score of the two texts when only comparing the apostrophe distribution of the stochastic matrix. This comparison can be useful, but as will be explored in the "Orthographic Extrema" chapter it is also very sensitive.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Tom Sawyer Abroad	HuckN	9.39	9.33	9.46	0.2	1.06
Tom Sawyer, Detective	HuckN	9.42	9.29	9.56	0.34	0.75
Adventures of Huckleberry Finn	Tom	9.93	9.97	9.89	0.61	0.83
Tom Sawyer, Detective	JakeDunlap	9.96	9.88	10.04	1.17	1.98
Tom Sawyer, Detective	Tom	9.98	9.99	9.96	0.69	1.18
Adventures of Huckleberry Finn	Huck	10.01	9.87	10.16	0.91	0.8
A Connecticut Yankee In King Arthur's Court	MorganNarrator	10.01	10.33	9.69	0.44	1.52
Tom Sawyer Abroad	Tom	10.23	10.42	10.04	0.76	1.5

Adventures of Huckleberry Finn	HuckI	10.24	11.5	8.97	1.22	1.29
The American Claimant	PollySellers	10.29	10.35	10.22	0.65	1.5
A Connecticut Yankee In King Arthur's Court	Clarence	10.31	10.53	10.09	0.69	1.9
Roughing It	Bemis	10.31	10.73	9.9	1.7	5.33
Pudd'nhead Wilson	Narrator	10.34	10.08	10.61	0.76	6.03
The American Claimant	Narrator	10.37	10.73	10.01	0.63	3.23

Table 3-2. A table comparing the stochastic matrix generated from Huck Finn's moments of narration in *Adventures of Huckleberry Finn* to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to this version of Huck.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
The Prince and the Pauper	EdwardTudor	10.15	10.19	10.11	2.53	6.39
The Prince and the Pauper	TomCanty	10.29	10.14	10.43	1.87	7.13
A Connecticut Yankee In King Arthur's Court	SmallpoxWoman	10.33	9.71	10.94	3.82	0.37
A Connecticut Yankee In King Arthur's Court	Sandy	10.33	10.4	10.26	3.85	5.69
The Prince and the Pauper	MilesHendon	10.37	9.91	10.83	2.32	6.12
A Connecticut Yankee In King Arthur's Court	Clarence	10.47	11.01	9.92	1.56	3.96
A Connecticut Yankee In	MorganNarrator	10.63	11.29	9.96	1.73	5.99

King Arthur's Court						
A Connecticut Yankee In King Arthur's Court	Morgan	10.68	11.45	9.91	2.41	5
The Prince and the Pauper	KingHenry	10.74	11.39	10.09	3.87	8.24
Tom Sawyer Abroad	HuckN	10.74	10.31	11.17	1.93	7.42

Table 3-3. A table comparing the stochastic matrix generated from King Arthur in *A Connecticut Yankee in King Arthur's* Court to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to Arthur.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Tom Sawyer Abroad	Jim	12	11.39	12.6	1.91	0.63
The Adventures of Tom Sawyer	Huck	12.74	12.23	13.25	2.32	2.77
Adventures of Huckleberry Finn	Jim	12.8	11.25	14.35	1.91	1.6
Adventures of Huckleberry Finn	Huck	12.94	12.54	13.34	2.92	4.33
Tom Sawyer Abroad	Huck	15.07	17.23	12.91	3.05	4.72
Tom Sawyer, Detective	Huck	15.59	17.77	13.42	4.54	6.28

Table 3-4. A table comparing the stochastic matrix generated from the character Jimmy from Twain's "Sociable Jimmy" to versions of Jim and Huck from Twain's other works.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
The Adventures of Tom Sawyer	Huck	12.92	12.01	13.82	1.39	3.68
Adventures of Huckleberry Finn	Huck	13.03	14.26	11.81	2.71	4.54
Adventures of Huckleberry Finn	HuckN	13.84	15.29	12.39	2.06	5.81
Tom Sawyer Abroad	HuckN	14.01	12.52	15.5	2.33	4.91
Tom Sawyer, Detective	HuckN	14.85	12.49	17.22	3.05	5.04
Tom Sawyer Abroad	Huck	15.69	11.56	19.81	3.24	6.08
Tom Sawyer, Detective	Huck	15.95	11.65	20.25	3.16	5.67

Table 3-5. A table comparing the stochastic matrix generated from the *Adventures of Huckleberry Finn* version of Jim to the various versions of Huckleberry Finn found in the corpus.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
The Adventures of Tom Sawyer	Tom	9.94	9.9	9.99	0.37	0.39
Adventures of Huckleberry Finn	Huck	10.13	10.24	10.02	1.76	0.7
Adventures of Huckleberry Finn	Tom	10.15	10.35	9.95	0.92	0.72
Tom Sawyer Abroad	Tom	10.54	10.4	10.67	1.48	1.33
Tom Sawyer, Detective	Tom	10.54	10.26	10.83	2.12	1.3
Tom Sawyer Abroad	HuckN	10.56	10.23	10.89	1.25	1.67
Adventures of Huckleberry Finn	HuckN	10.56	10.88	10.24	0.96	1.6
Adventures of Huckleberry Finn	SallyPhelps	10.6	11.09	10.11	2.5	1.13

The Adventures of Tom Sawyer	Polly	10.73	9.88	11.59	2.84	1.22
Tom Sawyer, Detective	JakeDunlap	10.77	10.37	11.16	2.36	1.97

Table 3-6. A table comparing the stochastic matrix generated from the *The Adventures of Tom Sawyer* version of Huck to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to this version of Huck.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Adventures of Huckleberry Finn	Tom	10.57	9.65	11.48	1.78	1.59
Adventures of Huckleberry Finn	Huck	10.59	9.53	11.64	1.78	1.54
Tom Sawyer Abroad	Tom	10.61	11.66	9.56	2.42	0.69
The Adventures of Tom Sawyer	Tom	10.92	9.7	12.14	1.98	2.24
Tom Sawyer, Detective	Huck	10.96	11.06	10.85	2.35	2.97
Tom Sawyer, Detective	Tom	10.98	9.92	12.04	2.25	1.43
Adventures of Huckleberry Finn	SallyPhelps	11.03	10.14	11.92	2.65	2.45

Tom Sawyer Abroad	HuckN	11.05	11.98	10.12	1.82	2.2
The American Claimant	PollySellers	11.06	12.25	9.87	2.08	1.23

Table 3-7. A table comparing the stochastic matrix generated from the *Tom Sawyer Abroad* version of Huck to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to this version of Huck.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Tom Sawyer Abroad	Jim	15.29	19.68	10.91	4.5	1.33
Pudd'nhead Wilson	Roxy	15.78	21.02	10.55	4.41	2.73
Adventures of Huckleberry Finn	Huck	17.81	22.89	12.72	4.9	4.35
Tom Sawyer Abroad	Tom	18.48	23.73	13.22	5.76	4.94
The Adventures of Tom Sawyer	Huck	18.53	23.64	13.41	5.12	4.92
Sociable Jimmy	Jimmy	18.73	23.92	13.53	4.62	3.03
The Adventures of Tom Sawyer	Tom	18.8	23.67	13.93	5.32	6.43
Adventures of Huckleberry Finn	Jim	18.9	25.73	12.06	5.1	3.16
Tom Sawyer Abroad	Huck	18.91	20.71	17.11	5.19	4.97

Adventures of Huckleberry Finn	SallyPhelps	19.05	22.17	15.93	5.4	6.18

Table 3-8. A table comparing the stochastic matrix generated from *Puddin'head Wilson*'s Chambers to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to Chambers.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Tom Sawyer Abroad	Jim	10.91	10.63	11.19	0.75	0.56
Adventures of Huckleberry Finn	Jim	11.16	11.34	10.99	0.69	0.4
The Adventures of Tom Sawyer	Huck	12.23	11.45	13.02	1.01	2.77
Adventures of Huckleberry Finn	Dauphin	12.33	11.88	12.78	1.55	1.75
The Adventures of Tom Sawyer	Tom	12.38	11.47	13.3	1.31	3.18
Sociable Jimmy	Jimmy	12.45	11.24	13.65	1.98	1.76
Adventures of Huckleberry Finn	Huck	12.46	11.48	13.43	2.42	3.2
Tom Sawyer Abroad	Tom	12.77	11.76	13.78	2.4	3.95
A Connecticut	Morgan	12.93	12.25	13.6	2.06	3.86

Yankee In King Arthur's Court						

Table 3-9. A table comparing the stochastic matrix generated from *Puddin'head Wilson*'s Roxy to other characters in Twain's corpus. The characters included here are those with the lowest perplexity difference scores when compared to Roxy.

Comparison Character	Comparison Text	Perplexity Difference	Perp. Divergence Base Chambers	Perp. Divergence Base Comparison
Big Abel	The Battle-Ground	18.85	12.11	25.6
Champe Lightfoot	The Battle-Ground	19.58	12.13	27.02
I	The Cavalier	18.48	12.15	24.81
Raoul	The Grandissimes: A Story of Creole Life	23.1	12.19	34.04
Julis	The Conjure-Woman	23.8	12.22	35.39

Table 3-10. A table comparing the stochastic matrix generated from *Puddin'head Wilson*'s Chambers to other characters in the overall corpus. The characters included here are the 5 that Chambers's stochastic matrix model best explains. These characters have the lowest perplexity divergence scores when using Chambers as the basis of comparison.

Eben Holden, Plantation Literature and the Fate of Dialect

Prelude: A Bird's-Eye View of the Region(alism)

While resisting the urge to default wholesale to a large-scale distant reading of texts, corpus driven analysis also recognizes that the zoomed-out features of a text can contribute to critical inquiry. Such a zoomed out view proves most useful when investigating similarly large-scale formations, especially, in the context of this corpus, genre. The late nineteenth century saw the rise of a loose genre that scholars have termed "regionalism" or "local color" fiction. Regionalist works typically depict the pocket cultures of the United States — areas that, in fiction, if not in fact, remain distant from and thus "untainted" by the growing cosmopolitanism of the burgeoning American metropoli. For many of these works employing an orthography deemed "local" in comparison to the emerging standards of the city-based national press provides a key part of this effect. **Figure 4-1** demonstrates the relative similarity of the characters found in the corpus.

When using this plot to evaluate the similarity of characters across texts the caveats discussed in the methodological introduction remain in force. These caveats aside, the pre-marked points on the plot denote the general location of some clusters of characters that share orthographic traits. The consistency of these groupings validates the intuitive accuracy of PCA as a comparison analytic. For example, narrators grouping together in a specific spot close to the cluster of characters that use "standard"

orthography accords with both the conscious and preconscious analyses of writing systems performed during the act of reading. Other characters that share some of the "typical" markers of non-standard orthography — race, class, nationality and region — also orbit around the same axes. Overall this speaks to consistency. The literary guardians of alphabetic knowledge gather those that speak into categories based on their putatively identifiable traits. Despite this, singularities remain. Certain characters cross the boundaries of their known orthographic regions and venture into territory held by other forms of subjects. One such character, Uncle Eben Holden from Irving Bacheller's late regionalist novel *Eben Holden: A Tale of the North Country* (1900), is marked on the PCA plot. Uncle Eb's orthographic profile pushes him away from the main grouping of backwoods/northeast/northwest speakers, leaving him in a small grouping of characters on the margins of the general region of speakers typically identified as African American. Eb's unusual orthographic position on the larger scale offers an inroad into the specific details of a text that has previously been mostly overlooked, one replete with promising insights about the state of dialect-employing literature at the end of the nineteenth century and beyond.

Media Threat in the late Nineteenth Century

By the late nineteenth century the relatively staid media landscape of the United States had exploded into a cacophony of new forms, each making its own claims on the notion of the real. Photography, the phonograph and eventually the motion picture all

present strident challenges to what Brian Rotman calls "alphabeticism," the fundamental "logic of representation" implied by alphabetic dominance and the notion of the self such dominance fosters by presenting means of representation that capture the real with more mimetic accuracy.[106] Even without, or perhaps in part due to, this outside pressure the status of writing dominance came under pressure from the advent of the dime novel. Newly literate audiences hungered for reading material, and publishers like the house of Beadle and Adams provided such material, printing numerous cheap commissions full of feeling and sentiment, but also paving the way for "cheap publication of full-length reprint novels."[107] Such sudden widespread access to the production and consumption of literature threatened the epistemological status of writing simply through volume. The written depictions found in newspapers and novels no longer held exclusive epistemological hold on the realities of the event — and even more importantly the realities of the soul. The prophylactic feedback loops of writing epitomized by type's ability to "forcibly transcribe the unconscious" and thereby render the nonconscious self subject to organized processes of knowledge shrank before media that carried with them an excess of noise not translatable into an alphabetic code.[108]

[106] Rotman, *Becoming Beside Ourselves*, 9.

[107] Mott, *Golden Multitudes*, 150.

[108] Kittler, "Dracula's Legacy," 80.

This challenge to writing's dominance did not go unmet. In response to the threat of the real late nineteenth-century readers developed an appetite for new forms of literature that made central the use of written dialect.[109] The works that clustered into the formations we now call regionalism, plantation literature and dialect poetry marshaled diacritical marks and "nonstandard" orthographies to the defense of the written, bulwarks against the encroachment on its traditional epistemological territory. This technique was not wholly novel. As explored elsewhere, earlier nineteenth century writers had already used dialect orthography in a variety of different works, most notably humor and historical writing. Refreshing this technique in service of ordered, regular knowledge required a new ideological appendage — accuracy. Like some of the modern critics and linguists who would follow them these writers derided such earlier works as examples of eye dialect, defined by one more modern critic as "orthographic changes that signal dialect but do not correspond to 'real' dialect features."[110] These new writers sought to tame a technique that, in their eyes, only signified "this was said in dialect" by yoking its mess of apostrophes and dashes to specific phonetic features found in the "actual" sound patterns of speakers.[111] In effect, they sought to recapture the noisy

[109] These new forms also served other political ends, such as policing national identity. See Foote, *Regional Fictions*.

[110] Holton, *Down Home and Uptown*, 58.

[111] As noted previously any such attempt at least partly founders when faced with the ambiguities of translation from grapheme to phoneme.

real — tics, sputters, pitches — revealed by the new media of the era and in doing so reassert the written word's status as the prime means of regularizing knowledge of the other and the self, a window into the neural medium itself.

Below I will examine this formation in the context of plantation and regionalist literature, with special focus on Irving Bacheller's bestselling late regionalist novel *Eben Holden: A Tale of the North Country*. Bacheller was a prolific writer of popular fiction during the fin-de-siecle period, but the success of *Eben Holden* eclipsed all of his other literary forays.[112] *Eben Holden* itself is a book more strange than it has any right to be. Part Yankee/northern New York local color, part bildungsroman, part children's story book and at times even part gothic novel, at its core the story follows the progression of our narrator William as he grows from being the child ward of Eben, and then the northern New York Brower family, into becoming a successful writer and reporter in New York City. This generic confusion has one straightforwardly identifiable source. Citing Bacheller's autobiographical work *From Stores of Memory*, Mott notes that *Eben Holden* was initially itself autobiographical in design, centering on "a character [Eb] formed upon that of a hired man the author had known in his boyhood" and was originally intended to be a "juvenile" tale.[113] Arguably Bacheller was just making do.

---

[112] Mott cites a handful of the emerging best-seller lists to come up with a sales figure of "at least" 750,000 copies. Bacheller had future success in writing, but *Eben Holden* captured a certain form of lightning in a bottle. Mott, 204.

[113] Mott, 203. Mott infers the original generic intent from Bacheller's submission of the tale in serial form to *St. Nicholas* magazine. This is a reasonable inference, but it is also reasonable to note that Bacheller

Faced with rejection on the level of genre, a statement that this would not do as juvenile literature, he flung the proverbial kitchen sink at the problem, resulting in a work with general interest across both age range and generic preference.

Bacheller's own words add some further texture to this relatively simple account. Reflecting on Eb's origins in *Stores* Bacheller supplements his statement that Uncle Eb was based on a farm-hand from his youth with the claim that he knew this original figure "as well as [he] know[s] the multiplication table" (Bacheller 1938, 144). This allowed him to approach the composition of his novel in a naturalistic fashion, plotting no specific "adventures for the journey" but instead relying on "a chain of them" already resident "in [his] own experience" as well as the escapades suggested by the "changing scenes" and their corresponding inhabitants encountered by William as he ages and travels the eastern United States (ibid). Bacheller claims a very specific form of innate expertise as the driver of his plot. His understanding of Eb's "type" is learned empirically, but in such a fashion that it becomes rote instinct more akin to mathematics than psychology. Bacheller appeals to this knowledge but does not claim he can make it explicit. The subconscious calculations he performs result in a comprehensive account of a particular subject, and this in combination with physical movement leads to a

---

himself points out that at least one of the editors of *Nicholas* also edited the more general-interest *Century Magazine* and that the refusal of a tale was somewhat tantamount to a rejection from New York publishing in general. Bacheller may have conceived the project flexibly in the first place. See Bacheller, *From Stores of Memory*, 161.

plotted product that then necessarily corresponds to some generally acknowledged truth, no matter the wild generic twists that might ensue.

This form of semi-unconscious knowledge immediately smacks of substructural orthographic style, and for good reason. Bacheller was also the principal of the Bacheller Syndicate, typically cited as the first modern periodical syndication scheme in the United States.[114] As an alphabetic guardian of the real, this leaves him to invest wholeheartedly in text's ability to capture the vocal stylings of his titular elder Vermonter. As per the ideology of orthographic accuracy, doing so should offer privileged access to even the nonconscious aspects of the regionally particular subjectivity exemplified by Eben but also theoretically resident in even the "real" humans who shared his subject position. Bacheller offers access to the very grounds of what might be termed the "Northeast consciousness," the subtle factors of upbringing and locale that reside in (especially vocal) behavior and remain largely unseen to even the subjects themselves. However, orthographic substructural style is a complicated beast. As previously explored, it is both produced and consumed somewhere between the conscious and nonconscious self, and allows for more creativity and variation than its cousins in dialect syntax and vocabulary. The latter two may well unambiguously point to a particular self, but the orthographic implementation of Uncle Eb's northeast twang bears a striking resemblance to characters from a very geographically distinct

[114] Uncertainties about the genesis of the Bacheller syndicate remain, but he is at least cited as such by his contemporaries and was likely at least attempting to syndicate fiction as early as 1883. See Johanningsmeier, *Fiction and the American Literary Marketplace*, 72.

tradition — plantation literature and its respondents. In examining these similarities through both computation and close reading I argue that this strange correspondence does in fact reveal the contours of a particular subject-position — Bacheller's own.

The Northeast Mind

Though Bacheller's novel initially revolves around the native Vermonter Ebenezer Holden, the first action of the story takes place in northern New York state, well removed from the bowered slopes of the Green Mountains. *Eben Holden* opens with the titular farmhand fleeing his native land, the young now-orphaned son of his former late employers (eventually our narrator, William) in tow. Eben, perhaps due to his lack of "home or visible property," is deemed a less worthy caretaker for young William than William's "dissolute uncle" thus provoking him to escape the Vermont authorities by removing himself and the boy to the relative safety of northern New York (32). There, the pair happen upon the Browers, family farmers who adopt William and employ Eben as their handyman and chief confidant. The Browers, themselves paradigmatic instantiations of a particular regional type (New York Dutch) contrast Eben's Yankee antics with their staid industry and piety, yet embrace him as the regional type he cannot help but be. What follows is a novel divided into roughly two parts. The first half consists of episodes from William's youth and features the stories and pranks of Eben prominently. The latter sees William mature and move to New York City where he works for Horace Greeley's *Herald*, fights in the Civil War, and, finally,

marries his adoptive sister Hope Brower. This latter section de-emphasizes Eben's role, relegating him to becoming the main mover of the B-plot. Therein he discovers that the first Brower son, thought deceased at sea, still lives, and that he has spent numerous years as the "night-man," a quasi-mythical wildman of the New York countryside who regularly visits the Brower farmstead. Paralleling William's development, he later becomes a charitable gentleman of leisure in New York. This once-missing Brower also serves as William's guardian angel, personally carrying the injured William off of the battlefield at the first Battle of Bull Run and rescuing the Browers from financial ruin by gifting them a substantial amount of his fortune, one originally generated from a seed investment made by Uncle Eb.

Bacheller works to associate Eben with the northeast context through dialect, but also through regionally-associated traits. Eb may be a citizen of the U.S. but in this novel he is abroad, a regional singleton examined through comparison to those who don't share his subject position. Statistical examination of Eben's dialect orthography does indeed reveal the fingerprints of the northeast. When compared to other members of my corpus using the perplexity measure, Eb's orthography emerges as most similar to a veritable "who's who" of northeastern orthographies, with characters from Joseph C. Lincoln's *Cap'n Eri*, Sarah P. Mclean's *Cape Cod Folks* and Sarah Orne Jewett's *The Country of the Pointed Firs* all making the list of close matches, alongside two other characters from *Eben Holden* itself (see **Table 4-1**).

In addition to the characters from these textual representations of the northeast, Eb, the Vermonter-turned-New Yorker, also orthographically resembles characters from two midwestern texts — Gene Stratton-Porter's *Song of the Cardinal* and Zona Gale's *Friendship Village*. All of these most-similar texts come from the latter part of the corpus, with *Cape Cod Folks*'s 1881 publication date being the earliest point. Just on first blush these results offer a macro-level narrative about the waning years of regionalism. Even as writers focus more and more closely on regionally specific settings — upstate New York, central Wisconsin, coastal Maine — certain "close enough" dialect treatments homogenize under the pressure of centralizing literary forces. Alternatively, this set of similarities might encode a historico-geographical reality. As the U.S. pushed its boundaries ever more westward the orthographic traits originally associated with the east transfer (along with eastern bodies) to the newly-defined central regions. Needless to say, the linguistic reality of this hypothetical shift is irrelevant. Even if the actual traits of midwestern spoken language of the period varied greatly from any potential eastern predecessors, the temptation to render them as orthographically similar in text reveals the desire to connect these new western folk to the known United States of the east, even as many settlers to the midwest region came from the Nordic and Bohemian regions of Europe. Hamlin Garland nods to this fact with the eponymous character of *A*

Little Norsk, even if Anson Wood and Bert Gearhart, the characters whose dialect most resembles Eb's, both seem to hail from Wisconsin.[115]

On its own this broader view establishes a much less surprising correlation than my earlier insinuation that Eb has the blood of plantation literature flowing through his veins. Though the imagined geographical reach of this particular form of "backwoods" orthography is, perhaps, unexpected, it is easily reconcilable to factors of nationalism or literary homogenization described above. Understanding Eb's peculiarity requires a closer look at two of the characters he resembles most closely not just orthographically, but also structurally, from the above lists of works — Almira Todd from *The Country of the Pointed First* and Calliope Marsh from *Friendship Village*.(See **Figures 4-2, 4-3** and **4-4**).

The texts that house these characters — *Eben Holden*, *The Country of the Pointed Firs* and *Friendship Village* — all feature a first person narrator with some sort of close, pseudo-familial relationship to a dialect-speaking guide character. Eben acts as an adoptive uncle to his narrator; Almira Todd hosts hers and serves as a capable Maine facsimile of Virgil to her narrator's Dante. Calliope lives apart from her narrator, but still catalyzes her being "abruptly launched" into Friendship Village society as her mentor

[115] Intuited from both their orthographic attributes and their tendency to refer to the birth-father of their young Norwegian ward as "the Norsk" indicating his status as a recent immigrant. Garland, *A Little Norsk: Or Ol' Pap's Flaxen*, 18.

and first local friend.[116] All three are also older than their narrative-scribbling charges, making it easy to slot them into the parental roles that are explicitly or implicitly absent in all three works. These figures all offer more than mentorship to their young(er) wards. They compound their usefulness by providing fodder for the writings of their narrator-proteges, tales suitable for re-transmission beyond their original provincial contexts. Jewett's narrator first hears the sad tale of "Poor Joanna" from Todd before setting off (with the Todd family as guides) to report on this unfortunate's island hermitage for herself.[117] Gale's narrator learns much of Friendship Village's small history in much the same way, being regaled by Calliope with the backstories of characters like the intriguingly-named Eb Goodnight in both prospective and retrospective fashion.[118] Both texts even adapt their formal structure to the cycle of narration and re-narration performed by these character pairs, utilizing an organization more akin to a closely connected cycle of short stories with a shared set of locations and characters than other novel forms.

The subset of regionalist, dialect-heavy literature frequently termed "plantation literature" also employs these structural norms. Joel Chandler Harris's *Uncle Remus, His Songs and Sayings* typifies both the standards — and the controversy — of the genre. The cycle of Remus stories, like the above texts, takes place in a hermetic regional

[116] Gale, *Friendship Village*, 27.

[117] Jewett, *The Country of the Pointed Firs* 62.

[118] Gale, 138.

bubble; in this case the confines of a southern plantation rather than a remote northern hamlet. Each individual chapter also centers on an elder mentor figure, Remus, supplying a younger listener, in this case a white child of the plantation, with a distinct unit of local folklore. Charles Chesnutt also draws on these elements in the composition of his *Conjure Tales*, building each chapter around a unique narrative-centric interaction between the aged Uncle Julius and the young northern purchasers of his home plantation. No doubt, as argued by other scholars, he does so to bend these generic norms to his own ends. Heather Gilligan provides the subtle but important point that Chesnutt does not solely draw on a tradition of African-American folklore that would be "largely illegible" to a contemporary white readership in order to effect his subversion, instead choosing to inhabit the "esoteric" halls of plantation fiction in order to make his point.[119] The verbal tones and narrative situations he employs take on the character of the very object of his critique. I, however, argue that the third formation of "local color" regionalist literature deserves a place in this dyadic structure. The close of the nineteenth century sees the three genres of (1) plantation literature, what might be termed (2) "anti-plantation literature" written in the model of Chesnutt and (3) regionalist texts conglomerate into a complex web, related by not only their structure but also the striking similarities in dialect orthographies attributed to characters of even vastly different racial and regional identity. *Eben Holden*, specifically Uncle Eb himself, embodies the homogenization of these groupings that at least in retrospect look so

[119] Gilligan, "Reading, Race, and Charles Chesnutt's "Uncle Julius" Tales," 196.

temptingly distinct. Eb serves as an index into the destabilizing effect of new media on the written word's ability to marshal knowledge of the inner life of the subject. His role as a point of confusion between genre formations that so explicitly delineate the subject positions they seek to understand (the plantation mind, the New England mind, the what-have-you mind) marks a shift in how dialect fits into linguistic strategies of knowing, one that must adapt to the destabilization of its privileged status as the very stuff of thought.

On the level of textual speech, Eb resembles Remus and Julius to a much greater degree than his most orthographically intimate backwoods familiars. Most tellingly, the closest similarity emerges from the similarity in usage of the elision apostrophe, one of the dialect writer's preferred tools. On the whole Eb speaks in harmony with the midwestern/northeastern standard present in the corpus, but when it comes to the distribution of this particular feature (compared using the KL divergence) his accent adopts a slightly different timbre. Comparing characters across just this feature increases Eb's similarity to Remus and Julius, making them the 39th and 50th most similar characters, respectively. Meanwhile, Todd's similarity with the two comes in at ranks 378 and 443 respectively, indicating a significant divergence in how each author deploys the elision apostrophe. Strikingly, he is tenth most similar on Remus's list when sorted by this feature comparison, and thirteenth on Remus's, appearing in the midst of other African-American identified characters and well away from most of his backwoods

kin.[120] **Table 4-2** illustrates the specificity of this difference.

By all measures, Julius and Remus align most closely. Dismantling the halls of plantation literature from within requires Chesnutt to use its own language — in this case meaning orthographic style. Chesnutt's version of the African-American storyteller orthography, if not necessarily his syntax or vocabulary choices, points mostly back to the generic system he seeks to undermine itself. This conclusion textures the import of Eb's substructural similarities. Bacheller's backwoods Yankee storyteller produces an orthographic form drawn more from his function than his origin. Although the empirical view of a character's vocabulary cannot match the explanatory power of the sort of modelling that forms the backbone of this study, it does contextualize the results such modeling provides. Inspecting the vocabulary of each character reveals that all four of these characters make heavy use of elision in the final position of words, most commonly removing a terminal "g" as seen in wordforms like "runnin'" or "walkin'". Remus, Julius and Eb also frequently elide final "d" and "t" graphemes, evidenced in "han'" and "mus'", in a way that Todd does not. Finally, this subset of three uses the elision apostrophe outside of the terminal position much more frequently than Todd. Words like "j'ints" and "reg'lar" enter their vocabulary with more frequency than Jewett's Maine elder. The model heatmaps associated with each character sum up the distinction between the odd grouping of Remus/Julius/Eb and Todd — the former three

[120] Note that Remus was split into two characters — one for his own attributed voice and one for when he adopts the voices of one of his stories. This does not significantly alter the results.

speak with more widely distributed elisions (see **Figures 4-4, 4-5** and **4-6**).

Eb retains enough of the style of his contemporaries to be most similar to them overall, and is even more similar to Todd in the way he uses the elision apostrophe than he is to Julius or Remus. However, his relative similarity to this pair of African-American storytellers shows his chimeric nature. He contains the backwoods, but he also contains something else. He utilizes features from both groupings in a way the other backwoods characters do not, engaging in a sort of surplus subjectivity that isn't explicable in the regional vocabulary. Whether any of the orthographic features used by this triad picks out some particular phonetic quality at all is beyond the point — the grouping is self-evidently not random. Accusations of eye dialect simply do not apply. Eb varies in a manner that is wholly describable from the third person view, even if it may only confuse or go unnoticed by the first. He contains something beyond what he first seems, a nature that becomes fully legible only when viewed as a self-referential function of a system of substructural style and genre rather than a reference to any sort of outside world.

Eb's surprising orthographic similarity to Remus and Julius also extends to his narrative function. Starting with Eb's first entrance as a character, Bacheller strives to associate Eb with the sort of oral storytelling traditions typified by his African-American counterparts. Recalling some of the earliest memories of his youth, William identifies Eb as "not a strong man" who "had never been able to carry the wide swath of the help in the fields" but who still earned the love of his employers-cum-family through "his

kindness and his knack of storytelling."[121] Eb's worth comes not from his labor, but from the way he "enriched the nomenclature" of William's country neighborhood through his verbal inventiveness (ibid). In turn, both Julius and Remus primarily perform emotional, rather than physical, labor. Julius occasionally acts as the plantation coachman, but largely works by revealing the history of the estate to its new owner. Now in the twilight years of a long life of forced labor, Remus spends most of his time sharing his stockpile of folklore with the plantation youth. Despite the undeniable difference of his subject-position, Bacheller associates Eb with these formerly enslaved storytellers through the shared form of value held by their labors. At times crafting this similarity leads Bacheller to deviate from the largely standard bildungsroman structure of *Eben Holden*. The scenes that feature Eb most heavily tend to mimic the episodic cycle of repetitive narration that undergirds both Harris's and Chesnutt's works. Almost all of these scenes occur early in the novel, and a large portion of them come during Eb's and William's daring escape from the New York authorities. Bacheller endows this episode with a particular rhythm. The days (as recalled by William) consisted of flight and peril. However, once the fugitives found safe camp for the night Eb would invariably draw upon his natural avocation of storytelling to amuse and educate his young fellow-traveler. These tales draw upon a mythology of his own homespun design and focus on human interaction with the animal world. He puts much emphasis on the "swift", a cryptid creature "sumthin' like a panther" that "lay in the edge of the woods at

[121] Bacheller, *Eben Holden*, 2.

sundown" and "makes a noise like a woman crying, to lure the unwary" (Bacheller, 3). He also uses the early escape scenes to supplement his fearful stories of the swift with more Aesop-like tales concerning lost intimacies between humans and panthers, bears, and crows; intimacies brokered by the animals' ability to speak human tongues and by their adherence to very anthropic social notions of kinship and morality. All of these animal figures, the swift as well as the expanded menagerie of the later tales, possess a form of liminality, a sense that they somehow exist on the edge of the human and the animal.

Problematic racial politics aside, the most enduring legacy of Harris's Remus stories is a similar centrality of an all-too-human animal folkworld. Remus's morality-centric Bre'r Fox and Rabbit yarns have endured well beyond the memory of Harris himself, even earning what is surely the most lofty position any piece of American culture can achieve — a now disavowed spot in the Disney canon. On the whole, Harris offers a more hermetic view of the animal-human intersection. For example, the interactions between Fox, Rabbit and human in the tale "Mr. Fox Gets into Serious Business" resemble, more or less, the typical human view of human-animal intimacy. The human farmer, christened, tellingly, "Mr. Man" by Remus, becomes irate at Rabbit for nibbling at his crops and attempts to extract revenge through the handle of a hickory switch.[122] Though Mr. Man addresses his victim as he strikes, Fox (because, naturally, Rabbit has tricked him into taking the drubbing on his behalf) refuses to

[122] Harris, *Uncle Remus His Songs and His Sayings*, 143.

reply, responding to Mr. Man's query of "W'at kin'er w'atzyname is you, ennyhow" only with silence, and allowing him to "talk on" in monologue (ibid). If it wasn't for the surrounding accounts of Rabbit convincing Fox to take the drubbing and his subsequent gloating the story would have no hint of animal language or society at all. This interaction between hominid and canid comes off as a near miss rather than a moment of congress. The farmer and the fox speak the same tongue, but the narrative situation denies them the ability to utter anything to the other at all. Adding to this the puzzle of "Miss Meadows and the gals" confuses the situation even more. Rabbit frequently references this "Miss Meadows," but the only answer Remus has to his youthful listener's inquiries about how this character could possibly be is a simple "Don't ax me" and a deferral of authority to his predecessor storytellers — she has simply always been in the tale (Harris, 67). Though she seems to interact with Rabbit on equal footing, she also seems to be human, rather than human-animal. Stella Brewer Brookes reports that this befuddled Harris's publishing partners as much as it did Remus's young audience. When faced with the task of visually depicting Meadows and the girls, Harris's illustrator Frederick Church attempted to disambiguate their nature by contacting Harris himself, who simply replied that as "the compiler" of the tales he had no notion of her nature either. Bereft of any other avenues, Church ultimately settled on a human depiction.[123]

[123] Brookes, *Joel Chandler Harris: Folklorist*, 27.

Chesnutt literalizes the trope of liminal animality more concretely than his generic fellow-travelers. Uncle Julius provides his young northern listeners with a set of tales that feature sorcery-induced transmogrifications of humans into animals (and often back again). Due to either their relatively advanced ages or the outlandishness of Julius's claims of literal sorcery (rather than Harris's folklore animal second-world or Bacheller's lost paradise of human—animal connections) Julius's interlocutors often approach these tales skeptically. When, in "The Conjurer's Revenge," Julius asks the pair if they "wouldn' wanter b'lieve" that the club-footed Primus was "oncet a mule," they at least initially reply in the affirmative — they certainly wouldn't want to believe it, and almost structurally cannot, as this allows Chesnutt an opportunity to let Julius unfold his tale.[124] Primus still interacts with the human world during his equine sojourn, finding time to get drunk on new wine and to injure a potential suitor of his presumptively widowed wife. More so than the other two tale-tellers, Julius likely does not intend for his animal tales to be believed. Despite his insistence that he tells only the truth, each tale usually ends with his interlocutors discovering that there is a perfectly earthly reason for Julius's seemingly rarified and magical yarns — normally some amount of profit to be earned by guiding the new plantation owners down a particular story-driven path. His double-talk strategy certainly, as critics have pointed out, could be read as Chesnutt's attempt to goopher his own audience in the same way that Julius goophers his, using it as his own strategy to undermine the power of racialized

[124] Chesnutt, *The Conjure Woman, and Other Conjure Tales*, 72.

stereotypes. This is a compelling argument, and one that has been discussed at great length from numerous points of view.[125] But what also emerges as noteworthy, and what binds these three texts together, is how the nexus of an unusually human form of animality and the oral forms that depict them work to undermine the hitherto unquestioned power of writing genres to delineate and define types of self.

Of these it is *Eben Holden*, as both the latest and most singular of the three texts discussed above, that wrestles most desperately with its position in an increasingly information-rich world. Bacheller, the newspaper man at the turn of the century, distrusts his framework-disorienting storyteller in a way Harris and Chesnutt do not. Remus and Julius inhabit a cyclical, almost mythical, position in their respective works. Their moments of storytelling recur ever onward, emanating from settings that rely on them to produce a sense of continuity in a changing world. In contrast (an especially curious one given the eponymity of the novel) Eben's star fades as the novel *Eben Holden* progresses. Eben, and the rural setting he partially embodies, recede into the background as William removes to New York City and, eventually, the bloody fields of the American Civil War. The novel finishes in a fashion unthinkable to those conditioned by the eternal nature of Chesnutt and Harris's storytellers — with William at the grave of the now late Eben Holden. William ponders Eben's epitaph, consisting of

[125] For example, Brodhead sees the complicated and arguably collaborative publication history of the tales as an indication that Chesnutt was varying his approach to how he depicts racial stereotypes; sometimes undermining them, sometimes playing into them. Brodhead, "Introduction," 19.

his last words, rendered, curiously enough, in his signature dialect orthography. The epitaph attests to Eben's honesty (for example, he "Never ketched a fish bigger'n 't was") before noting that Eben is now "Goin' off somewheres, Bill — dunno the way nuther" but that no matter the destination he "ain't afraid" (Bacheller, 432). This provides a tidy diagetic bow, connecting Eben's journey into the afterlife with the escape scene that opened the novel. It also tempts the conclusion, once popular with scholars of regional literature, that Bacheller uses this scene to lament not only the dear departed Eben Holden but also the specific parochial form of life he represents, an identity eradicated by the increasing spread of culture from the metropolitan centers of the United States to its antipodes. In this interpretation William would make a fitting successor to Eb. A country boy himself, his migration to the polestar of the nation to work at one of its leading newspapers allows him to wax nostalgic on the life he left behind both in private and in print. Arguably, this specific latticework of nostalgia — watching as a particular identity passes and mourning it as it goes, rather than simply looking back fondly — is what separates *Eben Holden* as regionalist literature from Chesnutt and Harris's works of (pseudo) plantation fiction, despite the orthographic and content similarities Eben shares with their own storytellers. Eben would then once more resemble a "properly" Yankee figure, one like Jewett's Almira Todd. Though *The Country of the Pointed Firs* does not end with Todd's passing, it does end with the narrator passing from Dunnet Landing's shores. As she sails away she takes a moment to take one final look at "Dunnet Landing and all its coasts" only to realize that "all its coasts were lost to sight" (Jewett,

131). She leaves unstated, but wholly palpable, that she will never be able to visit it again — at least in the vanishing regional form she most values.

Despite the similarity in the form of nostalgia employed by *Eben Holden* and *Country* the same original sticking point remains. Eb simply does not tell the same stories as Todd. Todd's stories, perhaps typified by her relation of Poor Joanna's self-imposed exile on Shell Island, utilize only a tinge of the mythical aspects Eb tends to incorporate into his yarns. He also does not tell these tales in a similar voice, slanting more towards the Julius/Remus side of the spectrum. Eben is a chimera, a glitch that defies easy characterization as a "plantation" or "regionalist" storyteller. His presence forces us to extend Stephanie Foote's argument that "conflating regionalism's concerns with its formal properties" leads to the easy error of assuming that the work that regionalist literature undertakes is itself nostalgic.[126] Recent work, Foote's own included, largely argues that despite their various conceits, regionalist works are less paeans for disappearing local ways of life and more trojan-horse vehicles that consolidate a national identity by depicting local custom as inherently out of time and ineffective or as a means of policing which identities count as "American."[127] Adding Eb's specifically linguistic component, his orthographic crossing into territory that is not "his," brings a non-narrative, media component to this conclusion. Despite Bacheller's later claim to accuracy and specificity in *Stores* Eb's actual narrative function and substructural

[126] Foote, "The Cultural Work of American Regionalism," 27.

[127] See Brodhead, *Cultures of Letters* for the former point and Foote, *Regional Fictions* for the latter.

composition makes him more a generic. The system of knowledge orthography seems to represent — the ability to "peg" a subject to a position by region, class or race — turns back in on itself in a manner that mars its own purpose. The hyperliterate, the readers and writers of regional novels, are instead themselves afflicted by a particular form of orthographic being, the result of writing's attempt to hold on to this form of knowledge. Being technology and not essence, the surplus information non-standard orthography provides in its elisions and modifications reflects more the cacography of an advanced system of dialect literature than any particular subject. Orthographic substructural style allows creativity, but one that stems from the writer's own hidden notions of what a deviant orthography might look like, so conditioned by the previous orthographic systems they have encountered. Eb's position somewhere between Remus, Julius and his Yankee contemporaries justifies the inclusion of his particular orthography on his gravestone. Eb is nothing but writing, the marker of a literary system's inward turn as it busily produces the object of its knowledge for its own private consumption.

In the combination of his orthographic and narrative positions, Eb somehow transcends genre. As a means of compensation, the novel turns inward on itself. Rather than tarry in the north woods with Eb, William travels to city to — what else — write. Henceforth the novel becomes a *bildungsroman*, one that eventually celebrates William's harnessing of his generically regionalist youth, and thus Eb, as a means rather than an end in itself. After securing his position at the *New York Tribune* William joins Horace Greeley for a celebratory dinner. As they dine, William regales the great

commoner with tales from his sylvan upbringing, a theme familiar to Greeley who shared a hometown with William's adoptive father. Greeley takes the opportunity to philosophize on the advantages the countryside holds over the city, claiming that only in the city does the "lie have many forms" and eventually concluding that the "great cities" all suffer from the sinful trifecta of "Vanity, Flattery and Deceit" (Bacheller, 335-336). Greeley assaults city life on the basis of verity. His triad of vices all rely on narrative prevarication, leading to a proliferation of subjects "pretending to be what [they aren't]" relying on the "many forms, unique, varied, ingenious" that the city provides (ibid). In contrast, finding "truth", embodied in "men... genuine, strong and simple," requires getting "back into the woods" (ibid). In this way, Bacheller's version of Greeley freely offers his own take on the convergence of region, identity and nostalgia. Greeley envisions regional identity as reducing the complexity of his own task as a writer. Those who speak in the tones of the backwoods are knowable subjects. Their voices assure their regional identity, allowing them to be fixed to the page as known subjects more readily than the cosmopolitan dwellers of the city who can speak in many tongues and adopt many identities. William shies away from accepting this account wholeheartedly. Without naming names, he states that there is "no Eden there in the north country" and that it holds plenty of liars (338). Despite his reluctance to identify any liar in particular, his reverence for Eb and tacit endorsement of the testament to Eb's honesty on his tombstone likely disqualifies Eb from that set of northwoods individuals. Even though Eb tells tall tales, he himself is not one. In general William agrees with Greeley, but his position takes a formal twist. The types of tales city-dwellers might tell seem to concern

identity. According to Greeley they use their cosmopolitan repository of forms to conceal the very truth of who they are. Even though Eben, in his own way, spins lies, he does not use them to obfuscate his self-narrative instead. Indeed, quite the opposite. Eb's wild tales of liminal animal beings actually reinforce external appraisals of his subject position by offering an "objective" measure of who he is in the form of his recorded orthography. The content matters less than the media-embodied form his tones take, a lasting record which becomes the truth of Eben's upbringing, life and self through analysis and comparison to other groups of recorded speakers, even simply across an unsystematic lifetime of reading.

William takes Greeley's advice mostly to heart. He records Eb as a function of his role as narrator, but he also uses the technique diagetically. Upon moving to New York City William secures lodging above the establishment of an elderly sailor-turned-shopkeep named Riggs. William depicts Riggs as a singular character from the first, hinting at the irony that a man blind with age should own a lantern shop and being sure to recount Riggs's Platonic musings on the dreamlike nature of the world, a temporary veil that conceals the true realities of "God and love and Heaven" (319). By the end of the chapter William reveals that he has used Riggs as the basis of his "first tale," a "brief account of what [he] had heard and seen at the little shop that evening" (320). Not content to share it solely with us, his retrospective audience, he sends it off to the *Knickerbocker* to be considered for publication the very next day. In including both this encounter and the moment of its recording Bacheller doubles the

purpose of this one small moment. The story of Riggs serves as both proof of mastery of the sort of writing typified by the novel itself and a pseudo-intertextual moment wherein William cites his own magazine-bound tale. Beyond just being a cheap moment of self reference, this moment of self-citation serves to define the power of writing through negation. Simply put, the *Knickerbocker* refuses this initial version of the tale, returning it to him with "ready-made thanks" on a preprinted slip leaving William himself "firmly, thankfully rejected" (321). The slip itself girded by the system of alphabetic knowledge it represents prevents William himself from identifying as the particular type of subject that can enter new truths into the written record. Its inky permanence becomes for William what Eb's particularly rendered orthography is for him — a marker of what kind of subjectivity the bearer holds. Only after Riggs finally wakes from his dream of life and enters the Platonic realm of death does William's tale earn validation. After discovering the lately deceased lamp-seller William claims that "his story of Riggs was now complete" leading to its subsequent publication "because it was true" (347). William does not explicitly relate making any emendations to the original story, implying that the death itself was the most vital change. Riggs ends up in his desired realm of Platonic forms, just not quite the one he expected. With his death he seals the "truth" of William's account and becomes a recorded typology, an immutable subject-form available to written knowledge modes of comparison and categorization. As the (anonymous) recipient editor of the story puts it "All good things are true in literature" — an odd choice of construction when placed next to the more conventional "All true things in literature are good" (ibid). The editor's particular version of this adage credits

the medium with the production of truth, rather than the narrative itself. The act of recording transforms the good of the oral/ephemeral world into the true of a written discourse that allows for comparison and identification across the time while also co-creating a subject with a fixed being that underwrites the truth of the medium. Writing does not just produce *a* truth — according to the editor (and through tacit agreement, William) it produces a specific notion of truth itself.

The final truth of Eb's gravestone recapitulates the manner in which William fixes Riggs to a certain narrative type, albeit in a more stonily permanent medium. William's narratorial rewriting of the epitaph's writing of Eb monumentalizes William's life just as much as Eb's, marking the moment he achieves full mastery over writing as a form of knowledge. Eb's disappearing act that began with William's legal majority reaches its logical end — Eb becomes written knowledge and the novel becomes a bildungsroman. Sitting in contemplation over the "perished forms" of his predeceased friends and family, William takes the opportunity to send a literary form to the charnel house as well by exorcising the ghost of Remus/Julius style cyclical narratives from the machine of written fiction (431). This moment recapitulates the narrative of development Nadia Nurhussein associates with the developing school system of the era. William initially commences his writerly education by taking lessons in dialect from Uncle Eb, but concludes by graduating to a "higher" level of standardized writing that claims the ability to crystallize and comprehend subjects like Eb. The importance of banishing Eb specifically is highlighted by the contrasting way in which Bacheller treats Jed Feary, the

other dialect-spouting tutor of William's youth. Feary plays the role of vernacular poet, regaling anyone who will listen with occasional compositions tinged with his own backcountry lilt. The apex of Feary's literary career has more to do with William than with the perceived quality of his own compositions, coming in the form of an ode to William on his removal to college. Feary means for his ode to be purely occasional. Although he produces a manuscript of the (quite lengthy) composition, he delivers the work in oral form, reading it to those assembled to see William off. This produces a situation isomorphic to the final recording of Eb's epitaph, featuring William remediating the ephemeral into the print material of his novel. The difference between the two moments inheres in William's ends rather than his means. Feary's composition serves both as an ode to William and a remonstrance concerning his future conduct in the wide world apart from the glades of northern New York. Feary uses the local figure of "Aunt Samanthy Jane" to drive home his point about the dangers of overvaluing worldly things at the expense of faith in God and the comforting knowledge it brings (226). Specifically, Feary denounces the power of the "lens er rule o' cipherin" to know the "soul" of the world, its ultimate truth (225). Embracing this rude Platonism allows such fortunate Samanthy Janes peace, no matter where "stormy Jordan flows", even if it is, in her case, flowing in the direction of the poorhouse (226). Feary's recitation parallels Eb's moments of storytelling in a few significant ways — he presents a narrative in dialect, if only a minor one, and does so in order to advance a viewpoint on the nature of the world. Despite these similarities, William dismisses Feary's wisdom directly after relating it, telling the reader that he "give[s] this crude example of rustic philosophy not

because it has [his] endorsement" but rather because "it is useful to those who may care to know the man who wrote it" (226). Similarities aside, Feary's numerous differences from Eb make his form of orality less threatening than Eb's wild cosmologies, and William simply claims the dominance of writing over him. Feary, William claims, is simply a rural type and merely recording his doggerel poem composed in nonstandard English demonstrates this fact. Feary's content level deviations from Eb's style of storytelling provide some justification for this straightforward dismissal. Unlike Eb, whose presentation of human—animal intimacies calls in the base fundamentals of the type of underlying subjectivity writing means to secure, Feary chooses to buy in to the Platonism of subjects. His poem's treatment of "Aunt Samanthy" resembles the way William attempts to handle Eb, presenting her as a "good old" regional type whose own local subjectivity should provoke reflection on the "mind", the "part o' God's creation" available regardless of context (226). This investigation of interior space assumes the same underlying fixity of mind, guaranteed by an ultimate Platonic God, that William then applies to Feary, using Feary's own demonstration as proof of its truth. William assures us that there can be no confusion here; Feary is straightforwardly a northern New York poet, and his words tell you that on simple face value.

Even more importantly, Feary's accent corroborates the solidity of his subject position. A major component of Eb's singular presence in the text is the unusual nature of his orthography, specifically its proximity to the African-American narrators of plantation literature. Feary ranks as incredibly dissimilar from Remus and Julius, being

about 1900th most similar to the former and 2040th to the latter. While Feary's dialect orthography itself does not overly resemble that of someone like Jewett's Almira Todd, his association with other "backwoods" characters (with top matches including Carrie from *Sister Carrie* and Ethan Frome, among others) locates him solidly in the tradition of oral storytellers used by a Jewett or Gale. Bacheller draws upon their regionalist toolkit in a way he (consciously or not) does not with Eb. Feary, like a Todd, intones in the voice of his region, as defined by the literary system surrounding him. He also, like a Todd and very much unlike Eb, stays locked in his context, disappearing when the narrative moves outside of its confines. He simply disappears from the narrative, with a note after the ode that he has "long passed the praise or blame of this world" (226) being his brief memorial.[128]

All of this stands in stark contrast to Bacheller's treatment of Eb, who is not so much praised (or remembered nostalgically) as buried. Unlike Feary, or for that matter Julius or Remus, Eb's final subjectivity must be insisted upon. His linguistic assault on orthographic accuracy, its companion genres, and their shared ability to identify the real forces a response. As Eb's creator, Bacheller is ultimately responsible for both this assault and its resolution. A thoroughly intention—bound reading in which Bacheller uses this seeming paradox to narrate the drama of orthography's fate would prove

[128] Though this moment memorializes him, it is not Feary's last appearance. He graces the novel with one more brief composition — but only re-enters when William returns home to northern New York for Christmas. Bacheller, *Eben Holden*, 401.

unsatisfying. However, the bare facts of orthographic production and interpretation prevent such a reading. Bacheller produces Eb's dialect system, but he does so from a perspective much the same as those who will eventually consume it. Creating Eb's particular accent requires Bacheller to inspect his own nonconscious embodiment of orthography from a first person perspective and combine what he finds there with invention on the phenomenological level. The resulting amalgam reflects less the reality of a certain northeastern subject position, and much more the reality of the situation Bacheller finds himself in. The noise of the world — the influence of innumerable texts and other media, seeps in from the lower levels and diverts Eb's narrative from any initially planned point. The attempt to employ orthographic meaning as a repository of stable knowledge only returns the hint that the very concept it seeks to buttress — the type of subject implied by writing as a medium—simply does not exist. The bildungsroman aspect of Bacheller's work, a deviation from previous texts with Eb-like characters, is itself diverted by this truth. William's reward is not access to truth, but the ability to create it as a certified functionary of the print universe. He does not record truth, instead he writes things that become true because they are good. The lessons of non-standard orthography were supposed to turn William into a writerly subject, but instead they may have revealed that this has always been nothing at all.

The Unlife of Orthography

Eben Holden cannot state this truth, even implicitly. In its final turning inwards on itself it chooses to shift the goalpost of writing's hold on the real. Mastery over subjects becomes mastery of the system itself, and Eb becomes another "good thing true in literature." Without the privileges afforded by a claim on some external reality, the attractiveness of a knowledge system begins to fade. This descent quite possibly contributes to the diminishing popularity of the traditional regionalist genres in the early twentieth century.[129] Arguably, the conspirators of text, the newspaper and magazine men like Bacheller who pursued the truth of the other through orthographic practice, simply lost their hold on the real. As the forest of media-forms grew more and more verdant, they could no longer see the forest for the trees. In addition, new literary forms claimed the technique of non-standard orthographic substructure for their own ends. Two accounts, the generally similar theses presented by Michael North and Houston Baker, argue that dialect strays from its popular roots to become the very basis of modernist experimental literature.[130] Of course these texts do not wholly resemble early dialect works. They rarely employ stable dialect narrators, a cast of characters

[129] As Foote puts it: "The heyday of regional writing was roughly between the Civil War and the early years of the twentieth century." Foote, "The Cultural Work of American Regionalism," 28.

[130] See Baker, *Modernism and the Harlem Renaissance* and North, *The Dialect of Modernism: Race, Language, and Twentieth-Century Literature*.

"tagged" with non-standard orthographies, or the cyclical structures that hold common in the regionalist era works. They instead transform substructural orthographic style, utilizing interpolation and quotation in order to smuggle non-standard orthography into "high" literature in zombie form. In this guise, the effects of orthographic dialect served as just that — an artistic effect — a form of style rather than a form of truth. These "high literary" successors adopt Eb's final truth as much as they do Bacheller's techniques, embracing substructural style as the form of inward literary reference it is.

The spread of orthographic similarities in the later period of the corpus as demonstrated in **Figure 4-7** shows that emerging "high" literary discourses were not the only ones to take advantage of dialect orthography's untimely demise. North and Baker's analyses of the use of dialect in the modernist avant-garde details one account of the "afterlives" of nonstandard orthography, but its use as a technique in the early twentieth century was not solely restricted to texts with an experimental agenda. Popular novels in the more or less regional vein continue to appear from the pens of authors like Edna Ferber and Gene Stratton-Porter. The Ring Lardners, George Ades and Peter Finley Dunnes of the world also continued apace, producing short self-contained newspaper sketches in "slang." The arguably new formation of "genre literature" also adopts non-standard orthography, with texts like Zane Grey's pioneering Western *Riders of the Purple Sage* employing it liberally. Even the American naturalists join in, freely peppering their works with dialect orthography in search of some form of reality function, and eventually the technique gets re-adopted by people of color for use

in the emerging minority literatures of the early twentieth century. As the 1890s spill over into the new century the use of non-standard orthographies seems to actually expand rather than contract — at least in one sense. However, I do not say this to challenge the critical line that dialect literature in its "original" nineteenth century form somehow diminishes as the new century commences. Dialect orthography remains popular in the new century when measured by volume, but Bacheller's handling of Eb typifies the diminished amount of explanatory power granted to it as it wends it way through its new generic homes.

Even so, surprisingly direct echoes of "high" regionalism persist even past the first World War. Despite its usual classification as horror or "weird" fiction, "The Shadow Over Innsmouth" bears more than a passing resemblance to a doppelganger version of the earlier works examined above. Lovecraft's story follows the narrator as he embarks on a genealogical tour of New England that culminates in an unexpected — and severely hostile — stay in the titular shambling and degraded port town. There, as in many Lovecraft stories, the narrator locks horns with the local cult, this one composed of fish-human hybrid followers of the water-god Dagon, before he finally undergoes the terrifying genealogical realization that he too springs from the same stock. This hybridity certainly stands in for Lovcraft's racist fear of miscegenation and obsession with purity. However, with the context of the earlier works examined above "Innsmouth's" particular approach to the subject, including its particular racialized viewpoint, continues their particular thread. As a locale, Innsmouth closely resembles the sort of fading New-England seaside port town typified by Jewett's Dunnet Landing.

Both towns have shuffled into their respective modern eras, burdened by the material signifiers that until recently demonstrated their value as portals to the wider world. To the modern eyes each work's narrator, these signs of faded cosmopolitan wealth read only as bizarre trinkets smuggled into a rapidly changing world. Such items — Elijah Tilley's memory-worn tea set from Bordeaux, the intricate south-sea crowns and jewelry favored by the Innsmouth town-folk — reveal little more than the disordered subjectivity of their owners and the environs they inhabit (Jewett 123).[131] In Lovecraft's hands, the nostalgia for the regions "left behind" transforms into horror.

Lovecraft, like these early writers, also peppers his story with dialect orthography using oral tale-tellers in the mold of Julius, Remus and Eb. Unlike most of the other texts in the corpus, "The Shadow Over Innsmouth" employs dialog sparingly. Lovecraft only endows a handful of characters with quotation-demarcated speech, letting the internal discourse of his narrator provide the rest of the story. However, the moments of direct speech he does employ are, invariably, profuse and non-standard. The drunken, aged ex-seaman Zadok Allen serves as Lovecraft's version of Captain Littlepage, one of Jewett's tale-tellers, down to the relative similarity of their orthographies (see **Table 4-3**).

Allen passes along a tale just as wild as Littlepage's relation of his journey into the frozen land of souls, even if the revelation that the rumors of the adoption of eldritch sea-god worship in Innsmouth are both true and very much real comes with more

[131] Lovecraft, *The Call of Cthulhu and Other Weird Stories*, 276.

immediate danger than Littlepage's pseudo-religious pontifications. The comparisons of *Pointed Firs* and "Innsmouth," Littlepage and Allen, are surprisingly, not of apples to oranges (or seafloor abominations to spiritualist apparitions). Rather, "Innsmouth" is at least in part the post-lapsarian version of *Pointed Firs*, the uncanny familiar of a regionalist text in a world now rife with the destabilizing presences dramatized in Bacheller's final treatment of Eb. Jewett's narrator treats Littlepage with kid gloves, not exactly disputing his story but not explicitly crediting it either. The possibility that Littlepage's orthography indicates a certain form of regional subject, the wild Yankee sea-tale teller, prone to fanciful untruths, remains fully open. However, after Eb cleaves the coupling of subjectivity, tales of the nonhuman, and orthography, Zadok must be taken seriously. Although Zadok's regional dialect (and purported insanity) tempts even the narrator to simply deem him another cracked Yankee, prone to "philosophizing in a sententious village fashion," all he says proves true (294). Zadok exceeds the markers of his subjectivity, short-circuiting the pathway between an "objective" marker of who he is and the very notions of the self that would allow any such evaluation to make sense.

Zadok's terrible revelation of the elder gods and their half-aquatic Innsmouth worshipers sits upon girding forged from a deeper truth of the post-Eb world. By the end of his story, Lovecraft shows that even the narrator has been bowing to his own unknown self the whole time. Rather than being what he might think he is — a human subject, a narrator relating some odd goings on in a queer and backward Yankee port town — his genealogical research reveals that he is a descendant of the Innsmouth

Marsh family, the originators of the Innsmouth fish-people, and thus, no human subject at all. Like the rest of Marshes he is part Dagonic fishman, a terminal form of non-human subjectivity that eventually overwhelms whatever human part remains. His story is subjectless, an emanation that reveals nothing at all about the truth of the speechless, gilled being that has lain beneath the veneer of "reputable Yankee scion" the whole time. Lovecraft makes explicit the fear that Bacheller begins to unearth through Eb's strange tales and even stranger orthography — that writing may not have the subject it claims to have at all. Even something like orthography, a facet of writing that appears to lay bare and codify the underlying processes that produce varied sounds of speech, cannot guarantee the transfer between the first person experience of a communication and the non-phenomenological neural inscriptions that undergird it. Lovecraft's exploitation of this insight pursues it to its absolute conclusion by positing the extinction of the subject as such. As the stages of fish-human hybridity progress the linguistic capabilities of those so endowed progressively atrophy, calling into question whether the narrator's musings on the unusual nature of the Innsmouth folk based on their habits and speech ever had any purchase at all. The narrator uses his final moments of lucidity to predict such a fate for us all, while at the same time hinting that this void has always been humanity's birthright — as much an ominous reference to life's emergence from the sea as it is a statement about the impenetrable depths that dwell below the manifest image of the subject.

The only bitter palliative Lovecraft offers comes in his choice of form. Just as *Eben Holden* supplements the typical regionalist formula with an injection of bildungsroman in order to mitigate the fears associated with writing's shift in epistemic status, "Shadow" manages to cram the cosmic horror of the great outside into the confines of a sort of bio-bildungsroman. Lovecraft's narrator undergoes a thoroughly singular coming of age with its own concomitantly extreme form of puberty. The narrator's physical changes, both neuronal and morphological, imply a reading of the form itself — when text and self are simply one medium among many all bildungsromans have always been bio-bildungsromans. Understanding these changes through their manifestation in narrative takes a backseat to non-phenomenological approaches. Lovecraft, as befits his racial ideologies, analyzes his subjects using a form of pseudo-scientific proto-genetics. The Innsmouth self moves generationally, leaving the narrator to wonder "if he is coming to resemble my grandmother and Uncle Douglas," both carriers of the trait (334). Even though he deduces that he is in fact undergoing the transformation, simply knowing this does not clue him in to the type of subject he will become. He initially plans to pyrrhicly preserve his humanity through suicide, but as his aquatic nature takes hold through unending, uninterpretable dreams he begins to awake from them with "exaltation, not terror" before deciding to embrace his nature and dwell below the sea "amidst wonder and glory for ever" (335). The proto-genetic aspect provides some certainty about the subjectivity the narrator possesses, but realizing its origin, progression and implications does not predict the subjectivity he eventually demonstrates. Pseudo-scientific discourse inherits the

epistemological mantle orthography once held, but wears it with a smile made grim by the qualified aspect of the knowledge it provides. In this odd way the epistemological void left by the reposition of print as just one medium among many others helps birth "weird fiction" as a partial generic successor to regionalism. Bereft of the certainty that simply inspecting orthography phenomenally can detect the contours of unseen subjectivity, it seems only horror remains.

Lovecraft's strange brew of horror, racial fear, and genre/orthographic reference finalizes the list of uncertainties *Eben Holden* began to enumerate. The regionalist taxonomy of human types flowers strange fruit when fed on the light of an environment where accents themselves shift and multiply in unsystematic ways. The genres that bloom as a result of this upbringing inherit the formal mission of orthographic taxonomy, if not the power it once held as a technique. Rather than writing the "real," or even welcoming the cultural noise that this real emerges from, texts like Lovecraft's embrace the potential emptiness of any form of circumscribed conscious or unconscious self, asking whether the processes that generate these noises and their encodings are comprehensible to a phenomenally-oriented culture of writing at all. William Brower and Lovecraft's narrator find themselves in oddly analogous situations. Staring into the orthographic void at the heart of their somewhat off-kilter regionalist narrations drives them away from the knowledge of other selves produced by more typical works of this genre. For both, their wild flight through generic norms turns inward. The endpoint of William's bildungsroman-as-writer provides some socially holistic palliative. Even if

some version of the system itself is all that lies at the bottom of the orthographic rabbit hole, this at least can be mastered and from this mastery produce a truth of its own. The semi-conscious production of orthography, a technology conditioned into each individual by their cultural system of writing as a whole, is at least a sort of system after all, even if its phenomenally invisible constituent parts only refer to the system itself and not some form of specific subjectivity. On the other hand, Lovecraft's narrator accepts becoming the non-human thing that writes as the end of his bildungsroman. The horror Lovecraft intends his narrator's eventually joyful conversion to provoke is a conservative (as it of course would be) step beyond. For Lovecraft, a part of the self becoming an inscribable medium itself is quite unbecoming. Discovering that the noise of a cultural technology resides within unseen and uncontrolled by the phenomenal self is a threat, a contagion that stains being, one that deserves to be warded away by references to racialized science and bloodline degeneracy.

Eben Holden is a previously unnoticed bridge between the world of Todd and Julius and the world of Zadok Allen. All three of these groups embrace the ends sought by regionalism's use of non-standard orthography, each with differing affect and implementation. *Eben Holden*'s own status as a somewhat atypical genre-bending regionalist novel lets the very gap it spans becomes visible. The comparison of an "Innsmouth" and a *Country of Pointed Firs* is only skin deep without the orthographically strange figure of Uncle Eb to solidify the bond. Bacheller's novel thus provides a new version of the history of regionalist literature, one mediated as much by

orthographic substructural style as by genre and subject. Eb's final scene, the bizarre decision to engrave lines in non-standard orthography on his tombstone, is strangely appropriate in this light.[132] These lines do not, as we are tempted to believe, record Eb, or even mourn him. Rather, they serve as a testament to the self-referential nature of the orthographic real, a discourse that makes sense all on its own.

[132] Bacheller later revists Eb in a short follow-up novella. See Bacheller, *Eben Holden's Last Day A-Fishing*. This work has been commented upon even less than *Eben Holden* itself, perhaps because it seems to be written mostly as a memorial volume in honor of A. Barton Hepburn, a New York state assembly member who was also possibly one of Bacheller's professors at St. Lawrence College. Thankfully, Bacheller avoids the temptation to retcon Eb's passing, instead choosing to set the novella during an unused period of his later years. The theme is mostly one of Christian comfort.

Figures

Figure 4-1. This plot is the PCA reduction of the stochastic matrix models of characters drawn from the corpus texts ranging from 1868-1930. Plotting the models in this fashion produces rough, visually available clusters.

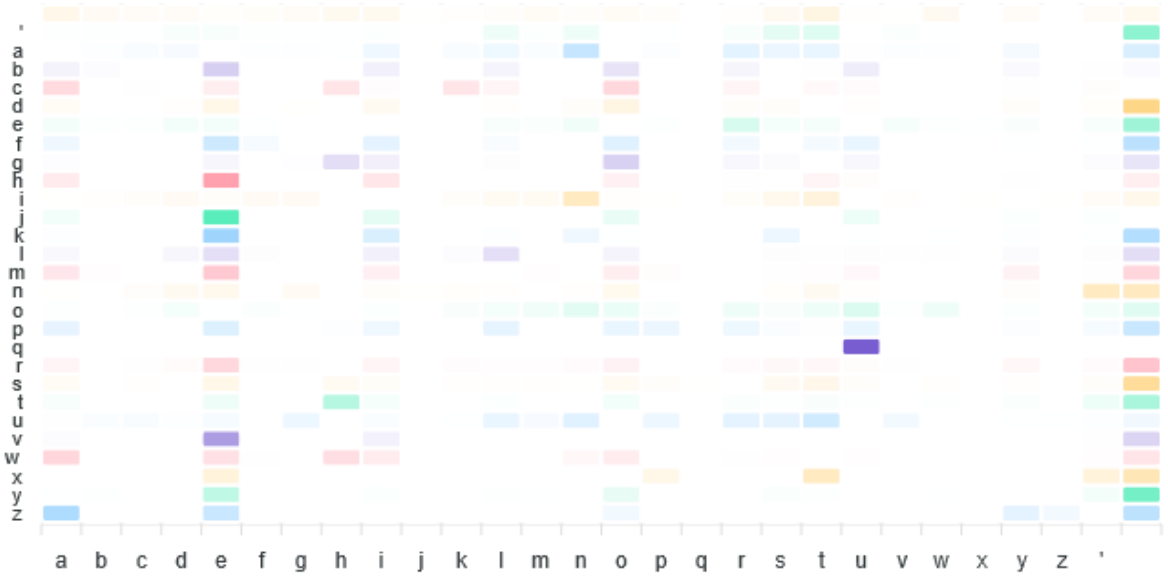

Figure 4-2. A heatmap that visualizes the stochastic matrix model associated with *Eben Holden's* Uncle Eb. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme.

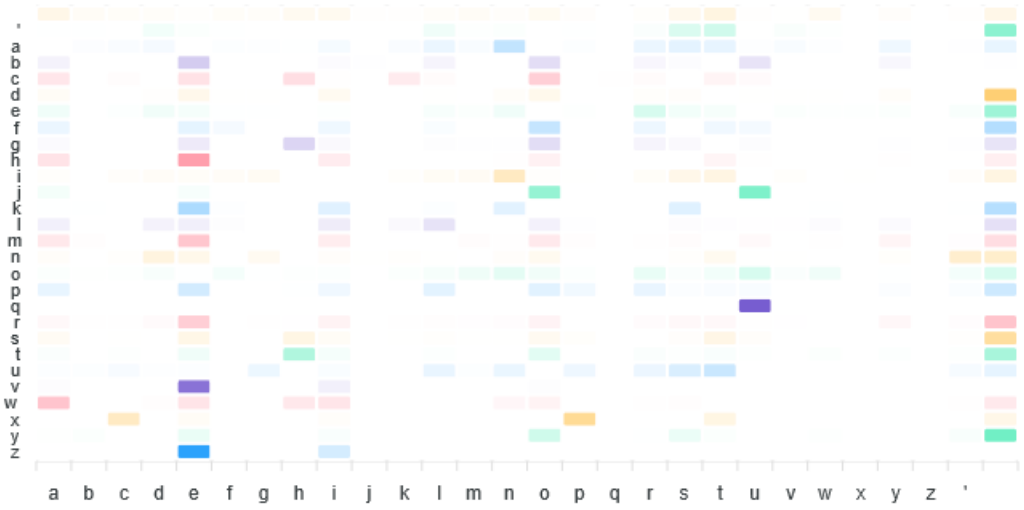

Figure 4-3. A heatmap that visualizes the stochastic matrix model associated with *The Country of Pointed Firs'* Almira Todd. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme.

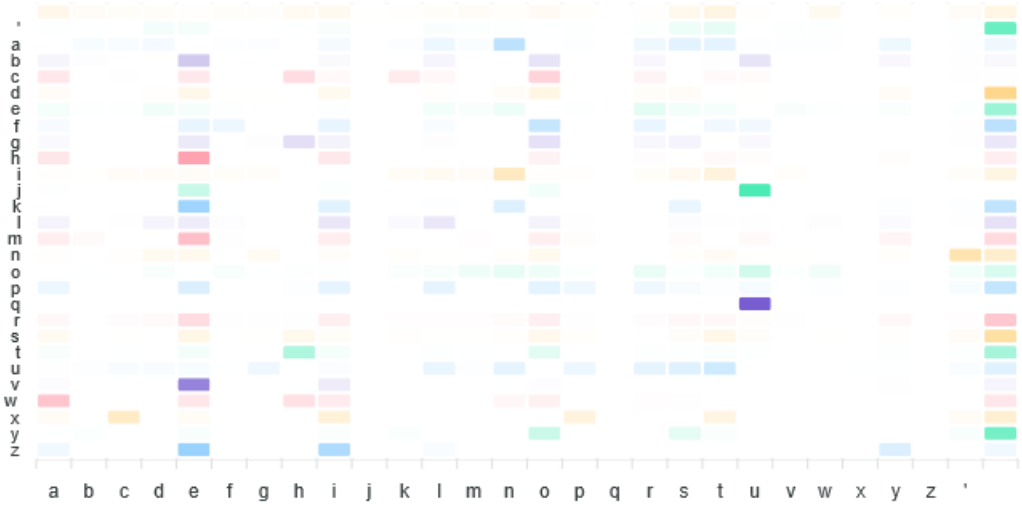

Figure 4-4. A heatmap that visualizes the stochastic matrix model associated with *The Country of Pointed Firs'* Almira Todd. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme.

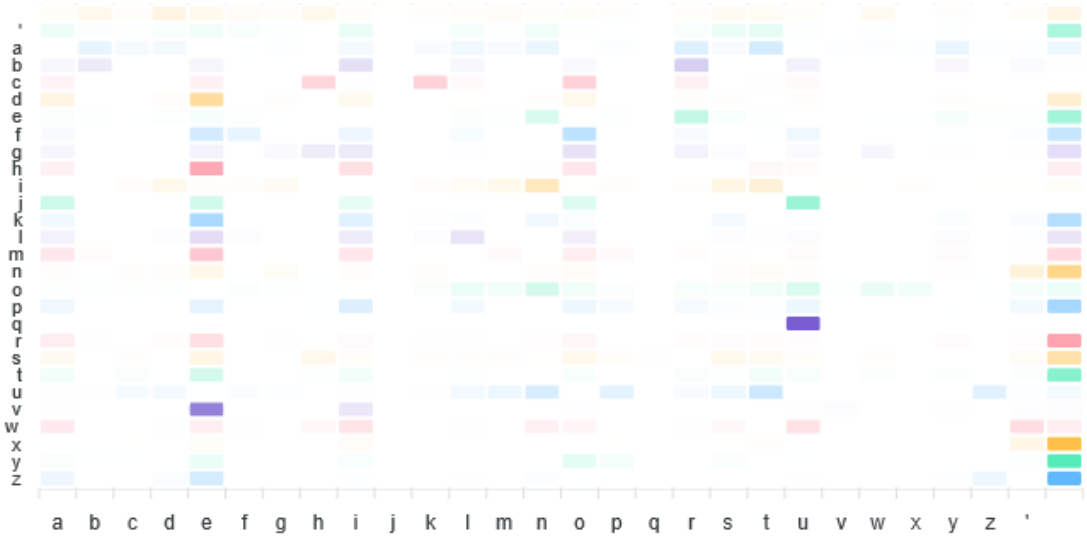

Figure 4-5. A heatmap that visualizes the stochastic matrix model associated with Joel Chandler Harris's Remus character. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme.

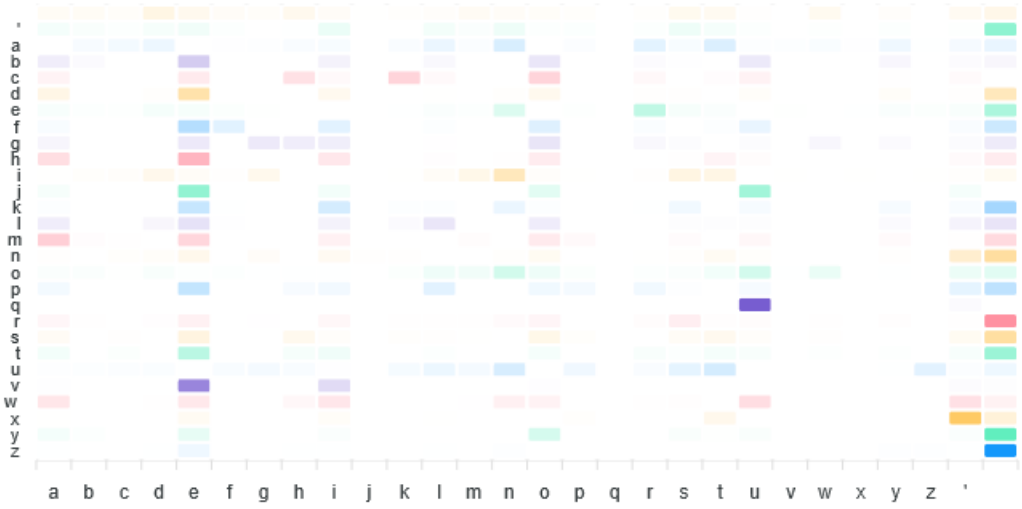

Figure 4-6. A heatmap that visualizes the stochastic matrix model associated with Charles Chesnutt's Julius character. Darker shading at the intersection of two graphemes indicates that the row grapheme often predicts the appearance of the corresponding column grapheme.

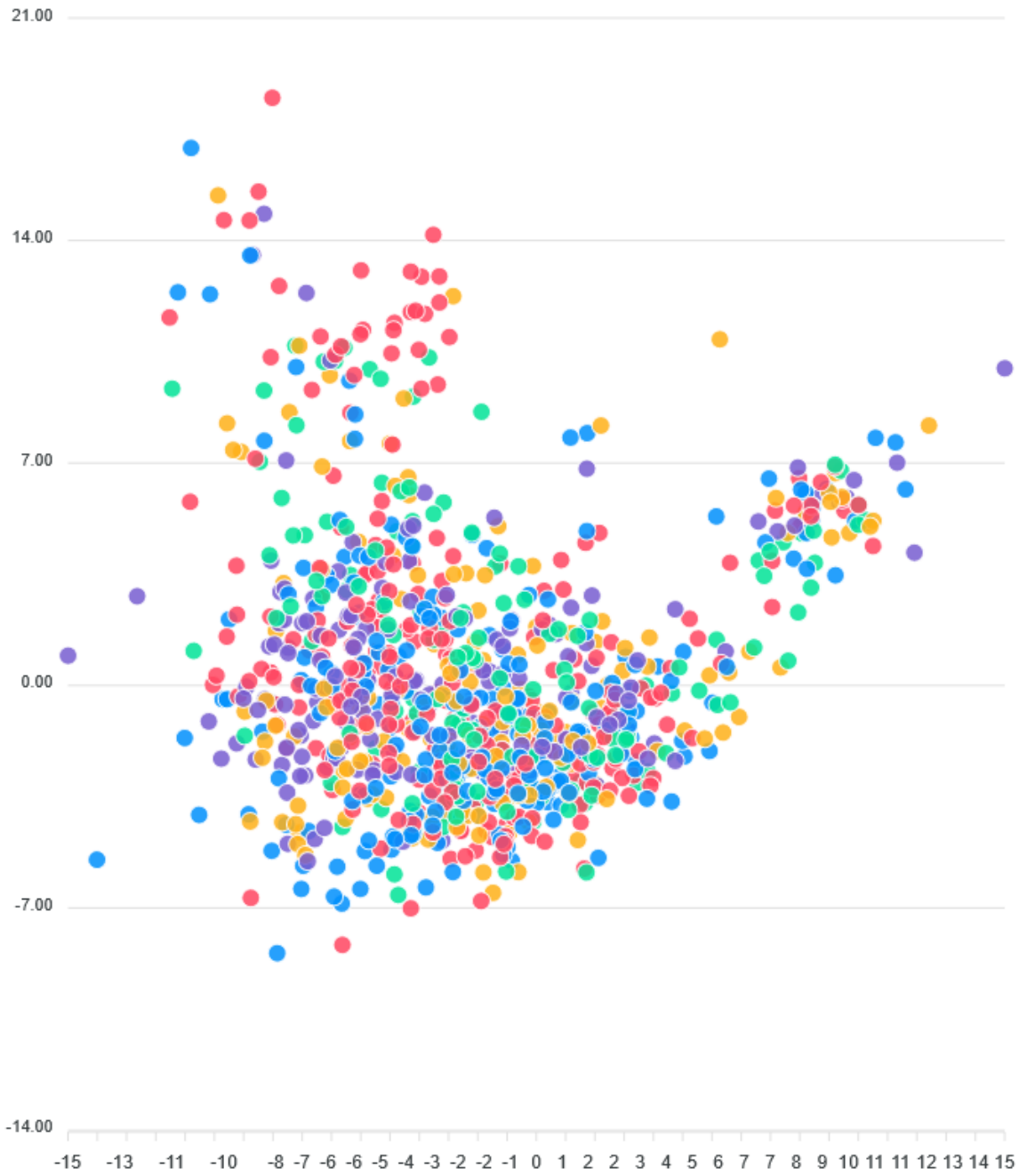
21.00
14.00
7.00
0.00
-7.00
-14.00
-15 -13 -11 -10 -8 -7 -6 -6 -5 -4 -3 -2 -2 -1 0 1 2 2 3 4 5 6 7 7 8 9 10 11 11 12 13 14 15

Figure 4-7. This plot is the PCA reduction of the stochastic matrix models of characters drawn from the corpus texts ranging from 1890a-1930. Plotting the models in this fashion produces rough, visually available clusters.

Tables

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Eben Holden: A Tale of the North Country	Dave Brower	11.03	11.63	10.43	1.12	0.38
Friendship Village	Peleg Bemus	11.03	11.69	10.37	1.13	0.54
Eben Holden: A Tale of the North Country	William Brower	11.05	10.66	11.43	1.46	0.43
Friendship Village	Calliope Marsh	11.08	11.25	10.9	1.05	0.57
The Country of the Pointed Firs	Todd	11.13	10.66	11.6	1.65	0.38
Song of the Cardinal	Abram	11.31	10.78	11.83	1.41	0.65
A Little Norsk; or Ol' Pap's Flaxen	Anson Wood	11.37	10.81	11.93	1.46	0.52
Riders of the Purple Sage	Jim Lassiter	11.37	12.1	10.65	1.65	1.3

Cap'n Eri: A Story of the Coast	Perez	11.37	10.8	11.94	1.34	0.7
Cap'n Eri: A Story of the Coast	Eri	11.37	10.96	11.78	1	0.7
A Little Norsk; or Ol' Pap's Flaxen	Bert Gearheart	11.38	10.38	12.37	1.55	0.98
Jerome, A Poor Man	Ozias Lamb	11.4	10.65	12.14	1.68	0.86
Fishin' Jimmy	Fishin Jimmy	11.4	9.98	12.81	1.85	0.49
Gabriel Conroy	Gabriel Conroy	11.43	11.56	11.31	0.98	0.65
Tiverton Tales	Nicholas Oldfield	11.49	10.47	12.51	2.26	0.64

Table 4-1. A table collecting the characters that score as most similar to Uncle Eb from *Eben Holden: A Tale of the North Country*. The table is sorted by the third column, perplexity difference. A lower score indicates that the character's orthography is more similar to Eb's.

Character 1	Character 2	Perplexity Difference	Perp. Divergence Base 1	Perp. Divergence Base 2	KLD on '
Uncle Eb	Remus	13.23	12.97	13.5	0.68
Uncle Eb	RemusCharacter	15.01	14.63	15.4	1.71
Todd	Remus	14.54	13.64	15.45	1.56
Todd	RemusCharacter	16.16	15.19	17.12	2.8
Julius	Remus	11.23	10.85	11.61	0.32
Julius	Todd	14.61	12.7	16.53	1.75
Julius	Uncle Eb	13.02	13.31	12.72	0.75

Table 4-2. A table collecting comparisons between a handful of characters – Charles Chesnutt's Julius, Joel Chandler Harris's Remus, Irving Bacheller's Uncle Eb, and Sarah Orne Jewett's Almira Todd.

Model 2 Filename	Model 2 Character Name	Perplexity Diff	Perplexity Div 1 2	Perplexity Div 2 1	KL	KL On Apostrophe
Friendship Village	Calliope Marsh	11.32	11.27	11.38	1.01	0.59
Tom Sawyer Abroad	HuckN	11.34	10.56	12.12	1.42	5.02
Deephaven	Captain Sands	11.42	10.31	12.54	2.08	2.92
The Country of the Pointed Firs	Todd	11.47	11.17	11.76	1.96	0.72
Friendship Village	Peleg Bemus	11.52	10.94	12.1	1.46	0.67
Fishin' Jimmy	Fishin Jimmy	11.55	10.3	12.79	1.78	0.49
Sweet Cicely – Or Josiah Allen as Politician	Narrator	11.6	11.39	11.8	1	1.1
The Country of the Pointed Firs	Littlepage	11.6	10.94	12.26	1.84	1.02
The Deerslayer: or The First Warpath	Hurry Harry	11.69	11.36	12.02	1.84	3.2

Sonny, a Christmas Guest	Narrator	11.69	11.96	11.42	0.92	0.65

Table 4-3. A table collecting the characters that score as most similar to Zadok Allen from "The Shadow over Innsmouth." The table is sorted by the third column, perplexity difference. A lower score indicates that the character's orthography is more similar to Allen's

Orthographic Extrema

Previous sections of this project have operated somewhere in the middle space between the "corpus-driven" and "corpus-based" poles elucidated in the general introduction. The provocation to understand Twain orthographically stemmed from the previous "top-down" discourse surrounding his use of non-standard English; the focus on *Eben Holden* emerged from the specific desire to understand a particular corpus feature (Eb's relatively oddly positioned orthography) in the light of of the top-down formation of genre. The advantage of mixing perspectives is twofold. On the theoretical level, such an approach respects the cognitively multiple nature of substructural style. As has been shown, the interpretation of orthographic meaning draws on both the contextually developed nonconscious processes associated with decoding and processing as well as "top-down" information from semantic meanings in the text at hand, or other previously held cognitive investments. Each of these poles roughly corresponds to the critical techniques that motivate studying a corpus in a "driven" or "based" fashion, and combining the two allows the previous studies to wander, if guardedly, in the clouded regions that obscure the connection between the two. Methodologically, such a mixed approach avoids error derived from the computational processing of orthographic sequences. Even applying a simple measure like the information-based ones used here is a transformation of the data from its native form, and will not capture all of the complexities it contains. Having a "top-down" component at hand helps guard against falling into computational mirage, and offers a check

against more routine sources of error (say, a systemic mis-attribution of a certain character's dialogue samples).

Be that as it may, there are some questions more suited to a bottom-up approach, where statistical discovery acts as the primary driver of interpretation. Operating chiefly on this level, where context is "baked in" to the measures themselves by virtue of their comparative nature, allows for conclusions that characterize more far reaching notions less available to phenomenal introspection. The subject of this chapter, orthographic extrema, is arguably one of these notions. Having such a wealth of information available foments the desire to apply superlatives, determinations of which character in the corpus has the most unique orthography, or the most random one, both questions that, in theory, could be answered without the additional traditional critical attention that was paid to the previous texts.

This study has resisted such an approach, both due to its theoretical commitment to the interplay between levels of orthographic meaning, and its desire to avoid the methodological pitfalls described above. Compound measures derived from the pointwise comparisons of our individual models (of which PCA is one) come with the additional methodological downfall of making them even more sensitive to input length. For example, one simple measure that would quantify how distinct a character is from the rest of the characters in the corpus would work by averaging all of one character's perplexity or KL comparisons to every other character. However, this compound measure would inherit the accumulated error in every single comparison, exacerbating

it by a factor equal to the number of other characters in the corpus. Practical solutions to this specific issue likely exist — say looking for outliers among groups of characters with similar dialogue length — but are still inelegant and potentially misleading.

Avoiding this sort of pitfall requires some amount of compromise in the form of extreme chosen for analysis. Difference from the rest of the characters in the corpus is but one potential extreme. Given its methodological issues, the simplest response might simply be to approach the problem through a rethinking of what "extreme" might mean. Below is an experiment that approaches the orthographic extrema of the corpus in a bottom-up fashion. Specifically, it employs the average entropy approach briefly elucidated upon in the methodological introduction. This experiment is twofold; it is both an exercise in determining what a particular measure, selected from the bottom up, might reveal, and then applying that conclusion to some of the extrema that emerge.

Extremity and Randomness

The "average orthographic entropy," as noted in the methodological introduction, averages the Shannon entropy of each individual row of the stochastic matrix associated with a given character. As a reminder, each matrix is associated with one character, and is built from the character pairs that make up their in-text utterances. Each row is a probability distribution giving the likelihood that this character will use the particular grapheme associated with that row given the history (to a certain length) of graphemes that precede it in a sequence. Applying Shannon entropy to each row of this matrix and

averaging the result will then yield some sense of how regular a particular character's orthography is; how predictable or surprising the next character in the sequence tends to be. This resultant average entropy score is generated by a single character's model alone, thus avoiding the multiplication of error that would come from averaging a character's comparative scores. These measures can then be compared in some fashion (here, largely by eye), keeping in mind that a character's input length with still have some, if less dire, effect on their score. Understanding such a comparison requires applying a different interpretive lens than the one used to understand the KL or perplexity difference scores. Those measures are direct comparisons of the specific distributions of two character models, a measure of closeness that provides an inherent "why" statement (why do these two character score as similar? Because each of their character distributions is similar). Comparing the average orthographic entropy of two characters, however, does not offer such a foothold. Two characters with similar scores may well have similar scores for very different reasons, making understanding what average orthographic entropy might actually convey — what "regularity" means in an orthographic sense, what sort of substructural meaning it might shed light on — the first and most necessary step of this experiment.

Although this experiment was motivated by the bottom-up question of what entropy, given its status as a known measure of regularity, could tell us about this corpus, prior research, both linguistic and literary, offer some top-down guidance concerning what the measure might mean. Specifically, studies in linguistics and

cognitive science offer some guidance on how to interpret such a measure, even if their object and goals are not themselves necessarily literary. As previously mentioned, Shannon himself applied a derivation of his methodology to estimate the per-character entropy of English.[133] Not long after, in response to a study concluding that the entropy of Russian characters stood at 1.9, A.N. Kolmogorov, the progenitor of his own famous branch of information theory, posited that the constraints of particular genres may well lower that number to 1.2 or 1.3 in literary, and especially poetic, contexts.[134] Modern studies continue to use straightforward entropy measures with some regularity. For example, Suzuki, Buck and Tyack utilized entropy methods to analyze a discretized version of whale song. Their conclusion that whale song sequences have significantly lower entropy than the maximum possible within the whale song vocabulary leads them to argue that there is a "strong structural constraint," a syntax, that regulates the songs.[135]

The latter two examples in particular offer a possible general interpretation for our own entropy measure. In both cases, a lower entropy value is associated with the imposition of control and regularization on systems which, when unchecked, have the ability to produce a greater diversity of sequential variations. American English

[133] See Shannon, "Prediction and Entropy of Printed English."

[134] See Kolmogorov, "Three approaches to the quantitative definition of information."

[135] See Suzuki, Buck and Tyack, "Information Entropy of Humpback Whale Songs."

orthography undoubtedly fits this description. The set of graphemes included in its purview is diverse enough to create a large number of novel sequences, but its use as a tool of language subjects any such sequence to constraints generated by a number of linguistic, semantic and pragmatic factors.[136] Drawing upon the conclusions noted above, it follows that orthographies with lower average entropies are subject to stricter constraints than those with higher average entropies. There are many possible ways to grant this conclusion more specificity. One might say that the lower entropy set is "structured" or "predictable" or "uniform." In contrast, the higher entropy set might be termed "unstructured" or "random" or "diverse." The difference between these phrasings is more than empty semantics. Each comes with its own implications, and judging which is the best linguistic interpretation has interpretive ramifications. Discovering which might be most apt requires returning to the discussion of information entropy in general, as originally introduced in the methodological introduction. There, the concept of Shannon entropy was illustrated with three example distributions collected as **Table 2-4**.

As noted in the original example, both of the first two distributions have relatively high entropies — 2.0 and 2.72 respectively. However, when inspected in the above tabular form, one might be tempted by the conclusion that the latter of the two is much more "random" than the former. It has more overall possibilities, more

[136] To be completely clear, the meaning of constraint used here is not the technical one associated with cognitive linguistics. It simply refers to some sort of exterior principle that produces a negentropc effect on orthographic sequences.

opportunity for uncertainty. The difference in entropy scores embodies this intuition, but in a qualified sense. Viewed through the lens of information, the latter is more random, but not orders of magnitude so. A better way to understand it might be "widely distributed."

Viewed through the lens of an intuition trained by the conclusions from the entropy-based studies referenced above, one might expect the orthographies of this project's corpus to fall into roughly three groups. The first group, comprising the orthographic models with the lowest average entropy score, might be composed of the characters that have limited vocabularies but encode the words they use in an orthographically predictable fashion. Their vocabularies may be limited due to a number of factors — the character could have a small sample of dialogue, or could be repetitive in their word usage. The second group might contain characters with larger vocabularies, leading to them producing a larger amount of potential sequences. However, these sequences are still relatively predictable — there is some rule in play that causes certain graphematic pairings to occur more frequently than others. Hypothetically, this might contain the narrators and users of standard American English orthography; the bulk of the corpus. Finally the third group might contain the characters who also have large samples of dialogue and a large vocabulary but who also don't produce graphematic sequences as systematically from word to word. They may, for example, switch between the ending sequences "ing" and "in'" seemingly without reason. Alternatively, this group might include characters with small vocabularies that

are simply random — perhaps swapping between orthographic systems, or utilizing different graphemes in different higher-level contexts.

Examining the calculated average entropy scores both confirms and complicates this hypothesis. Excluding some extreme outliers, the range of average entropies stretches from 2 to 3. The characters gathered in the lower entropy ranges do tend to have smaller model lengths, indicating that part of their placement might be due to their relatively restricted vocabulary. However, this grouping also contains some lengthier outliers. Most notably, the selection of characters labelled "HuckI" appear in the lower entropy region despite their relatively large dialogue samples. This tag was used to mark the interstitial moments of dialogue Huck Finn uses in the novels where he serves as narrator. These moments are essentially non-diagetic, consisting of functional phrases like "he said." All three score as low entropy, with even the most extensive sample (Huck's interstitial moments from *Huckleberry Finn*) coming in as the 109th least entropic model out of all 2426 characters. The orthography associated with this character tag is predictable insofar as it fulfills the utilitarian duty of keeping the flow of dialogue and narration on track. This role constrains Huck's effusive speech to a small set, rendering his graphematic sequences predictable. However, the sought-after higher entropy groups do not emerge so clearly. For example, barring some outliers, the narrators begin to appear in earnest at rank 1463 (average entropy of approximately 2.6) and then continue to be fairly evenly distributed throughout this entire higher range. While this result is at least somewhat encouraging, as the narrators do have generally

higher average entropies (perhaps, as posited due to their expanded vocabularies) it is hard to draw any conclusions from this overly broad clustering that don't simply return to model length.

One conclusion to draw from this result is a fairly straightforward methodological one. Simple averages are sensitive to outlying values. Returning to the third example distribution provided in **Table 2-4** helps demonstrate this.

This distribution produces a far lower entropy value than the other two examples reproduced above. The earlier two were relatively high entropy — 2.0 and 2.72. This distribution has a much lower entropy, coming in at .47. Reconceptualizing these three distributions as members of the same stochastic matrix, each a row associated with the following probabilities of a different grapheme (say 'a', 'b', 'c') allows for the application of the average entropy measure employed above on the actual corpus of models. The process from this point is straightforward — simply add all three entropies together (5.19) and divide by the number of members (3) to arrive at 1.73 as the average entropy for this particular hypothetical stochastic matrix. This resultant average measure cannot help but seem unrepresentative — only one of the distributions is relatively non-entropic, but it has pulled the average down significantly. Furthermore, it is impossible to tell whether this final score is the result of an outlier without inspecting all of its constituent entropies. An average of 1.73 could just as well be the result of three entropies that cluster around 1.7 instead of the effect of an outlier on two other generally

higher entropy models. The individual randomness of a distribution becomes hidden information when it is conglomerated into such an average.

There are many ways to compensate for outlying terms, and smuggling some top down *a priori* knowledge about the nature of orthography into this generally bottom-up study allows the use of one of the more simple approaches available. Understanding the effect of an outlier is a straightforward task if the potentially outlying factor has already been identified. In the constructed example, manual inspection revealed that the third distribution and it's associated entropy were responsible for skewing the average. Recalculating the average with this element removed yields the much more intuitively appropriate average entropy of 2.36. Taking the difference of this new average value and the original average that includes the outlying distribution produces a measure that reflects how much the outlying distribution skews the overall average. Another interpretation of this measure is the amount of entropy the outlying distribution adds to the average. In this case, the skew value is -0.63. This value is directional in that it shows that including the outlying third distribution in the average pushes the overall average heavily towards the negentropic, it lowers the overall entropy of the distribution. The opposite also holds — a positive value would indicate that the inclusion of a particular distribution makes the overall average more entropic.

This skew measure is quite useful; it reveals how much a particular distribution impacts the average generated from a particular stochastic matrix, to what degree, and in which direction. However, this constructed example does not wholly comprehend the

reality of applying the technique to the average entropy scores generated from the actual corpus. The advantage granted by using a constructed example lies in knowing beforehand which particular distribution is likely to be the outlier. In contrast, manual inspection of all the stochastic matrices associated with the 2426 characters present in the corpus would prove a fool's errand. Here, however, is where top-down knowledge can come to bear on this bottom-up approach. In the context of orthography, this de-skewing method can be employed to detect the entropic effect of a specific grapheme distribution on a character's average entropy. Here, it serves as a measure of the relative randomness (in some sense we have yet to fully determine) of one particular grapheme in comparison to the others they employ.

For a corpus composed of entirely standard American English orthography this approach may not prove very illuminating. In such a corpus the typical alphabetic graphemes ('a','b', etc.) should vary quite slightly from character to character. While one may be an outlier when compared to the rest of the graphemes in the orthographic set associated with the character, this particular grapheme should vary in a substantially similar matter no matter which character is under scrutiny. The strictures of standardized orthography regulate the use of graphemes uniformly no matter the speaker. Putting aside differences in vocabulary, an 'e' should be consistently entropically distinct from the other graphemes any given character uses so long as that these characters all employ a similar orthographic standard.

This project's corpus proves a more appropriate fit for this approach. The character models extracted from the corpus are, as has been shown in other chapters, quite orthographically diverse. Those who employ an orthography not entirely consistent with some approximate nineteenth century American standard may well use certain graphemes in a manner that adds more or less entropy to their average than those who do utilize the standard. Even more intriguing is the prospect that they do not. Even if a character utilizes a particular grapheme in a manner distinct from all of the other characters in the corpus, the way they use it could still be just as entropic or negentropic as any of the other characters in relation to their own orthographic model. The actual manner of use — which graphemes this particular grapheme tends to predict — may well be different, but the level of consistency they achieve with this different set of rules could be very much the same. The skew approach could, for example, be applied to a particular alphabetic grapheme, say, 't'. If one character more consistently follows this grapheme with 'o' and another replaces that grapheme with 'e' (to form 'ew') the skew score for each character's orthography would reflect this fact, even if the ultimate result is that both are consistent in different ways. For the purposes of this project, the single apostrophe (""") was selected as the grapheme to examine using the skew approach. While the differing usages of alphabetic graphemes would no doubt prove illuminating and could be the study of future work, the single apostrophe's place of pride in nineteenth-century dialect literature makes it a logical target. The elision apostrophe is a potentially wild thing. Being less shackled to the linguistic communicative function of language allows it great latitude and throws the substructural meaning it might

convey into sharp relief. To this end, an average entropy eliding the apostrophe distribution was recalculated from each character's model, and the skew measure was generated for each using the original average entropy. Additionally, the entropy of the apostrophe distribution alone was calculated in order to provide finer-grained detail, a procedure already used in the KL comparative context in earlier chapters.

The addition of these two supplementary measures helps shed some light on the specific sense of randomness captured by the average entropy scores. Just inspecting the average entropy scores associated with the narrators included in the corpus was relatively un-enlightening. The narrators were distributed fairly evenly throughout the upper half of the average entropy rankings. It was theorized that this is due to vocabulary-related factors. While most of the narrators utilize a similar form of standard orthography, the wide-ranging nature of their role gives them the opportunity to be linguistically diverse. This in turn leads to the subtle accentuation of different orthographic regularities. If one narrator tends to attribute dialogue with "stated" and another with "said" this will have large scale effects on that character's 's' distribution. Sorting the data by the apostrophe skew score validates this conclusion. When organized in this fashion, the narrators cluster in the lower ranks, typically recording negative apostrophe skew scores. **Table 5-1** offers a sample of three narrators, ranked as the 127-129th lowest skewed.

The effect of the apostrophe distribution on these characters' stochastic matrices is heavily negentropic. The apostrophe distributions themselves are quite low entropy,

so much so that, in relation to the other distributions included in the stochastic matrix, they produce an overall negentropic shift. These narrators use the apostrophe as a regulatory, rather than potentially substructurally meaningful, feature. The third entry in the **Table 5-1** is the narrator of Charles Brockden Brown's 1799 novel *Edgar Huntly*. Examining the heatmap of his stochastic matrix shows that he uses the apostrophe almost exclusively as the possessive contraction (See **Figure 5-1**).

The apostrophe almost exclusively predicts the coming of a terminal "s," making its appearance highly predictable and thus not entropic. Though this usage of the apostrophe may have originated as an elision of some possessive final particle, in its modern (and eighteenth century form) is serves as the messenger of some higher order semantic/pragmatic restraint, in this case an indication of ownership. The substructural meaning of this grapheme has been captured by an outside system of meaning — it serves an external role so dominant that in standard forms of the era's orthography, it becomes wholly predictable. Only a few narrators gain entropy or lose an insignficant amount with the addition of the apostrophe distribution. One such example is the narrator of Sarah Orne Jewett's novel *Deephaven*, whose model gains entropy when the apostrophe is reintroduced. Examining the text reveals a straightforward reason for this — the automatic attribution process misattributed a number of sections of dialogue featuring interpolated quotations ("'an interpolation', he said"). This was enough to skew her overall distribution towards high entropy. In general then, characters with tightly controlled top-down usage of the apostrophe should lose entropy when the distribution is included and gain entropy when it remains a part of the average. Given

nineteenth-century dialect literature's fondness for the apostrophe, this means that these characters likely use some form of standard orthography, where one or two usages of the grapheme dominate all others. Again, this does not mean that the grapheme dominates in the exact same fashion from text to text — it is wholly relative to the character's usage. The heatmap of the letter writer/narrator Jack from Ring Lardner's *Busher's Letters* series demonstrates this (see **Figure 5-2**).

Jack is ranked similarly to the narrator of *Edgar Huntly* but for an entirely different reason. Here his main usage of the apostrophe is the "'t" contraction rather than the "'s" possessive. Yet since it is just as dominant as *Huntly*'s "'s" possessive it too pushes his overall average towards negentropy. In both cases, the shift is associated with a usage of the apostrophe that may have once been "just" an elision, but has become a specifically controlled linguistic unit in its own right, a standardized and consistently implemented tool.

In theory, there could be a model that loses entropy with the inclusion of the apostrophe distribution that does not fit into this case. Such a character would use the apostrophe for one consistent purpose, but not one associated with these forms of contraction. Instead, they may well choose to elide, say, the "h" in "'uman" and nothing else, leaving that usage as dominant as the "'t" and "'s" contractions above. Rather than manually inspect every stochastic matrix for this possibility, we can add a new tool to our repertoire — an automatic calculation of the two apostrophe-following graphemes with the highest probability. Among those character who lose entropy with the addition of the apostrophe distribution, almost all of them fit into the Jack/Huntly mold, with

"'s" or "'t" dominating all other possibilities.

These usages account for one extreme — those characters who use the apostrophe in a manner more regulated than their general speech. At the far end of the other extreme, those who use the apostrophe much more randomly than their general speech, we find that the skew measure predicts a different form of use. First and foremost, the most extreme outliers are those characters who do not use the apostrophe at all. Notably, this group contains many characters associated with a particular historical era — Inez Middleton from Cooper's *The Prairie* and Morgan LeFay from Twain's *A Connecticut Yankee in King Arthur's Court*, to name two. It also contains some Native American characters in the form of Tamenund from Cooper's *The Last of the Mohicans* and Occonestoga from William Gilmore Simms' *The Yemassee*. In these authors' hands the hyper-pedantry associated with using no contraction or elision whatsoever equates itself with antimodernity in addition to, perhaps, a touch of regality. Much of the group just preceding this extrema includes characters who use the apostrophe to interpellate dialogue. When compared to their general speech this usage, logically, registers as random. Since such interpellations require a matching pair of apostrophes, their successor character (especially in the case of the first apostrophe in the set) might be any character whatsoever. Polly Ochiltree from Chesnutt's *Marrow of Tradition* is the exemplary figure in this set. Her orthography is largely standard — so much so that she tends to even avoid contractions — leaving interpellation to be her chief use of the apostrophe. This set mirrors the group of characters that lose entropy when the apostrophe distribution is included. That set of characters used the apostrophe

in the manner of a linguistically controlled grapheme, but under the higher level restriction of specific semantic meaning, e.g. ownership. Because of this, their apostrophe usage did not resemble the usage of their other graphemes whatsoever. This new group uses the apostrophe under a different higher order sign. In the hands of Mrs. Ochiltree the apostrophe is largely non-linguistic, providing a demarcation function common to literary convention more than anything else.

Though they have interest in their own right, these examples primarily serve to demonstrate the functioning of the skew measure. Even more vitally, they have demonstrated the need to re-evaluate our initial concept of an extreme. The extremes of each individual measure — average entropy, entropy on the apostrophe and skew — have generated some insight on well-known high level structures of meaning that can impact orthography, but have not allowed us to access substructural meaning at large. Instead, it is the difference between the measures that proves most fruitful. This is a different notion of extreme, and a far more revealing one. **Table 5-2** contains these various measures generated on the stochastic matrix associated with Peter Finley Dunne's character Mr. Dooley.

Dooley is one of the most absolutely entropic characters in the corpus, with a rank of 2440. Despite this, the entropy of his apostrophe distribution ranks as quite low at 1267th overall. This, logically, leads to him gaining a powerful negentropic skew from his apostrophe distribution, a measure in which he ranks 828th. His usage of alphabetic characters is quite unpredictible, so much so that the addition of the apostrophe grants him a relative amount of order. In point of fact, Dooley, like Mrs. Ochiltree, is a

storyteller. His sections of dialogue are peppered with apostrophe-demarcated interpellations of previous conversations told as tales within the tale. Despite this similarity, Mrs. Ochiltree is ranked 996th overall by average entropy, her general speech is controlled, leaving her interpellation apostrophes to add, rather than subtract, entropy in relation. A factor that "seems" random — literary demarcation — in the context of Ochiltree is actually an organizing force for Dooley, one that, relative to his overall orthographic usages, evinces order and predictability. As a consequence, we could say that Dooley's range in his use of alphabetic graphemes is perhaps the broadest in the corpus. Each individual grapheme holds relatively little predictive power, and reveals relatively little about which grapheme might follow.

As **Table 5-3** reveals, The Deerslayer/Natty Bumppo from James Fenimore Cooper's *The Deerslayer: Or the First Warpath* operates in almost the opposite fashion: [137]

Natty's average entropy is relatively controlled, coming in at rank 1673. However, both the skew and entropy of his apostrophe distribution rank quite high, at 2370 and 2371 respectively. The inclusion of the apostrophe distribution adds a relative amount of entropy in this case, pushing the overall average provided by the distribution of his alphabetic graphemes towards disarray. Another way to express that is, relative to the other distributions that compose his stochastic matrix, the apostrophe has a broader

[137] It should be noted that this text was not hand attributed or corrected. However, it has been inspected by eye.

distribution. It enjoys the privilege of a different functionality than the other graphemes.

Inspecting an automatically generated list of Natty's actual vocabulary, the actual sequences of graphemes he employs, explains this phenomenon to some degree. As an initial grapheme, the apostrophe replaces the 'e' in 'ea' and 'i' in 'it' ("'arliest,'tis") but also serves to replace all of the initial graphemes of "raccoon" ("'coon"). In the medial position it seems able to stand in for almost any vowel at all, but does so inconsistently ("arr'nd" and " ar'n'd" for "errand"). It also replaces some terminal characters, most especially 'g', but also sometimes does not ("becoming", "becomin'").[138] Compared to the rest of his orthography, the apostrophe is guided by some different purpose.

The titular figure of Harriet Beecher Stowe's *Uncle Tom's Cabin* fits a similar profile (see **Table 5-4**). Uncle Tom ranks 1421st in average entropy, 2215th in skew and 2240th in apostrophe distribution entropy. Like Natty, the inclusion of the apostrophe distribution in his stochastic matrix serves to make his grapheme sequences less predictable overall. Again, this can be glimpsed more concretely in his vocabulary — "chil'en" and "children", "takin'" and "taking." Despite their relatively similar metric-level position, with both gaining a decent amount of entropy from their apostrophe distribution, they do so in somewhat distinct ways. As an initial grapheme, for example, Tom tends to use the apostrophe to replace 'a' and 're.' Even more

[138] These conclusions were checked against print versions of the novel to minimize transcriber error/textual variations. Some amount of variety does seem to exist which is perhaps an implicit commentary on the apostrophe's general use in the work.

prominent is its use in replacing the medial "te" ("mas'r"). Despite both characters using the apostrophe in a distinct fashion, their relatively similar average entropies and skew values reveal that their attitude towards the use of the apostrophe is still similar. Both have a few general use-cases that repeat throughout the their respective portions of dialogues, but they also both deviate at times, distributing the apostrophe more freely than the other graphemes they employ. Casual inspection of each character's vocabulary offers an even more compelling possibility. Natty's approach to the apostrophe seems both more extensive and broadly distributed than Tom's, an observation supported by his higher apostrophe distribution entropy. This is evidence that the skew value might serve as some sort of metric rather than just a point of comparison among characters. Both high and low values of the term would indicate that the apostrophe is employed differently than the other graphemes in play, allowing the characterization of what "sort" of apostrophe user any given stochastic matrix is likely do be without recourse to messy comparison.

The final unexpected extreme that proves salient is the extremely moderate. Take the scores from the two models collected in **Table 5-5**. The first row is generated from the stochastic matrix of Polly Sellers from *The American Claimant* and the second Jim from *Tom Sawyer, Abroad*. These two characters register remarkably similar scores across the board, have extremely similar model lengths, and were even crafted by the same author (Twain). They are only a handful of ranks distant when ordered by skew score and apostrophe distribution entropy, and perhaps two hundred when sorted by average entropy with Jim tending towards the more entropic side. Relative to the usage

of their other graphemes, the apostrophe has a very mildly negentropic effect on the model averages of both. To those familiar with these characters, this is a striking result. Here is a sample of text drawn from Polly's utterances:

"Laws! The idea. They would if they could, poor old things, and perhaps they think they do do some of it. But it's a superstition. Dan'l waits on the front door, and sometimes goes on an errand..."

And one from Jim:

"I knows it perfectly well, Mars Tom—'deed I knows it perfectly well. But ef we takes a' axe or two, jist you en me en Huck, en slips acrost de river to-night arter de moon's gone down..."

Numerous other means that demonstrate their actual orthographic sequence level differences might follow. Their heatmaps appear generally dissimilar(see **Figures 4-3** and **4-4**). Some general notes from the listings of their vocabulary help reveal this disjunction as well. Polly uses largely standard spellings with the addition of a few particular interstitial elisions (Dan'l) and an expanded amount of contractions as compared to other characters — she uses all the borderline standard "'ll", "'re", "'ve" and

"'d" contractions she can get her hands on. Jim most frequently uses the apostrophe to elide initial and final graphemes or grapheme sequences, and, as can be glimpsed from his sample, utilizes an overall different system of spelling.

One more comparison is required before unpacking the consequences of this similarity. Jim's overall average entropy is almost identical to Deerslayer's, varying only by a few thousandths. When sorted by this particular metric, they are roughly a dozen ranks apart. Deerslayer's apostrophe skew, however, is quite entropic, while Jim's and Polly's are very slightly negentropic. He does not group with them when sorted by this metric. From this set of comparisons, as well as the general characterization of the skew metric performed above, a few conclusions emerge. What unifies Jim and Polly is their commitment to using the apostrophe in a manner consistent with the other graphemes they employ. This does not speak to the consistency of their orthography in general, how regularly any particular grapheme will predict the next. This factor is to some degree captured more accurately by the overall average entropy. Jim is more entropic in the aggregate than Polly, effectively equal to Deerslayer. However, Deerslayer's much more entropic apostrophe skew score shows that unlike Jim he gains this additional entropy in the particular way he uses the apostrophe. Jim and Polly use this grapheme more or less like any other in their set, while Deerslayer employs it distinctively.

This means that as the apostrophe skew score approaches either extreme — regardless of it being an extreme point of high entropy or low entropy — the usage of the apostrophe becomes less and less guided by the principles underlying the associated

character's general use of orthography and becomes more and more guided by some other outlying structure of meaning. The extremity of the divergence, rather than the direction (entropic or negentropic) is the important factor in this determination. Both Ochiltree and Dooley veer away from a skew value of 0 for the same literary reason, the use of interpollated dialogue, but do so in different directions as their overall entropies are massively different. They, as well as the narrators and very-standard orthography users like Jack that reside in the lower skew regions, apply an organizing principle to their apostrophe usage that they do not apply to their other graphemes in general.

This realization implies one final useful measure. Taking the absolute value of the skew score for each character generates an ordering based on how distinctively a character uses the apostrophe compared to their overall grapheme use. The further away from zero a character ranges, the more distinct their usage. The characters lowest on this scale guide the usage of an apostrophe by a set of rules akin to those that structure their general orthography. The characters in the highest ranges use the apostrophe quite differently. For some, like narrators who only use the possessive, or characters who interpollate dialogue, this is due to the extremity of their contributing exterior structure. For others it could be a sign that, in comparison to their general orthographic principles, their use of the apostrophe is seemingly random.

As a potential "in" to the meaning of a particular form of orthographic substructural style based around the apostrophe, the most intriguing possibilities are found in the middle areas of this ranking. These characters employ the apostrophe just distinctively enough to make its separation from the character's general speech

palpable, yet does not wind up so predictable (or unpredictable) to the point where it can carry no meaning of its own. In the midpoint between total in-distinction and total predictability or unpredictability, meaning can flourish.

Possibly, one would be tempted to condemn the dwellers of this middle region as users of "eye dialect." Barring, as we previously have, definitions of this term that cling to phonology in even some small degree, such a conclusion does not quite follow.[139] For one, this conclusion bears only on the apostrophe. A comprehensive analysis of the consistency of all a character's graphemes would be an entirely different and much larger task. More importantly though, the randomness evinced in these regions is not cacography, it is self-exception. It is the utilization of one of the main capacities of orthography as a tool — its flexibility — to both use it as a tool to transmit semantic meaning and, whether unconsciously or consciously, point towards a substructural meaning potentially distinct in its message. An absolute difference in entropy, in this case, does not necessarily expose randomness in the manner it is commonly conceived, but instead the potential that a different form of order might condition the phenomenon under question, even if this different form of order is not as apparent as, say, the possessive. Randomness is relative to the context in which it is examined.

Returning to *The Deerslayer* illustrates the potential inherent in this re-interpretation. So far as Cooper's use of nonstandard orthography goes, Twain's famous 1895 essay on Cooper's "Literary Offences" seems to still hold on to the last

[139] Contra, for example, Minnick, *Dialect and Dichotomy*.

word.[140] Twain argues that the commandments of literature "require that when a personage talks like an illustrated, gilt-edged, tree-calf, hand-tooled, seven- dollar Friendship's Offering in the beginning of a paragraph" they should not pivot suddenly to "talk like a negro minstrel in the end of it."[141] Putting aside the accuracy of his specific comparison ("negro minstrel") Twain's statement simply demands the question "why not?" Self-consistency may be important if one wishes to use orthography to point to one stable foothold of meaning, say a regional racial or class identity, but it need not do so. An inconsistency from this perspective may well be a consistency, or simply an isolated moment of meaning, from another.

Late in the novel, Natty receives two proposals of marriage. As is his wont, he declines both, but does so in somewhat different terms. His issues his first denial to Judith Hutter:

"A woman like you that is handsome enough to be a captains lady, and fine enough, and, so far as I know, edication enough, would be little apt to think of *becoming* my wife."[142]

---

[140] Future MLA President Louise Pound did take up the question some 30 years after Twain, producing a linguistically-informed study that deems "his departure from standard speech" to consist mostly of "archaisms" and thus relatively "dignified" compared to the modern use of slang. See Pound, "The Dialect of Cooper's Leather-Stocking."

[141] Twain, "Fenimore Cooper's Literary Offences," 60.

[142] Cooper, "The Deerslayer," 420.

And the second to a Native American woman:

"The tarms are onadmissable, woman; and, though I feel for your losses, which must be hard to bear, the tarms cannot be accepted. As to givin' you ven'son, in case we lived near enough together, that would be no great expl'ite, but as for *becomin'* your husband, and the father of your children, to be honest with you, I feel no callin' that-a-way."[143]

Perhaps a distance of 60 pages is enough to avoid the linguistic whiplash Twain takes general exception to, but the self-relative orthographic inconsistency evident in these two passages offers a different conclusion. Most tempting is the notion that Deerslayer is "code switching," changing his spoken dialect to meet the appropriateness of the situation. With an educated white woman he makes sure to include his final "g," while in the relatively "informal" backwoods conversation with a native woman he feels no such pressure. Such a conclusion affords Cooper much agency. In this account of the difference, Cooper is in tune with the orthographic meaning the inclusion or elision of the "g" provides. He is a producer of linguistic knowledge; tuning Deerslayer's tones to fit the formality of the occasion. While such an argument could be made, it places a great burden on the ability to psychologize Cooper himself. It would require making an argument about the specific import of this particular inconsistency (among others) to

[143] Cooper, 494.

Cooper as a conscious writing being, that it was planned specifically for some specific effect.

This project's commitment to the independence of a form of orthographic meaning that occurs somewhere between the automatic processing of decoding and the higher-level interventions that impinge on (especially) encoding makes adopting this potentially tendentious conclusion unnecessary, even if one wishes to avoid returning to the notion that Cooper's usage of orthography is somehow random. Cooper's relatively entropic orthography could as much be the sign of the intervention of multiple higher order restrictions at once as it could the sign of forgetfulness or carelessness. In this case, genre serves as a likely source of these constraints. In his response to Judith, Deerslayer inhabits the role of a male lead in a novel of manners.[144] The standard of matrimonial appropriateness he sets himself against ("captain") is one appropriate to this genre and would not be out of place in a historical romance as well. Deerslayer's profession of his own lack of social status would fit right in to a Jane Austen or Walter Scott b-plot, and imagining a line where he eventually becomes the chosen betrothed of Hutter regardless is not hard to do. In contrast, the second refusal hints at the gothic horror embedded in the euro-America anxiety towards racial mixing.[145] In this moment Deerslayer professes fear rather than some sort of insufficiency. When confronted by

[144] And indeed Cooper did launch his literary career with a novel of manners entitled *Precaution*.

[145] A common trope of early American literature well explored by critics as far back as Fiedler's 1960 *Love and Death in the American Novel*.

what he sees as the simulacrum of marriage, an unhealthy intimacy with the trappings but not, in Deerslayer's view, the soul of this union, he recoils. Cooper, like other early American writers, blend these genres — the novel of manners, the historical romance, gothic horror — to develop an early take on uniquely North American literature.[146] In the example of these two instances Cooper blends through mixing rather than incorporating. Both moments are an instance where intimacy is proposed, but each is encoded in the trappings of their own particular genre. Cooper's orthographic deviation is, in turn, a symptom of the differences between these higher order structures. The elided "-ing" in the second instance must be seen as the hallmark of Cooper's employment of one particular set of generic structures (gothic adventure) rather than another (novel of manners) to encode this particular situation. The elision is not random, but not intentional. Instead it itself *means* this genre, at least when catalyzed through Cooper's own semi-conscious writerly production. Just as dedication to genre norms might order literary choices on a structural or semantic level, this too has an impact on the orthographic plane of meaning, even if such semi-consciously produced details only emerge when viewed from the third-person perspective. The bottom-up computational corpus approach renders moments such as these visible, providing

[146] Charles Brockden Brown's 1799 *Ormond* serves as a stellar example of the commingling of (especially) elements from gothic fiction with those from the novel of manners, while authors like Lydia Maria Child in her 1824 novel *Hobomok* draw more on the historical romance-gothic commingling (even if her ends are certain distinct from Cooper's own).

accounts of their significance that allow them to fit into more standard literary critical narratives.

Figures

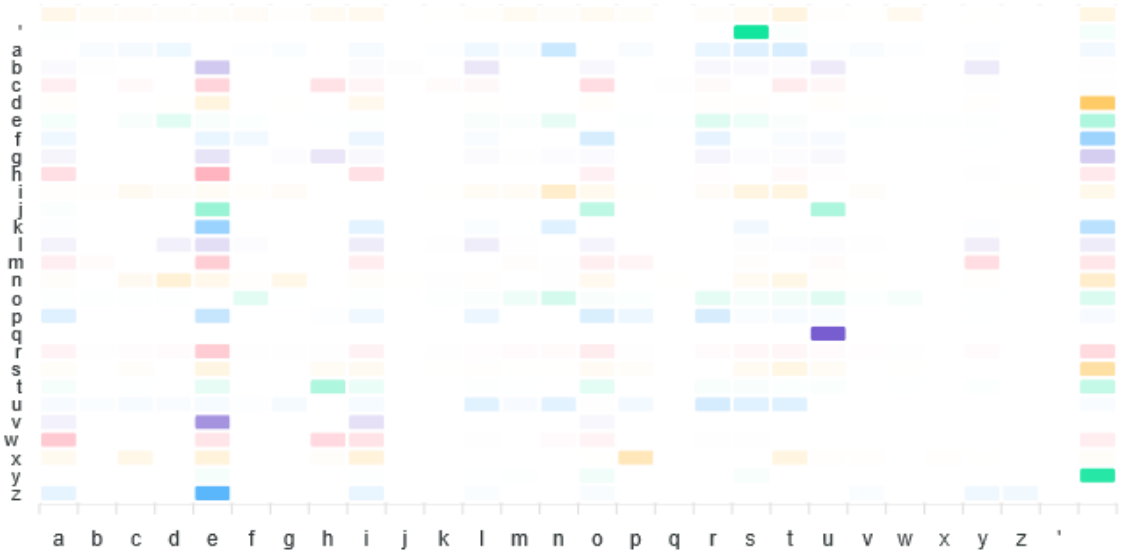

Figure 5-1. A heatmap visualizing the stochastic matrix generated from the narrator of Charles Brockden Brown's 1799 novel *Edgar Huntly*. The darkly-shaded green segment at the intersection of the apostrophe row and "s" column indicates that this narrator primarily uses the apostrophe to indicate the possessive.

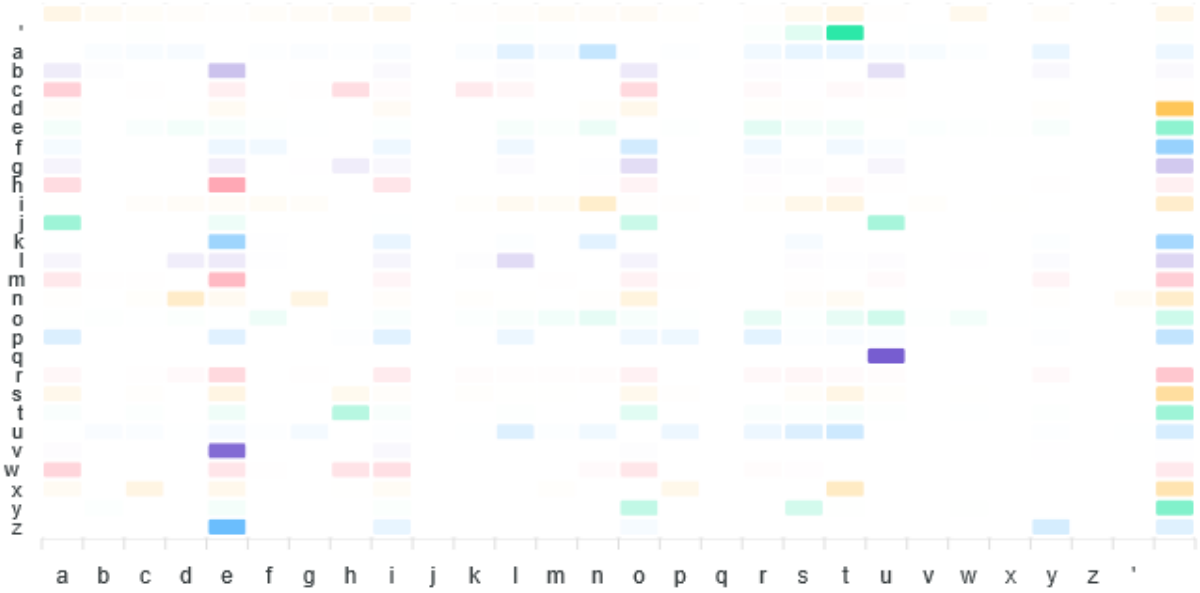

Figure 5-2. A heatmap visualizing the stochastic matrix generated from Jack, the letter writer/narrator of Ring Lardner's *Busher's Letters*. The darkly-shaded green segment at the intersection of the apostrophe row and "t" column indicates that Jack uses the apostrophe primarily for the "'t" contraction found in wordforms like "ain't" or "can't."

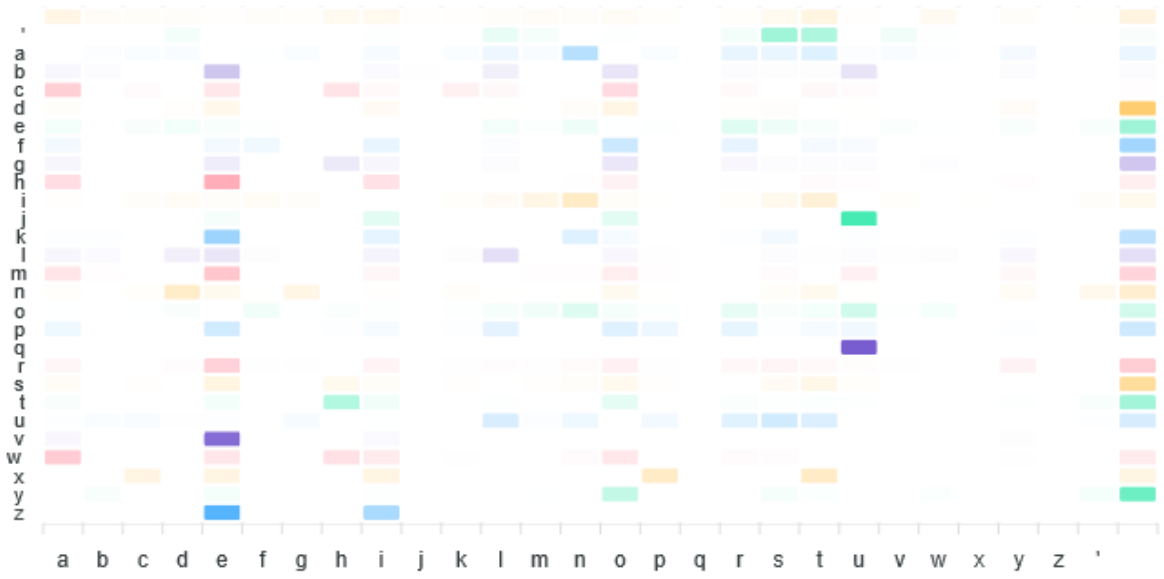

Figure 5-3. A heatmap generated from the stochastic matrix of Polly from Mark Twain's *American Claimant*.

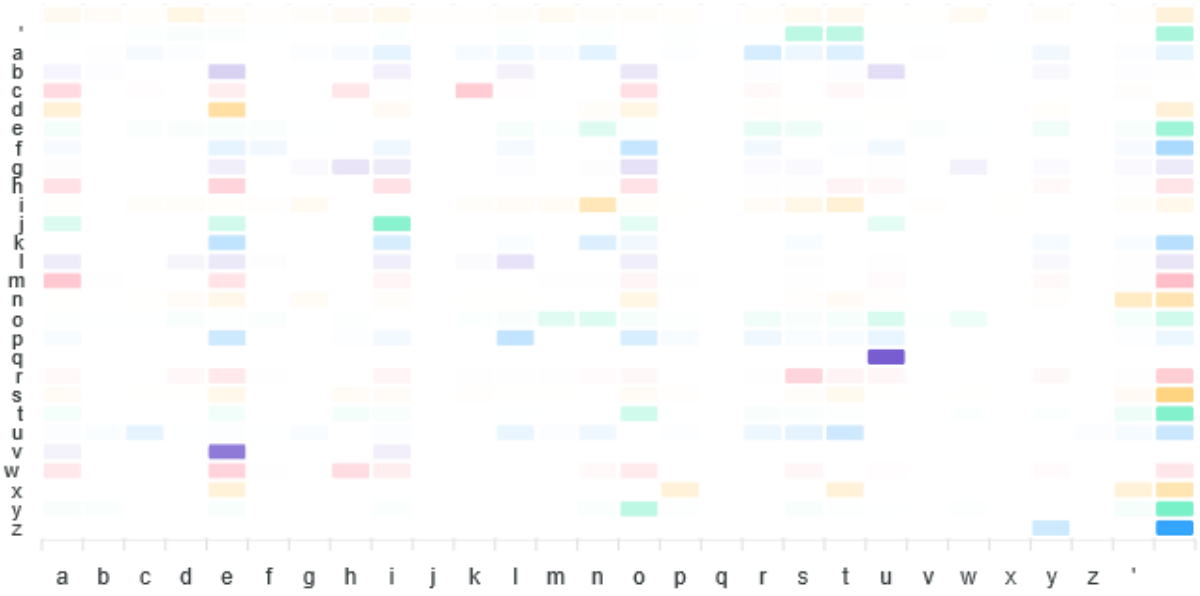

Figure 5-4. A heatmap generated from the stochastic matrix of Jim from Mark Twain's *Tom Sawyer Abroad*.

Tables

Character	Average Entropy	Apostrophe Entropy	Average Without Apostrophe	Skew
Narrator	2.65	0.4	2.74	-0.083
Narrator	2.68	0.43	2.76	-0.083
Narrator	2.68	0.45	2.77	Narrator

Table 5-1. This table offers a sample of three narrators, ranked as the 127-129th lowest by skew scores. The first column is the average entropy of the character's stochastic matrix model and the second the entropy of the apostrophe row of that matrix. The third is the model's average entropy with the apostrophe distribution removed and the fourth is the skew score.

Character	Average Entropy	Apostrophe Entropy	Average Without Apostrophe	Skew
Dooley	2.948	2.34	2.97	-0.023

Table 5-2. A table collecting the average measures generated from the stochastic matrix model of Peter Finley Dunne's Mr. Dooley character.

Character	Average Entropy	Apostrophe Entropy	Average Without Apostrophe	Skew
Deerslayer	2.67	3.39	2.64	0.026

Table 5-3. A table collecting the average measures generated from the stochastic matrix model of Cooper's Deerslayer version of Natty Bumppo found in *The Deerslayer:Or the First Warpath*.

Character	Average Entropy	Apostrophe Entropy	Average Without Apostrophe	Skew
Tom	2.63	3.02	2.62	0.014

Table 5-4. A table collecting the average measures generated from the stochastic matrix model of the titular character of Harriet Beecher Stowe's *Uncle Tom's Cabin*.

Character	Average Entropy	Apostrophe Entropy	Average Without Apostrophe	Skew
Polly Sellers	2.63	2.44	2.64	-.007
Jim	2.67	2.5	2.68	-.006

Table 5-5. A table of average entropy scores generated from the stochastic matrices of Polly Sellers from *The American Claimant* and Jim from *Tom Sawyer, Abroad*.

Coda

The previous chapters of this dissertation have varied in their particular focus. While all are united by a commitment to the particular approach to orthographic technology developed throughout and the shared utilization of a specifically collected corpus of works, their individual topics have ranged from the specific examination of a small set of texts to broad methodological examinations that touch on many. As a result, this project has offered equally numerous moments of conclusion. This final section will not attempt to offer their total sum. In part, this is because this dissertation is a step towards further work rather than a full end in itself. The density of texts collected in the corpus and additional precision gained by honing the methodological approach used to analyze orthographic sequences would both benefit from improvements further work could provide. At its core, this project is inherently a living one. New additions and revisions to the corpus have the potential to radically change any historical or literary conclusions derived from the analyzed data, and can point to novel inroads that were previously unseen.

The living nature of this project aside, the conclusions that have been drawn in the above chapters are in point of fact quite diverse, and might be separated into three categories (with the normal caveat that there will be some inevitable overlap): methodological conclusions, literary conclusions and theoretical conclusions. Each will be explored below in its own section.

Methodology

Computational methodology is not new to literary criticism. Beyond current approaches made popular by the digital humanities, historical examples of employing computers for literary processing abound. As mentioned in a previous chapter, the earliest days of mainframe computing saw Fr. Roberto Busa utilizing an automated approach to concordance-making as a labor-saving measure. However, even in the late 1970s, the dawning era of the personal computer, critics like Hugh Kenner were expanding the literary uses of computation, working to employ them in a creative and analytic capacity rather than as time-saving drudges.[147] As surprising as it is to say given the popularity of the topic in modern discourse, many of the avenues by which computation can contribute to literary study have already been traversed. New digital approaches might extend the power of one of the these avenues — labor saving, information analysis, creativity — but it is difficult to argue that they have moved beyond the tasks envisioned for them by early pioneers like Busa and Kenner. To that end, this dissertation did not seek to introduce a fully novel set of computational techniques. Instead, it sought to import a set of methodologies from corpus linguistics that have only rarely been employed towards specifically literary critical ends.[148] The

[147] For example, his work on automatic poetry generation in collaboration with Charles O. Hartman. See Hartman and Kenner, *Sentences*

[148] Not to say this move is without precedent, see for example the aforementioned work of David Hoover.

corpus and information-based techniques drawn from this scholarly discipline have great power to render features of texts more available to more standard literary critical approaches, but have, in modern times, been underutilized in comparison to machine learning and statistical techniques drawn from computer science. In part, this seems to stem from the theoretical attachments corpus techniques carry with them. Linguistics disciplines are seen as mind sciences, and the conclusions they draw from corpus means are appropriate to this particular empirical end. Studies from linguistics sub-fields often hope to discover more about the invariant aspect of linguistic phenomena, their invariant brain-bound underpinnings, and less about the sort of exceptions and oddities that tend to excite literary critics.[149]

This dissertation hoped to demonstrate that literary criticism can employ these techniques without adopting their ends. Orthography offers a stellar use-case for such an approach. Being at its core extra-linguistic, a tool bent towards language rather than one of language's instinctual apparatuses, corpus-level analysis of non-orthography is

[149] For example, the corpus approach to "construction grammar," possibly best represented by the work of Stefan Th. Gries, uses corpus-based research to draw conclusions about, especially, the nature of syntactic processing and cognitive semantics. Though this dissertation does not take up such an end, it should be noted that scholars employing construction grammar approaches are the natural linguistic co-travelers of this sort of literary project. Chomskyean linguistics insists on "deep structure," rendering the analysis of surface emanations without reference to what might produce them from underneath fruitless. Construction grammar, on the other hand, argues that evidence found in the surface expression of language is very much meaningful, a point this project has tried to stress.

less about accumulating samples of invariant features in order to build an evidential case for some particular lower-level functioning and more about inspecting the near infinite meaning-holding variations different forms of orthographic expression contain. The linguistics approach towards orthography does not exhaust its possibilities. More than other elements of written language, new additions to a corpus have the potential to bring whole new phenomena with them — new exceptions to a general set of rules, new semi-systemic nonstandard orthographic forms. Despite its innate nature, orthography is still subject to variation introduced by top-down principles, "intrusions" from phenomenal levels of cognition. In such an environment, the search for invariant principles can only go so far.

This is where literary critical approaches can step in. Close reading and historical analysis flourish in domains where exceptions and variety multiply, and importing corpus approaches renders the orthographic versions of these moments visible to such traditional literary critical techniques. On a methodological level, this project aimed to demonstrate how this transfer can proceed while retaining analytical goals appropriate to literary criticism.

Literary

This dissertation has offered a handful of specific conclusions related to individual texts or groups of texts collected in the corpus. These individual moments have all relied on the major literary critical tenet this project has embraced — the

contention that orthography provides its own set of literary meanings distinct from those associated with the words or syntax they encode as well as the encoding process itself. This is, to some degree, an importation from sociolinguists like Sebba and Harris. It also draws upon previous literary studies that note, even if implicitly, this phenomenon but do not fully flesh out its implications.[150] By applying this general insight to the literature of the nineteenth and early twentieth century United States this project aimed to shed new light on the texts of this era while at the same time using the widespread orthographic creativity found in the texts of this period to further the overall point about orthographic meaning's independence. This era seems an especially apt one to investigate through the orthographic substructural lens. The uncodified process that continually defines a sense of what counts as standard American English orthography accelerated into and through this era, producing new orthographic distinctions and similarities that, in turn, produce novel orthographic meanings. By unshackling orthography from syntax and diction, a general understanding of these meanings that does not simply recapitulate dialect-centric connections to racial, class or regional archetypes begins to emerge. Orthographies have their own meaningful literary histories, and this dissertation aimed to serve as a prolegomena to a larger project that could detail the life of many such systems evident in nineteenth-century America with greater scope and precision.

[150] For example, Walpole, "Eye Dialect in Fictional Dialogue."

Theoretical

Finally, on a theoretical level, this dissertation has argued that phenomena only fully available to third-person means of investigation are a core part of the literary experience. While literary criticism has always had an interest in depths, the symptomatic approach that proceeds from phenomenal experience "downward" simply may not capture large swaths of meaningful experience. As the tool that lives inside us, orthography is the prime example of such a phenomenon. It is the ultimate "at-hand" tool, the prime "supplement" or "prosthetic," the technology large swathes of childhood-literate humanity literally cannot imagine living without.[151] Despite its seemingly innate nature, writing and orthography are historically specific means to an end, subject to variation, change, and intervention from other cognitive apparatuses. The historical nature of orthography in general is hard to grasp from the first person perspective, but when viewed from askance its potential contributions to literary meanings become more clear. This project has used orthographic substructural style as an example of just one of these phenomena, meaningful elements of the reading process that are recalcitrant when faced only with the traditional literary critical toolkit but that

[151] See Heidegger, *Being and Time* and Simondon, *On the Mode of Existence of Technical Objects*. Heidegger's work on both dasein and tool being have earned some popularity in the cognitive sciences, often times catalyzed through philosophers of mind like Herbert Dreyfus. See Kiverstein and Wheeler, *Heidegger and Cognitive Science*.

become available to such means when other methods are first employed. Many more such phenomena likely exist. Rather than cede these elements wholly to the mind sciences, this project has argued that such phenomena are inherently literary — inherently historical, inherently contextually sensitive — and that their study and elucidation belongs in departments of literature as much as it does in departments of linguistics, cognitive science, psychology or neuroscience.

Appendix One: Corpus Listing

Author	Text	Year	Full Text (Typed or corrected OCR)	Downloaded	Hand?
Harris, George Washington	Sut Lovingood. Yarns spun by a natral born durn'd fool.	1867	Documenting the South	1	
Longstreet, Augustus Baldwin	Georgia Scenes	1835	Documenting the South	1	
Holley, Marietta	Samantha Among the Brethren	1890	Gutenberg	1	
Holley, Marietta	Sweet Cicely – Or Josiah Allen as Politician	1885	Gutenberg	1	
Holley, Marietta	Samantha at Saratoga	1887	Gutenberg	1	
Hooper, Johnson Jones	Some Adventures of Captain Simon Suggs	1845	Documenting the American South	1	
Caruthers, William Alexander	Cavaliers of Virginia	1835	Vol 1 - Gutenberg Vol 2. Gutenberg	1	

Caruthers, William Alexander	The Knights of the Horse-Shoe	1835	Documenting the South	1	
Kennedy, John Pendleton	Horse-Shoe Robinson: A Tale of the Tory Ascendancy	1835	Gutenberg	1	
Kennedy, John Pendleton	Quodlibet	1840	Gutenberg	1	
Kennedy, John Pendleton	Rob of the Bowl	1838	Documenting the South	1	
Phelps, Elizabeth Stuart	Gypsy Breynton	1866	Gutenberg	1	
Matthews, Brander	Vignettes of Manhatten/Studies in Local Color	1894	Gutenberg	1	
Simms, William Gilmore	The Yemassee	1835	Documenting the American South	1	
Simms, William Gilmore	Guy Rivers: A Tale of Georgia	1834	Gutenberg	1	
Simms, William Gilmore	The Sword and the Distaff	1852	Documenting the American South	1	
Cahan, Abraham	Yekl: A Tale of the New York Ghetto	1896	Gutenberg	1	1

Holmes, Mary Jane	Tempest and Sunshine	1852	Gutenberg	1	
Holmes, Mary Jane	Lena Rivers	1856	Gutenberg	1	
Holmes, Mary Jane	The English Orphans; or A Home in the New World	1855	Gutenberg	1	
Holmes, Mary Jane	Maggie Miller: The Story of Old Hagar's Secret	1860	Gutenberg	1	
Thanet, Octave (Alice French)	Stories of a Western Town	1892	Gutenberg	1	
Cummins, Maria Susanna	The Lamplighter	1854	Gutenberg	1	
Garland, Hamlin	Jason Edwards: An Average Man	1892	No	1	
Garland, Hamlin	A Little Norsk; or Ol' Pap's Flaxen	1892	Gutenberg	1	
Garland, Hamlin	Main-Travelled Roads	1891	Gutenberg	1	
Garland, Hamlin	Prairie Folks	1893	Gutenberg	1	
Stowe, Harriet Beecher	Uncle Tom's Cabin	1852	Gutenberg	1	1
Stowe, Harriet Beecher	Palmetto-Leaves	1873	Gutenberg	1	

Webb, Frank J.	The Garies and and their Friends	1857	Gutenberg	1	
Crane, Stephen	The Red Badge of Courage	1895	Gutenberg	1	
Crane, Stephen	Maggie: A Girl of the Streets	1893	Gutenberg	1	
Hay, John	The Bread-Winners	1883	Gutenberg	1	
Stratton-Porter, Gene	Freckles	1904	Gutenberg	1	
Stratton-Porter, Gene	Song of the Cardinal	1903	Gutenberg	1	
Stratton-Porter, Gene	A Girl of the Limberlost	1909	Gutenberg	1	
Woolson, Constance Fenimore	Anne	1880	Gutenberg	1	
Woolson, Constance Fenimore	Jupiter Lights	1889	Gutenberg	1	
Frederic, Harold	The Damnation of Thereon Ware	1896	Gutenberg	1	
Frederic, Harold	The Market-Place	1899	Gutenberg	1	
Brown, Alice	Meadow-Grass: Tales of New England Life	1896	Gutenberg	1	

Brown, Alice	Tiverton Tales	1899	Gutenberg	1	
Slosson, Annie Trumball	Fishin' Jimmy	1889	Gutenberg	1	
Lincoln, Joseph C.	Cap'n Eri: A Story of the Coast	1904	Gutenberg	1	
Greene, Sarah Pratt McLean	Cape Cod Folks	1881	Gutenberg	1	
Greene, Sarah Pratt McLean	Vesty of the Basins	1892	Gutenberg	1	
Wharton, Edith	Ethan Frome	1911	Gutenberg	1	
Jewett, Sarah Orne	Deephaven	1877	Gutenberg	1	
Jewett, Sarah Orne	Strangers and Wayfarers	1890	Gutenberg	1	
Jewett, Sarah Orne	Old Friends and New	1879	Gutenberg	1	
Jewett, Sarah Orne	The Tory Lover	1901	Gutenberg	1	
Jewett, Sarah Orne	Betty Leicester: A Story for Girls	1890	Gutenberg	1	
Jewett, Sarah Orne	The Country of the Pointed Firs	1896	Gutenberg	1	
Jewett, Sarah Orne	A Country Doctor	1884	Gutenberg	1	

Freeman, Mary Wilkins	Pembroke	1894	Gutenberg	1	
Freeman, Mary Wilkins	Jerome, A Poor Man	1897	Gutenberg	1	
Freeman, Mary Wilkins	Jane Field	1892	Gutenberg	1	
Freeman, Mary Wilkins	The Wind in the Rose Bush and Other Tales of the Supernatural	1903	Gutenberg	1	
Page, Thomas Nelson	Two Little Confederates	1888	Gutenberg	1	
Page, Thomas Nelson	The Burial of the Guns	1894	Gutenberg	1	
Page, Thomas Nelson	Marse Chan	1887	Documenting the American South	1	1
Allen, James Lane	The Blue-Grass Region of Kentucky	1892	Gutenberg	1	
Murfree, Mary Noailles (Charles Egbert Craddock)	The Prophet of the Great Smoky Mountain	1885	Gutenberg	1	
Murfree, Mary Noailles (Charles Egbert Craddock)	The Phantoms of the Foot-Brdge, and Other Stories	1895	Gutenberg	1	

Murfree, Mary Noailles (Charles Egbert Craddock)	Down the Ravine	1885	Gutenberg	1	
Murfree, Mary Noailles (Charles Egbert Craddock)	The Frontiersmen	1904	Gutenberg	1	
Chopin, Kate	The Awakening	1899	Gutenberg	1	
Glasgow, Ellen	The Deliverance: A Romance of the Virginia Tobacco Fields	1904	Gutenberg	1	
Glasgow, Ellen	The Battle-Ground	1902	Gutenberg	1	
Glasgow, Ellen	The Romance of A Plain Man	1909	Gutenberg	1	
King, Grace Elizabeth	Monsieur Motte	1888	Documenting the American South	1	
Dunbar-Nelson, Alice	The Goodness of St. Rocque, and Other Stories	1899	Gutenberg	1	
Austin, Mary Hunter	The Land of Little Rain	1903	Gutenberg	1	
Gale, Zona	Romance Island	1906	Gutenberg	1	
Gale, Zona	Friendship Village	1908	Gutenberg	1	

Robinson, Rowland Evans	A Hero of Ticonderoga	1898	Gutenberg	1	
Stuart, Ruth McEnery	Sonny, a Christmas Guest	1896	Gutenberg	1	
Stuart, Ruth McEnery	Moriah's Mourning and Other Half-Hour Sketches	1898	Gutenberg	1	
Bird, Robert Montgomery	Nick of the Woods	1837	Gutenberg	1	
Griggs, Sutton	Imperium in Imperio	1899	Gutenberg	1	
Lowell, James Russell	The Biglow Papers	1848	Gutenberg	1	
Chesnutt, Charles W.	The Conjure-Woman	1899	Gutenberg	1	1
Chesnutt, Charles W.	The Marrow of Tradition	1901	Gutenberg	1	1
Chesnutt, Charles W.	The House Behind the Cedars	1900	Gutenberg	1	
Chesnutt, Charles W.	The Colonel's Dream	1905	Gutenberg	1	
Chesnutt, Charles W.	The Wife of his Youth	1899	Gutenberg	1	
Cable, George Washtington	The Grandissimes: A Story of Creole Life	1880	Gutenberg	1	1

Cable, George Washtington	Madam Delphine	1881	Gutenberg	1	
Cable, George Washtington	Dr. Sevier	1882	Gutenberg	1	
Cable, George Washtington	The Cavalier	1901	Gutenberg	1	
Cable, George Washtington	Strong Hearts	1899	Gutenberg	1	
Cooper, James Fenimore	The Last of the Mohicans: A Narrative of 1757	1826	Gutenberg	1	
Cooper, James Fenimore	The Deerslayer: or The First Warpath	1841	Gutenberg	1	
Cooper, James Fenimore	The Prairie	1827	Gutenberg	1	
Cooper, James Fenimore	The Spy: A Tale of Neutral Ground	1821	Gutenberg	1	
Cooper, James Fenimore	The Pathfinder, or the Inland Sea	1840	Gutenberg	1	
Cooper, James Fenimore	The Pioneers; or the Sources of the Susquehanna	1823	Gutenberg	1	
Bangs, John Kendrick	Ghosts I Have Met and Some Others	1898	Gutenberg	1	
Bangs, John Kendrick	The Dreamers: A Club	1899	Gutenberg	1	

Harte, Bret	The Luck of Roaring Camp and Other Sketches	1870	Gutenberg	1	
Harte, Bret	Gabriel Conroy	1876	Gutenberg	1	
Tarkington, Booth	The Man from Indiana	1899	Gutenberg	1	
Tarkington, Booth	Monsieur Beaucaire	1900	Gutenberg	1	
Howells, William Dean	The Rise of Silas Lapham	1885	Gutenberg	1	1
Howells, William Dean	A Modern Instance	1882	Gutenberg	1	
Howells, William Dean	A Hazard of New Fortunes	1889	Gutenberg	1	
Howells, William Dean	The Lady of the Aroostook	1879	Gutenberg	1	
Dreiser, Theodore	Sister Carrie	1900	Gutenberg	1	
Twain, Mark (Samuel Clemens)	Roughing It	1872	Gutenberg	1	1
Twain, Mark (Samuel Clemens)	Pudd'nhead Wilson	1894	Gutenberg	1	1
Twain, Mark (Samuel Clemens)	The Gilded Age: A Tale of Today	1873	Gutenberg	1	1

Twain, Mark (Samuel Clemens)	Adventures of Huckleberry Finn	1884	Gutenberg	1	1
Twain, Mark (Samuel Clemens)	The Adventures of Tom Sawyer	1876	Gutenberg	1	1
Twain, Mark (Samuel Clemens)	Tom Sawyer, Detective	1896	Gutenberg	1	1
Southworth, E.D.E.N.	The Hidden Hand; or Capitola the Madcap	1859	Gutenberg	1	
Southworth, E.D.E.N.	Ishmael; or In the Depths	1876	Gutenberg	1	
Southworth, E.D.E.N.	The Missing Bride	1855	Gutenberg	1	
Southworth, E.D.E.N.	The Lost Lady of Lone	1890	Gutenberg	1	
Southworth, E.D.E.N.	Self Raised; or From the Depths	1878	Gutenberg	1	
Southworth, E.D.E.N.	Tried for Her Life	1871	Gutenberg	1	
Thompson, George	Venus in Boston	1849	Gutenberg	1	
Buntline, Ned (E.C.Z. Judson)	Wild Bill's Last Trail	1896	Gutenberg	1	

Eggleston, Edward	The Hoosier Schoolmaster	1871	Gutenberg	1	
Eggleston, Edward	The Mystery of Metropolisville	1873	Gutenberg	1	
Eggleston, Edward	The Graysons: A Story of Illinois	1888	Gutenberg	1	
Eggleston, Edward	The Faith Doctor: A Story of New York	1891	Gutenberg	1	
Brown, William Wells	Clotel; or, the President's Daughter	1853	Documenting the South	1	
Dixon, Thomas	The Clansman	1905	Gutenberg	1	
Norris, Frank	McTeague	1899	Gutenberg	1	
Norris, Frank	The Octopus: A Story of California	1901	Gutenberg	1	
Twain, Mark (Samuel Clemens)	Tom Sawyer Abroad	1894	Gutenberg	1	1
Bacheller, Irving	Eben Holden: A Tale of the North Country	1900	Gutenberg	1	
Wister, Owen	The Virginian: A Horseman of the Plains	1898	Gutenberg	1	
Grey, Zane	Riders of the Purple Sage	1912	Gutenberg	1	

Thomson, Daniel P.	The Rangers; Or The Tory's Daughter	1851	Gutenberg	1	
Wilson, Harriet E.	Our Nig	1861	Gutenberg	1	
Page, Thomas Nelson	Gordon Keith	1902	Gutenberg	1	
Stoddard, Elizabeth	The Morgesons	1862	Gutenberg	1	
Page, Thomas Nelson	P'laski's Tunament	1891	Gutenberg	1	
Jacobs, Harriet	Incidents in the Life of a Slave Girl	1861	Gutenberg	1	
Ames, Nathaniel	Old Sailors Yarns	1835	Gutenberg	1	
Fern, Fanny	Ruth Hall	1854	Gutenberg	1	
Twain, Mark (Samuel Clemens)	The American Claimant	1892	Gutenberg	1	1
Chandler, Joel	Uncle Remus	1881	Gutenberg	1	1
Twain, Mark (Samuel Clemens)	The Innocents Abroad	1869	Gutenberg	1	1
Twain, Mark (Samuel Clemens)	The Prince and the Pauper	1882	Gutenberg	1	1

Twain, Mark (Samuel Clemens)	A Connecticut Yankee In King Arthur's Court	1889	Gutenberg	1	1
Harte, Bret	Harte's Condensed Novels	1867	Gutenberg	1	
Melville, Herman	Moby-Dick	1851	Gutenberg	1	
Hawthorne, Nathaniel	The Scarlet Letter	1850	Gutenberg	1	
Rowson, Susanna	Charlotte Temple	1794	Gutenberg	1	
Brockden Brown, Charles	Wieland	1798	Gutenberg	1	
Brockden Brown, Charles	Edgar Huntly	1799	Gutenberg	1	
Twain, Mark (Samuel Clemens)	Sociable Jimmy	1870		1	1
Lovecraft, H.P.	The Shadow over Insmouth	1930	Wikisource	1	1
Dunne, Finley Peter	Mr. Dooley in Peace and War	1898	Gutenberg	1	1
Ade, George	Fables in Slang	1900	Gutenberg	1	1
Lardner, Ring	You Know Me Al A Busher's Letters	1916	Gutenberg	1	1

Bibliography

General Introduction

Baker, Houston A. Jr. *Modernism and the Harlem Renaissance*. University of Chicago Press, 1987.

Baron, Dennis. *Grammar and Good Taste: Reforming the American Language*. New Haven: Yale University Press, 1984.

Best, Stephen, and Sharon Marcus. "Surface Reading: An Introduction." *Representations* 108, no. 1 (2009): 1–21. doi:10.1525/rep.2009.108.1.1.

Blair, Walter. *Native American Humor: 1800-1900*. New York: Harper Collins, 1960.

Blair, Walter and Raven I. McDavid, eds. *The Mirth of a Nation: America's Great Dialect Humor*. Minneapolis: University of Minnesota Press, 1983.

Bonfiglio, Thomas Paul. *Race and the Rise of Standard American*. Walter de Gruyter, 2010.

Bridgman, Richard. *The Colloquial Style in America*. Oxford: Oxford University Press, 1966.

Carkeet, David. "The Dialects in Huckleberry Finn." *American Literature* 51. 1979. 315–32.

Chomsky, Noam. *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. Cambridge: Cambridge University Press, 1966.

Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes." *The Journal of Philosophy* 78, no. 2 (1981): 67-90. doi:10.2307/2025900.

Clark, Andy. *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: The MIT Press, 1997.

Coetzee, J.M. "The English Fiction of Samuel Beckett : An Essay in Stylistic Analysis." PhD. diss., University of Texas Austin, 1969.

Cole, Roger. "Literary Representation of Dialect: A Theoretical Approach to an Artistic Problem" *University of South Florida Language Quarterly* 24 (1986): 3-4. 3-8.

Dillard, J. L. *A History of American English*. Routledge, 2014.

Eisenstein, Elizabeth L. *The Printing Press as an Agent of Change*. Cambridge: Cambridge University Press, 1980.

Fischer-Starcke, Bettina. *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*. Bloomsbury Publishing, 2010.

Fish, Stanley. *Is There a Text in This Class? The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press, 1980.

Fishkin, Shelley Fisher. *Was Huck Black?: Mark Twain and African-American Voices*. First edition. Oxford University Press, 1994.

Foote, Stephanie. *Regional Fictions: Culture and Identity in Nineteenth-Century American Literature*. Madison: University of Wisconsin Press, 2001.

Harris, Roy. *Rationality and the Literate Mind*. Routledge, 2009.

Harris, Roy. *Rethinking Writing*. A&C Black, 2005.

Hayles, Katherine. *Unthought: The Power of the Cognitive Nonconscious*. Chicago: University of Chicago Press, 2017.

Holton, Sylvia Wallace. *Down Home and Uptown: The Representation of Black Speech in American Fiction*. Madison, NJ: Fairleigh Dickinson University Press, 1984.

Horne, Benjamin D, and Sibel Adali. "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News", The 2nd International Workshop on News and Public Opinion at ICWSM, 9.

Ives, Sumner. "A Theory of Literary Dialect." In *A Various Language: Perspectives on American Dialects*, edited by Juanita V. Williamson and Virginia M. Burke, 145–77. New York: Holt, Rinehart and Winston, 1971.

James, Henry. *The Bostonians: A Novel*. London: Macmillan, 1921.

Jones, Gavin. *Strange Talk: The Politics of Dialect Literature in Gilded Age America*. Berkeley: University of California Press, 1999.

Katz, Leonard, and Ram Frost. "The Reading Process Is Different for Different Orthographies: The Orthographic Depth Hypothesis." In *Advances in Psychology*, edited by Ram Frost and Leonard Katz, 94:67–84. North-Holland, 1992. https://doi.org/10.1016/S0166-4115(08)62789-2.

Kittler, Friedrich A. *Discourse Networks 1800/1900*. Stanford: Stanford University Press, 1990.

Kittler, Friedrich A. *Gramophone, Film, Typewriter*. Stanford: Stanford University Press, 1999.

Krapp, George Philip. *The English Language in America*. United States: Century Company, for the Modern language association of America, 1925.

Labov, William. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press, 1972.

Lepore, Jill. *A Is for American: Letters and Other Characters in the Newly United States*. Reprint edition. New York: Vintage, 2003.

Looby, Christopher. *Voicing America: Language, Literary Form, and the Origins of the United States*. University of Chicago Press, 1998.

McLuhan, Marshall. *The Gutenberg Galaxy*. Toronto: University of Toronto Press, Scholarly Publishing Division, 2017.

Meletis, Dimitrios. "The Grapheme as a Universal Basic Unit of Writing." Writing Systems Research 11, no. 1 (January 2, 2019): 26–49. https://doi.org/10.1080/17586801.2019.1697412.

Metzinger, Thomas. *Being No One: The Self-model Theory of Subjectivity*. Cambridge, MA: The MIT Press, 2004.

Michaels, Walter Benn. *The Shape of the Signifier: 1967 to the End of History*. Princeton, NJ: Princeton University Press, 2013.

Minnick, Lisa Cohen. *Dialect and Dichotomy: Literary Representations of African American Speech*. 1st Edition. University Alabama Press, 2007.

Nagel, Thomas. "What Is It Like to Be a Bat?"" *The Philosophical Review* 83, no. 4 (1974): 435-50. doi:10.2307/2183914.

North, Michael. *The Dialect of Modernism: Race, Language, and Twentieth-Century Literature*. New York: Oxford University Press, 1998.

Nurhussein, Nadia. *Rhetorics of Literacy: The Cultivation of American Dialect Poetry*. Columbus: Ohio State University Press, 2013.

Olson, David R. *The Mind on Paper: Reading, Consciousness and Rationality*. Cambridge University Press, 2016.

Ong, Walter J. *Orality and Literacy: 30th Anniversary Edition*. 3rd edition. New York: Routledge, 2012.

Pinker, Steven. *The Language Instinct: How the Mind Creates Language*. Reprint edition. New York: Harper Perennial Modern Classics, 2007.

Rourke, Constance. *American Humor: A Study of the National Character*. New York: New York Review of Books, 2004.

Sampson, Geoffrey. *The 'Language Instinct' Debate*. Revised edition. New York: Continuum, 2005.

Schmandt-Besserat, Denise. *Before Writing, Vol. I: From Counting to Cuneiform*. University of Texas Press, 1992.

Sebba, Mark. *Spelling and Society: The Culture and Politics of Orthography around the World*. Reissue edition. Cambridge University Press, 2012.

Sproat, Richard. *Language, Technology, and Society*. OUP Oxford, 2010.

Stubbs, Michael, Wolfgang Teubert, Michaela Mahlberg and Michael Hoey. *Text, Discourse and Corpora: Theory and Analysis*. New York: Continuum, 2007.

Trachtenberg, Alan. *The Incorporation of America: Culture and Society in the Gilded Age*. New York: Hill and Wang, 1982.

Treiman, Rebecca. "Learning to Spell Words: Findings, Theories, and Issues." *Scientific Studies of Reading* 21, no. 4 (July 4, 2017): 265–76. https://doi.org/10.1080/10888438.2017.1296449.

Tognini-Bonelli, Elena. *Corpus Linguistics at Work*. Netherlands: J. Benjamins, 2001.

Venezky, Richard L. *The American Way of Spelling: The Structure and Origins of American English Orthography*. New York: Guilford Publications, 1999.

Woolley, Samuel C., and Philip N. Howard, eds. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, 2018.

Methodological Introduction

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. "Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach." *Frontiers in Artificial Intelligence* 3 (2020): 73. https://doi.org/10.3389/frai.2020.00073.

Bradley, Stephen. "Quotation Parsing and Speaker Attribution in Narrative Texts." B.A. Thesis, Trinity College Dublin.

Chang, Kent K. and Simon DeDeo. "Divergence and the Complexity of Difference in Text and Culture." *Journal of Cultural Analytics* 4.11 (2020): 1-36. doi: 10.22148/001c.17585

Elson, David K. and Kathleen R. McKeown. "Automatic Attribution of Quoted Speech in Literary Narrative." AAAI Publications, Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

Gamallo, Pablo, José Ramom Pichel, and Iñaki Alegria. "From Language Identification to Language Distance." *Physica A: Statistical Mechanics and Its Applications* 484 (October 15, 2017): 152–62. https://doi.org/10.1016/j.physa.2017.05.011.

Heuser, Ryan and Long Le-Khac. "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method" Pamphlets of the Stanford Literary Lab. 4. 2012.

Hoover, David L. *Language and Style in The Inheritors*. Ann Arbor: The University of Michigan Press, 1999.

Landauer, Thomas K, Peter W. Foltz, and Darrell Laham. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25, no. 2–3 (January 1, 1998): 259–84. https://doi.org/10.1080/01638539809545028.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso, 2005.

Sampson, Geoffrey. *Writing Systems: A Linguistic Introduction*. Stanford: Stanford University Press, 1985.

Shannon, Claude E. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27, no. 3 (1948): 379-423.

Shannon, Claude E. "Prediction and Entropy of Printed English." *The Bell System Technical Journal* 30, no. 1 (1951): 50-64.

Sproat, Richard. *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press, 2000.

Underwood, Ted. "Why Literary Time is Measured in Minutes." *ELH* 85, no. 2 (2018): 341-365. doi:10.1353/elh.2018.0013.

Twain's Orthography in Context

Baker, Houston A. Jr. *Modernism and the Harlem Renaissance*. University of Chicago Press, 1987.

Bonfiglio, Thomas Paul. *Race and the Rise of Standard American*. Walter de Gruyter, 2010.

Briden Earl F. "Twainian Epistemology and the Satiric Design of 'Tom Sawyer Abroad'" *American Literary Realism*, 1870-1910, Fall, 1989, Vol. 22, No. 1 (Fall, 1989). 43-52.

Carkeet, David. "The Dialects in Huckleberry Finn." *American Literature* 51. 1979. 315–32.

Fishkin, Shelley Fisher. *Was Huck Black?: Mark Twain and African-American Voices*. First edition. Oxford University Press, 1994.

Galison, Peter. *Einstein's Clocks and Poincare's Maps: Empires of Time* New York: W.W Norton, 2004.

Jones, Gavin. *Strange Talk: The Politics of Dialect Literature in Gilded Age America*. Berkeley: University of California Press, 1999.

Kittler, Friedrich A. *Gramophone, Film, Typewriter*. Stanford: Stanford University Press, 1999.

Leigh, Philip. "Literary forensics: fingerprinting the literary dialects of three works of

plantation fiction." *Texas Studies in Literature and Language* 54, no. 3 (2012): 357-380.

Lepore, Jill. *A Is for American: Letters and Other Characters in the Newly United States*. Reprint edition. New York: Vintage, 2003.

Lott, Eric. *Love and Theft: Blackface Minstrelsy and the American Working Class*. Oxford University Press, 1995.

Michelson, Bruce. *Printer's Devil: Mark Twain and the American Publishing Revolution* Berkeley: University of California Press, 2006.

North, Michael. *The Dialect of Modernism: Race, Language, and Twentieth-Century Literature*. New York: Oxford University Press, 1998.

Pederson, Lee A. "Negro Speech in the Adventures of Huckleberry Finn." *Mark Twain Journal*, n.d., 5.

Sebba, Mark. *Spelling and Society: The Culture and Politics of Orthography around the World*. Reissue edition. Cambridge University Press, 2012.

Sundquist, Eric J. *To Wake the Nations: Race in the Making of American Literature*. Cambridge, MA: Belknap Press of Harvard University Press, 1993.

Tamasi, Susan. "Huck Doesn't Sound like Himself: Consistency in the Literary Dialect of Mark Twain." *Language and Literature* 10, no. 2 (May 1, 2001): 129–44. https://doi.org/10.1177/096394700101000201.

Twain, Mark. *Adventures of Huckleberry Finn: An Authoritative Text, Contexts, Criticism*. Ed. Thomas Cooley. New York: W.W Norton, 1999.

Twain, Mark. *A Connecticut Yankee in King Arthur's Court: An Authoritative Text, Contexts, Criticism*. Ed. Henry B. Wonham. New York: W.W. Norton, 2018.

Twain, Mark. "Concerning the American Language" in *Tom Sawyer Abroad, Tom Sawyer, Detective and Other Stories*. New York: Harper, 1924. 407-411.

Twain, Mark. *Pudd'nhead Wilson*. New York: Penguin, 1969.

Twain, Mark. *The American Claimant* in *The Gilded Age and Later Novels*. Ed. Hamlin Hill. New York: Library of America, 2002. 457-644.

Twain, Mark. *Tom Sawyer Abroad* in *The Gilded Age and Later Novels*. Ed. Hamlin Hill. New York: Library of America, 2002. 645-740.

Eben Holden, Plantation Literature and the Fate of Dialect

Bacheller, Irving. *Eben Holden: A Tale of the North Country*. New York: Lothrop, 1900.

Bacheller, Irving. *Eben Holden's Last Day A-Fishing* New York: Harper and Brothers Publishing, 1907.

Bacheller, Irving. *From Stores of Memory*. New York: Farrar & Rinehart, 1938.

Baker, Houston A. Jr. *Modernism and the Harlem Renaissance*. University of Chicago Press, 1987.

Brodhead, Richard H. *Cultures of Letters: Scenes of Reading and Writing in Nineteenth-Century America*. Chicago: University of Chicago Press, 1993.

Brodhead, Richard H. "Introduction" in *The Conjure Woman, and Other Conjure Tales*. Ed. Richard H. Brodhead. Raleigh: Duke University Press, 1993.

Brookes, Stella Brewer. *Joel Chandler Harris, Folklorist*. Greece: University of Georgia Press, 2009.

Chesnutt, Charles Waddell. *The Conjure Woman, and Other Conjure Tales*. Ed. Richard H. Brodhead. Raleigh: Duke University Press, 1993.

Foote, Stephanie. "The Cultural Work of American Regionalism." In *Blackwell's Reader in Regional Literatures of the United States*. Ed. Charles L. Crow. London: Blackwell, 2003.

Foote, Stephanie. *Regional Fictions: Culture and Identity in Nineteenth-Century American Literature*. Madison: University of Wisconsin Press, 2001.

Gale, Zona. *Friendship Village*. New York: MacMillan, 1908.

Garland, Hamlin. *A Little Norsk: Or Ol' Pap's Flaxen* New York: D. Appleton and Company, 1893.

Gilligan, Heather Tirado. "Reading, Race, and Charles Chesnutt's "Uncle Julius" Tales." *ELH* 74, no. 1 (2007): 195-215. http://www.jstor.org/stable/30029551.

Harris, Joel Chandler. *Uncle Remus His Songs and His Sayings*. New York: D. Appleton and Company: 1898.

Holton, Sylvia Wallace. *Down Home and Uptown: The Representation of Black Speech in American Fiction*. Madison, NJ: Fairleigh Dickinson University Press, 1984.

Jewett, Sarah Orne. *The Country of the Pointed Firs and Other Stories*. New York: Signet Classics, 2009.

Johanningsmeier, Charles. *Fiction and the American Literary Marketplace: The Role of*

Newspaper Syndicates in America, 1860-1900. Cambridge: Cambridge University Press, 2002.

Kittler, Friedrich A. "Dracula's Legacy" in *Literature, Media, Information Systems* ed. John Johnson. Amsterdam: OPA, 1997. 50-85.

Lovecraft, H. P. *The Call of Cthulhu and Other Weird Stories*. Ed. S.T. Joshi. New York: Penguin Publishing Group, 2016.

Mott, Frank Luther. *Golden Multitudes: The Story of Best Sellers in the United States*. New York: Macmillan Company, 1960.

North, Michael. *The Dialect of Modernism: Race, Language, and Twentieth-Century Literature*. New York: Oxford University Press, 1998.

Rotman, Brian. *Becoming Beside Ourselves: The Alphabet, Ghosts, and Distributed Human Being*. Durham: Duke University Press, 2008.

Orthographic Extrema

Brown, Charles Brockden. *Ormond, or, The Secret Witness*. New York: Broadview Press, 1999.

Child, Lydia Maria. *Hobomok* in *Hobomok and Other Writings on Indians*. Ed. Carolyn L. Karcher. New Brunswick: Rutgers University Press, 1986.

Cooper, James Fenimore. *The Deerslayer, or, The First War-Path*. New York: Modern Library, 2002.

Fiedler, Leslie A. *Love and Death in the American Novel*. Dalkey Archive Press, 1997.

Kolmogorov, A. N. "Three approaches to the quantitative definition of information." *Problemy Peredachi Informatsii*, 1 no. 1 (1965): 3–11.

Minnick, Lisa Cohen. *Dialect and Dichotomy: Literary Representations of African American Speech*. 1st Edition. University Alabama Press, 2007.

Pound, Louise. "The Dialect of Cooper's Leather-Stocking," *American Speech* Vol. 2, No. 12 (Sep., 1927): 479-488.

Shannon, Claude E. "Prediction and Entropy of Printed English." *The Bell System Technical Journal* 30, no. 1 (1951): 50-64.

Suzuki, Ryuji, John R. Buck, and Peter L. Tyack. "Information Entropy of Humpback Whale Songs." *The Journal of the Acoustical Society of America* 119, no. 3 (March 1, 2006): 1849–66. https://doi.org/10.1121/1.2161827.

Twain, Mark. "Fenimore Cooper's Literary Offences" in *Humorous Stories and Sketches*. United States: Dover Publications, 2012. 59-69.

Coda

Hartman, Charles O., and Hugh Kenner. *Sentences*. United States: Sun & Moon Press, 1995.

Heidegger, Martin., Macquarrie, John., Robinson, Edward. *Being and Time*. New York: HarperCollins, 2008.

Kiverstein, Julian and Michael Wheeler, eds. *Heidegger and Cognitive Science*. New York: Palgrave Macmillan, 2012.

Simondon, Gilbert. *On the Mode of Existence of Technical Objects*. New York: Univocal Publishing, 2017.

Walpole, Jane Raymond. "Eye Dialect in Fictional Dialogue." College Composition and Communication 25, no. 2 (1974): 191–96. https://doi.org/10.2307/357177.