Voicing Imperial Subjects in British Literature: A Corpus Analysis of Literary Dialect, 1768-1929

David West Brown

Thesis submitted for the Degree of Doctor of Philosophy

Department of Linguistics and English Language
Lancaster University

2016

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Paul Baker. This thesis has not only been shaped by his attentive feedback, but it stands upon a foundation of work that he and others in the department at Lancaster have built. Without that, this thesis would not have even been possible. I would also like to thank my former colleagues at the Centre for English Language Education and the University Town Writing Programme at the National University of Singapore. Although I have moved on since this project began, Paul Nerney and others were the first to encourage me to pursue it. Finally, my family – and particularly my partner Jasmine – has been my unwavering support throughout a process complicated by jobs and travels and life's predictably unpredictable disruptions. I love you all.

Declaration

I declare that this thesis represents my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualification.

Abstract

This study investigates nonstandard dialect as it used in fictional dialogue. The works included in it were produced by British authors between 1768 and 1929 – a period marking the expansion and height of the British Empire. One of the project's aims is to examine the connections among dialect representation and the imperial project, to investigate how ventriloquizing African diasporic, Chinese, and Indian characters works with related forms of characterization to encode ideologies and relations of power. A related aim is to explore the emergence and evolution of these literary dialects over time and to compare their structures as they are used to impersonate different communities of speakers.

In order to track such patterns of representation, a corpus was constructed from the dialogue of 126 novels, plays, and short stories. That dialogue was then annotated for more than 200 lexical, morphological, orthographic, and phonological features. That data enable statistical analyses that model variation in the voicing of speakers and how those voicings change over time. This modeling demonstrates, for example, an increase in the frequency of phonological features for African diasporic dialogue and a countervailing decrease in the frequency and complexity of coded features generally for Indian dialogue.

Trends like these that are surfaced though quantitative methods are further contextualized using qualitative, archival data. The analysis ultimately rests on connecting patterns of representation to changes in the imperial political economy, evolving language ideologies that circulate in the Anglophone world, and shifts in sociocultural anxieties that crosscut race and empire. The combined quantitative and qualitative analyses, therefore, expose representational systems – the apparatuses that propagate structures and the social attitudes that accrue to those structures. It further demonstrates that in such propagation, structures and attitudes are complementary.

Table of Contents

List of Figures

List of Tables

List of Appendices

Chapter 1
Introduction

1.1 Introduction

This thesis has its origins in a presentation that I gave more than a decade ago at the Modern Language Association convention. As part of that presentation, I outlined a kind of cross-racial and cross-linguistic performance that I termed "linguistic minstrelsy" (Brown, 2005). Linguistic minstrelsy, I proposed, was the practice of white Americans burlesquing or appropriating African American English. My focus was on instances of this practice in television and film, particularly as it is used in the performance of masculinity and as an instrument of self-actualization. However, I also endeavored to link more current examples in movies like *Bulworth* to histories of mimicry not just in performance, but also on the page. Thus began my interest in literary dialect.

In the intervening years, I also became increasingly familiar with corpus linguistics and began thinking about ways of using corpus approaches to explore more thoroughly the kinds of linkages that I had only broadly sketched out in my initial analysis. Additionally, I spent time living and working in Southeast Asia, where I was exposed to debates about identity and identity performance in writing outside of North American contexts. The thesis that has formed out of those experiences and curiosities is an investigation of literary dialect that is much expanded from its original incarnation.

The thesis uses computational techniques to model literary dialect structure, structural variation, and changes over time. Also, rather than examining representations of only a single vocal culture, it investigates representations of three: African diasporic, Indian, and Chinese. The data are drawn from British novels, plays, and short stories that were published between 1768 and 1929. The analysis aims not only to describe structural patterns in that data, but also to explore their intersections with ideologies related to language and empire. In setting out the specific scope of the project, this chapter begins with a brief discussion of literary dialect and race in the eighteenth century, a discussion that serves to frame the project and also to explain the date that is the jumping off point for the corpus – 1768. The next section presents the research questions. Those are followed by two sections that establish, in turn, this

study's approach to quantitative and qualitative analysis. In addition to descriptions, these sections include some illustrations of the analysis itself. These illustrations are only brief profiles, but they more clearly articulate this study's goals and orientation to data than the descriptions by themselves. Next, there is a short explanation of why this study does not engage in evaluations of literary dialect's accuracy, but instead focuses on its representationality. The chapter ends with an outline of the thesis' organization.

1.2 Literary dialect and race in the eighteenth century

Figure 1.1: A close-up of "The Nabob Rumbled or A Lord Advocates Amusement" by James Gillray with the full print inset at bottom (image courtesy of the British Museum, ©Trustees of the British Museum).

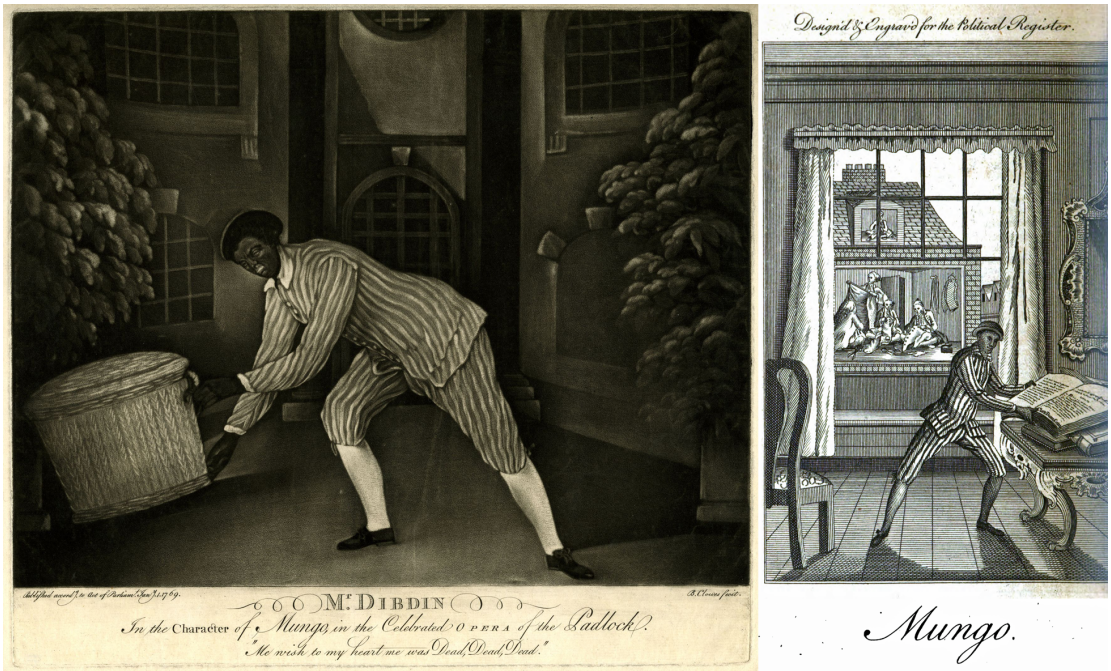
The 1783 engraving "The Nabob Rumbled, or A Lord Advocates [*sic*] Amusement" by James Gillray (see Figure 1.1) is critique of British greed in India and the nouveau riche equation of wealth with manners (Smylitopoulos, 2011). In it, we see Sir Thomas Rumbold vomiting guineas into a pot held by Henry Dundas, Lord Advocate of Scotland. Gazing at the stream of money, Dundas remarks, "I weel tak them to Lochabar and wash them in the Brook." The use of nonstandard spelling and grammar to mimic regional accents like Dundas' Scottish one has a long history in English, of course. The Scottish playwright Allan Ramsay (1725) used the very same respellings (*weel* for *well* and *tak* for *take*) in *The Good Shepherd* more than a half-century earlier:

(1) Daft are your dreams, as daftly wad ye hide
Your weel-seen love, and dorty Jenny's pride:
Tak courage, Roger, me your sorrows tell,
And safely think nane kens them but yoursel.

Ramsay's narrative use of such techniques – what has come to be called literary dialect – traces its roots at least to the fifteenth century and the northern English of the clerks in Geoffrey Chaucer's "A Reeve's Tale" (Elliott, 1974; Nielsen, 2005).

But Gillray also ventriloquizes another figure using literary dialect – the Indian servant on the back of the elephant piled with rupees. While he and the British driver gallop off out of frame, the servant exclaims, "Me and Massa leave England He! He! He!" While regionalized literary dialects had been around for hundreds of years at this point, racialized literary dialect was a relatively new phenomenon in the late eighteenth century. The very construct of "race," Wheeler (2000) argues, was evolving during the early empire, as social and cultural taxonomies that foreground religion were replaced by those that foreground skin color. And it is at this same moment that we begin to find examples like Gillray's voicing of the Indian servant.

Figure 1.2 (left): Portrait of the actor Charles Dibdin as Mungo (image courtesy of the British Museum, ©Trustees of the British Museum). **Figure 1.3 (right)**: Jeremiah Dyson caricatured as Mungo from *The Political Register, and Impartial Review of New Books* (Almon, 1769, p. 193).



Such racialized voices are the subject of this study. More specifically, it focuses on their appearance in British novels and plays from 1768 to 1929. The year 1768 marks the premier and publication of Isaac Bickerstaff's *The Padlock*. This play features Mungo, one of the earliest African diasporic characters whose dialogue is rendered in literary dialect – a nineteenth century history of "the stage negro" calls him "the first of his race" (Hutton, 1889, p. 132). Here is an example:

(2) Go, get you down, you damn hamper, you carry me now. Curse my old Massa, sending me always here and dere for one something to make me tire like a mule – curse him imperance – and him damn insurance.

According to Charles Dibdin, who performed the role in blackface (see Figure 1.2), the inspiration for Mungo's dialogue came from John Moody, an acquaintance of Bickerstaff's, who had spent time in the Caribbean "and knew, of course, the dialect of the negroes" (Kitchiner, 1823, p. 13). Following the play's debut, Mungo was a sensation and the name became a kind of generic index for any African diasporic person (much like Sambo), as well as a comic sobriquet for a number of public figures including the politician Jeremiah Dyson (see Figure 1.3). As a watershed both in form and influence, *The Padlock* makes a rational starting point.

The study goes on to cover roughly the period that is sometimes referred to as "the long nineteenth century" (e.g., Hobsbawm, 1987) – a period of British imperial expansion and the beginnings of decline. Included in the study are representations of African diasporic, Chinese, and Indian speakers. As the authors are not members of the speech communities they mimic or ventriloquize, one of this study's primary areas of interest is the ways in which writers of imaginative works use dialect to represent the linguistic, social, cultural, and ethnic Other. Such representations are predicated on acts of linguistic appropriation (i.e., the adoption of a dialect or style that is not part of a speaker's or writer's customary linguistic repertoire) and the use of nonstandard orthographies, morphosyntax, and lexicons. That appropriation has implications for the identities, ideologies, and power relations instantiated by literary dialects – implications that can be complex and sometimes contradictory. On the one hand, written representations of nonstandard varieties have the potential to subvert dominant language ideologies (Jaffe, 2000). They can be powerful resources in asserting complex (and sometimes highly individual) identities that are otherwise rendered invisible by standard language (Gupta, 2000). On the other hand, the social signaling of nonstandard features in the creation of literary dialect "is dependent upon the existence of certain social conventions and stereotypes" (Zanger, 1966, p. 40). Exploring the patterns and structures of literary dialects, therefore, engages "twin issues that are central to sociolinguistic inquiry: linguistic variation and linguistic inequality" (Jaffe, 2000, p. 498).

1.3 Research questions

In its investigation of literary dialects, the study addresses the following:

- The over-arching research question is: how is literary dialect used as an imaginative tool to represent the language of African diasporic, Chinese, and Indian speakers?

  This breaks down into 4 more specific operable questions:

  - What are the patterns of lexical, morphosyntactic, orthographic and phonological features that distinguish specific, imagined language varieties?
  - In what ways, if any, do such patterns evolve over time?
  - To what extent and in what ways are there any shared patterns of features between or among varieties?
  - How are patterns of linguistic representation implicated in evolving understandings of race, culture, and empire?

These questions can be separated into two broad categories: those that address structure (the first three questions) and those that address the social meanings that accrue to those structures (the fourth question). For the former, a corpus was compiled of dialogue from 126 novels and plays. That data were coded according to a taxonomy that groups features into four primary, superordinate categories: lexical (or features related to vocabulary), morphosyntactic (or features related to grammar and grammatical marking), orthographic (or variant spellings that have no relationship to nonstandard pronunciation), and phonological (or variant spellings that are imitations of accent). Those categories organize more than 200 different feature types. Although the corpus, itself, is modest in terms of its token count (a little more than 50,000 words), the data were assigned more than 18,000 codes and produced a matrix with almost 30,000 points. The analysis of that data matrix focuses on measurements along three dimensions: frequency, complexity, and similarity. The goal of the analysis is to identify synchronic and diachronic patterns, in the pursuit of which it employs a variety of statistical techniques including composite frequencies, diversity indices, dispersion measures, regression analysis, and cluster analysis. Such analysis builds from previous corpus-based research into literary dialect that has spotlighted the practices of a single author or small group of authors (e.g., Burkette, 2001; Minnick, 2007; Tamasi, 2001). It also bridges that work with scholarship in the digital humanities that has used computational techniques to model diachronic changes in genre, theme, and authorial style (e.g., Jockers, 2013; Moretti, 2005).

The second set of questions relate to the connections between dialect representation and the imperial project – investigating how ventriloquizing African diasporic, Chinese, and Indian characters works with other forms of representation to encode ideologies and social relations. This latter analysis rests on linking patterns of representation to changes in the imperial political economy, evolving language ideologies that circulate in the Anglophone world, and shifts in sociocultural anxieties that crosscut race and empire. Drawing those connections necessitates extensive reading of the imperial archive. Thus, the project combines the quantitative analysis of corpus approaches with sociolinguistics and colonial discourse analysis (or the study of Western representations of the non-Western Other, what Anand (2007, p. 23) terms "a critical political analysis of the Western imagination").

1.4 Corpora and culture

In order to demonstrate the potential connections between quantitative linguistic patterns and cultural currents, as well as to outline this study's specific approach to computational data, I want to consider a couple of illustrative trends. Both examples are drawn from the Google Books data tables.[1] To be clear, this is not the data that are used for the study, but they provide a convenient jumping off point for a discussion of these issues. That data function as a large (if not entirely unproblematic) historical corpus, which can be truncated to the approximate timeframe covered by the study (1770-1929). Further, its English language data can be separated into British, American, and Fiction sub-corpora containing 49,469,663,940 words, 65,235,115,030 words, and 6,136,772,010 words, respectively.

Although the Google Books data are simply expedient for the purposes of demonstration, the two tokens that serve as examples are, as will become clear, important constituents of the literary dialect under study: *MASSA* and *SAHIB*.[2] Both are address forms – the first primarily, though not exclusively (as the Gillray print

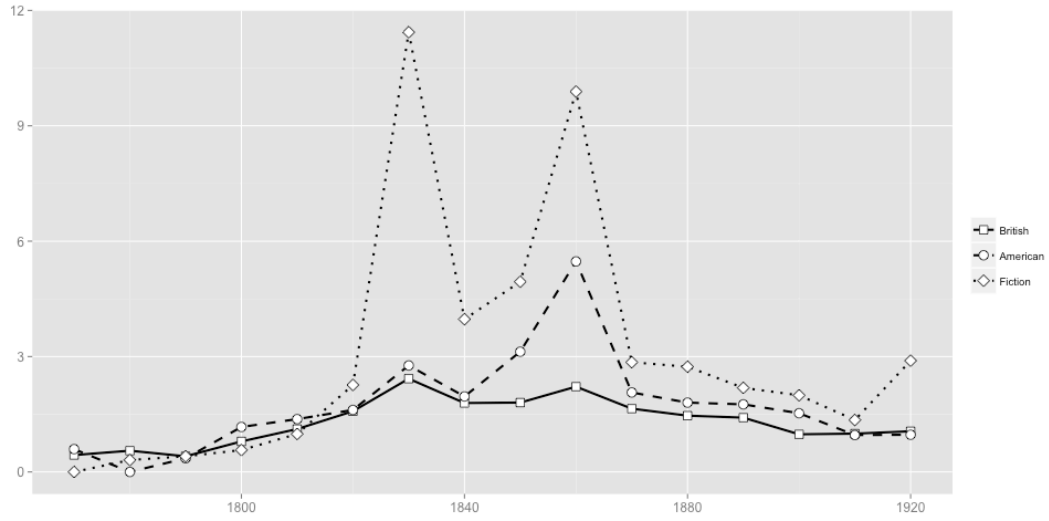
[1] The data from Google Books presents a host of problems. It comes in large data tables without any surrounding context. It is generated through automated optical character recognition, introducing error into both the data and the metadata, and it is not controlled for representativeness, so writers whose works are reprinted have their language replicated and its salience amplified. Matthew Jockers (2013) nicely summarizes these critiques. Mark Davies (2013), however, uses some case studies to show that Google Books can compare favorably to the Corpus of Historical American English, primarily because its tremendous size is capable of reducing error to near zero.

[2] Italicized small capitals indicate that a token is lemmatized. For example, *LOVE* would include all forms of that token (i.e., *love*, *loves*, *loved*). Regular italics (e.g., *love*) indicates a specific token in a specific form (i.e., only *love*).

illustrates), associated with fictive African diasporic dialogue and the second with Indian dialogue. The trends that they generate prompt just the kind of questions that advocates of such lexically based inquiry promote, whether using the Google Books data (Michel et al., 2011) or another historical corpus like the Corpus of Historical American English (Davies, 2012): What, for example, does it mean that the frequency of the word *god* has declined over the last two hundred years? Or that *teenager* has increased in the last fifty?

In the case of *MASSA* (see Figure 1.4), the plots produce a number of provocative inflection points. In British English, the frequency peaks in the 1830s, which aligns with the passage of the Slavery Abolition Act. The highest frequency in American English occurs in the 1860s, which coincides with the American Civil War and the passage of the Thirteenth Amendment outlawing slavery. The two peaks in English fiction mirror those in the national Englishes, only extending more dramatically. In this particular case, at least, it would appear that fluctuations in frequency correspond to historical periods of disruptions in established racial hierarchies.

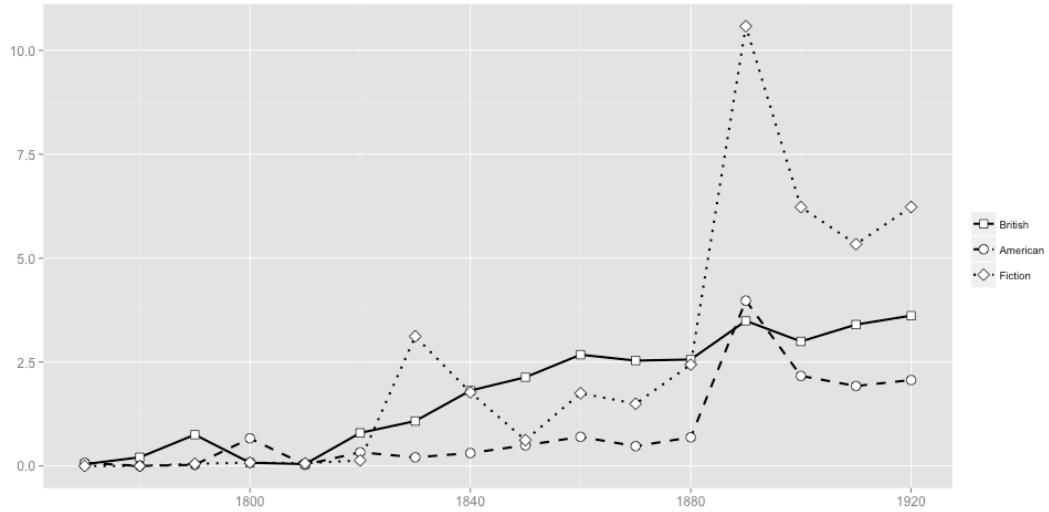
Figure 1.4: Frequencies (normalized per million words) of lemmatized *MASSA* in the Google Books data tables from 1770-1930.



The trends for *SAHIB* (see Figure 1.5) have very different trajectories than those for *MASSA*. There is an increasing frequency in British English that parallels the growth of Britain's imperial involvement in India over the course of the nineteenth century. The frequency in American English mostly trails the frequency in British English, which seems predictable in light of the fact that the United States was less politically, economically, and socially tied to India. The frequency in fiction sees a

small spike in the 1830s then a larger one in the 1890s. The smaller peak in the 1830s coincides with fundamental changes in the relationship between Britain and India codified by the passage of the Charter Act of 1833 and the English Education Act of 1835.

Figure 1.5: Frequencies (normalized per million words) of lemmatized *SAHIB* in the Google Books data tables from 1770-1930.



The larger peak may seem less immediately explainable by contemporaneous political, military, or economic events. It does, however, align with the growth of juvenile fiction and its genric demand for exotic locales, military adventures, and imperial heroes. Events in India, like the Indian rebellion of 1857 (sometimes called the Sepoy Mutiny), were particularly common sources of inspiration for adventure stories during the 1890s (Erll, 2006). Other types of Anglo-Indian fiction were similarly gaining in popularity during the decade, most notably the works of Rudyard Kipling like *The Jungle Book*. The frequencies of *SAHIB*, therefore, appear to have been mediated by changes in consumption and circulation – by the emergence of new reading populations like juveniles, the development of genres meant to appeal to them, and the tropes, character-types, and plots that go along with those genres. Additionally, frequencies are likely influenced by prominent authors and wider cultural and artistic engagement with their language and themes.

This analysis is, of course, just a thumbnail sketch. Still, in its broad outlines, it hints at the rich confluence of factors – material, aesthetic, and ideological – that can contextualize diachronic changes in feature frequencies. As interesting as such trajectories are, however, the motivating idea for this study was to model multivariate data rather than looking at a single variable (like the changing frequency of a word).

A multivariate approach treats literary dialect as a system. As such, we can track changes to literary dialect practices as a whole. Does literary dialect realize more or fewer features over time? Does it become more simple? Perhaps as a result of conventions fossilizing into a kind of stereotypical shorthand. Or more complex? Maybe because of authors' use of an ever greater array of available templates and developing expectations for verisimilitude.

Additionally, a systemic approach enables comparisons of features or groups of features. How, for example, do the frequencies of lexical features (like forms of address) compare to frequencies of features used to imitate phonology (like the *t*/*d*-for-*th* substitution in *dere* for *there*)? Do the relative relationships between such classes of features remain stable or change over time? Similarly, we can compare the patterns of features used to represent groups of speakers or the practices of specific authors. Remember in the Gillray print, the artist uses *massa* as an address from when he ventriloquizes the Indian servant. The question might be raised, then, whether there are structural similarities between eighteenth century representations of Indian voices and African diasporic ones that go beyond a shared lexical item. Related questions could be raised about the influence of authors and whether or not they leave stylistic traces in the practices of others who follow them.

1.5 Reading the "imperial archive"

The quantitative analysis, however, only tells part of the story. As the examples from Figures 1.4 and 1.5 demonstrate, computational techniques can point to intriguing patterns, but those patterns call out for explanation. For such explanations this study marshals qualitative evidence from the imperial archive. The "imperial archive" can be an elastic term, but here it is posited in its broadest sense to include documents and images produced in domestic Britain and the colonies that engage with the empire practically or notionally (see, e.g., Richards, 1993). This includes artifacts as wide ranging as speeches delivered in Parliament, Bibles translated into native languages by missionaries, travelogues produced for domestic consumption, and, of course, the literary works that are the sources for the corpus.

One way to think about the relationship between the quantitative and qualitative analysis is to imagine each artifact in the imperial archive as node in a vast network. Those nodes are connected by themes, tropes, and ideas all constituted though discourse. At first, however, all of those nodes and connections are darkened.

The computational analysis lights up connections among the source works and in those connections are clues – structural throughlines and disruptions, likenesses and contrasts. From those clues, tracing possible connections to additional nodes, requires reading new artifacts in a process similar to what Foucault (1972) describes in *The Archeology of Knowledge*. In its simplest terms, the analysis begins with a machine identification of patterns, which, in turn, informs a human one.

In order to illustrate both the process and the potential of the latter, let us consider some dialogue and surrounding description from one of the source works. The dialogue comes from Robert Ballantyne's (1888) *The Middy and the Moors*, and is attributed to Mrs. Lilly, an African diasporic slave. Here is a short sample:

(3) No, massa, I don' know of no white slabe as hab took refuge wid any ob our neighbours.

In total, the African diasporic dialogue in Ballantyne's adventure novel realizes 42 different features, but I will highlight only a few here, which are evident in (3): *t*/*d*-for-*th* substitution (e.g., *wid* for *with*), *b*-for-*v*/*f* substitution (e.g., *slabe* for *slave*), and address (e.g., *massa*). Those features have long histories of association with impersonations of African diasporic vocal culture. Below is a 1747 letter circulated in newspapers both in the North American colonies and in Britain – a letter which is supposedly written by a slave in Maryland to his "Masser Frankee" (see Figure 1.6).

Figure 1.6: A letter circulated both in North American and British newspapers that purports to be written by a slave in Herring Bay, Maryland ("Domestic news," 1747).

From the *MARYLAND GAZETTE*.

The following LETTER and Poſtſcript, is aſſur'd to be the Genuine Copy of one ſent by an old Negro Man at Herring Bay, to his Maſter. Pieces of this Nature are entertaining to moſt Readers who can enter into the Humour of them; and as there appears in this a great deal of Sincerity, and pure natural Simplicity, we cannot avoid giving it a Place.

Maſſer Frankee, From de grate Houſe. Merrylun.

WE oll berry well onely oal Sew Ded. and our oal blak Hoſs *Cudge*. God bleſs you Maſſer *Frankee*.

From your oal Negur.

TOBY.

Poaſs. My Wiff member he luff to you Maſſer *Frankee*. he ſay he got wun big bag of Puttato for you. and me got Tree oal Stockkin full Cheznutt for you touo. laſs Nile puttem in Chimne for Smoke berry well dat Time you cum. me dahter *Benus* ſay he got too mokkin Burd bigg oll wun cum Oalwun for you. ſhe ſay you pleeſe bring him wun topnot. and Maſſer *Frankee* bring fore Duſſen Pipe for me. me mak poor cropp Corn dis yeer. dat dam oal Hoſs *Cudge* fore dy broke Corn feel fenſe. he pooll down oll me Corn. Won dam oal Buk gump in Weet Patſh cberry Nite. poor oal Wooman no mak Pee nuff for by wun Callico Peticote dis yee. and Mole rute upp oll dahter *Benus* Putato. Maſſer *Richard* ond *Quaſh* member he luff to you. he ſay he got too clebber young Flyen Quurrel an wun oal Poſſum and fix young wun for you. God bleſs you Maſſer *Frankee*. Wun Saylor Man tell me you bin haff Small Pox. I berry glad you well. tak care you no ketch toder Maſſer *Frankee*.

From your oal Man.

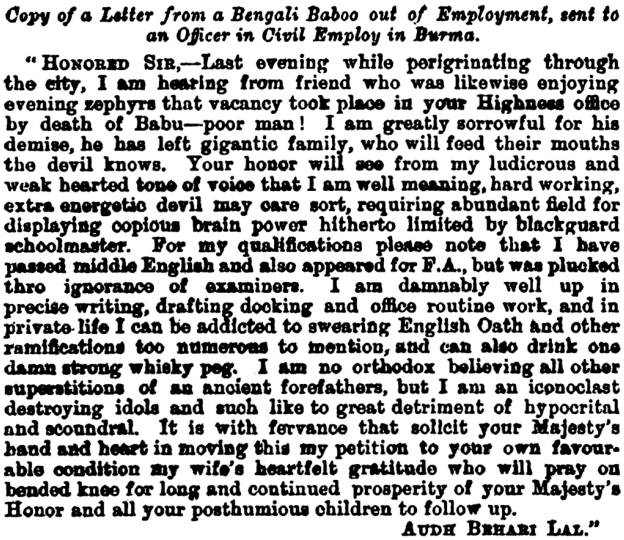
TOBY chew.

Like Mrs. Lilly's dialogue, the letter realizes *t*/*d*-for-*th* substitution (e.g., *dat* for *that*), *b*-for-*v*/*f* substitution (e.g., *ebery* for *every*), and address (e.g., *masser*), among its features. The connections between the texts, however, go deeper than their shared features. Both link similar social values to language. The preface to the letter positions linguistic difference as "entertaining" and as a source of "humour." Additionally, it connects the linguistic features present in the letter to specific traits: "Sincerity" and "pure natural Simplicity." Thus, the letter writer's linguistic practices are figured not only as a mark of difference, but also one of romanticized deficiency.

Ballantyne describes Mrs. Lilly's voice in strikingly similar terms – as possessing "childlike simplicity" – and both the novel and the letter are participating in a larger pattern. Over time, texts from a wide variety of genres including travel narratives, missionary tracts, magazines, and fiction juxtapose dialogue of African diasporic speakers with descriptions of speakers' "good-humor and simple interest" (Parrish, 1908, p. 26), "simplicity and lack of learning" (Murphy, 1899, p. 663); and with metadiscourse regarding "the simple language of these poor blacks" (Ramsden, 1841, p. 240), "the most simple negro dialect" (Barnard, 1882, p. 305), their "own simple language, most pathetically expressed" (The London Missionary Society, 1847, p. 164), "their simplicity, peculiar dialect, broken English, and quaint similes" (Baptist Missionary Society, 1832: 85). It is through such patterns of exemplification and recirculation that social meanings accrue to ventriloquized voices – a process Agha (2003, 2007) terms enregisterment.

The processes of exemplification and recirculation are carried in an array of text-types and artifacts. The Gillray print and the Toby letter are by no means unusual. Artists like George Cruikshank and Gabriel Shear Tregear make frequent use of literary dialect in their caricatures. Parodic letters are similarly common forms of racist humor throughout the period covered by this study. They also mimic a range of speakers and are printed in publications as diverse as *The Spectator* (see Figure 1.7) and the *British Bee Journal* (see Figure 1.8).

Figure 1.7 (left): A letter printed in *The Spectator* purportedly written by a "Bengali Baboo" (An Anglo-Indian, 1907). **Figure 1.8 (right)**: Excerpts from a letter printed in the *British Bee Journal* supposedly written by a Chinese-American beekeeper (Lung, 1893).



Copy of a Letter from a Bengali Baboo out of Employment, sent to an Officer in Civil Employ in Burma.

"HONORED SIR,—Last evening while perigrinating through the city, I am hearing from friend who was likewise enjoying evening zephyrs that vacancy took place in your Highness office by death of Babu—poor man! I am greatly sorrowful for his demise, he has left gigantic family, who will feed their mouths the devil knows. Your honor will see from my ludicrous and weak hearted tone of voice that I am well meaning, hard working, extra energetic devil may care sort, requiring abundant field for displaying copious brain power hitherto limited by blackguard schoolmaster. For my qualifications please note that I have passed middle English and also appeared for F.A., but was plucked thro ignorance of examiners. I am damnably well up in precise writing, drafting docking and office routine work, and in private life I can be addicted to swearing English Oath and other ramifications too numerous to mention, and can also drink one damn strong whisky peg. I am no orthodox believing all other superstitions of an ancient forefathers, but I am an iconoclast destroying idols and such like to great detriment of hypocrital and scoundral. It is with fervance that solicit your Majesty's hand and heart in moving this my petition to your own favourable condition my wife's heartfelt gratitude who will pray on bended knee for long and continued prosperity of your Majesty's Honor and all your posthumious children to follow up.

AUDH BEHARI LAL."

A CHINAMAN'S BEE-KEEPING.

Mister Newsplaper Man,—Me long time thinkee send you of my sugar-flies—Melican man call him "bee," eh?

Well, in flirst place, you sabbe I come from China where I was born, and work in lice fields, and thlee year washee clothes in Slan Flanclisco. Bimeby we laise legitables close by near Oakland. One day heep lot sugar-fly come my house and go in tea-box. My plartner, Jim, he no likee—too muchee bite. I likee sugar-fly heep muchee—him sugar-fly belly nice, heep sweet.

I fixee tea-box in sun and watchee honey-fly go and come. Thlee week him heep fly out—me thinkee allee my sugar-fly go away. What for? Me no sabbe. Me belly good to him. No hurtee him.

• • • • • • • •

What you thlinkee me as a sugar-fly-keeper? You likee hear flom me—me give you lot news. Me likee to hear 'bout Dr. Piller, Mr. Doonothing, Mr. Loot, Plofessor Clook, Mr. Gleen, Dr. Tlinker, and Hutchlinslon, Slecor, Heddlon, Dlibbern, Flance, Lallabee, Dlemalee, and other big sugar-fly writers. Me hopee they will be glad to hear flom me, and enjoy me experience as I enjoy theirs.—WONG LUNG, *Slan Flanclisco, Clalifornia.—Am. Bee Journal.*

These are all nodes in that vast network that I alluded to earlier, and they intersect with the source works in ways that are sometimes direct (like the replication of metalinguistic evaluations of "simplicity"), but are sometimes more opaque. For example, one of the unexpected attributes of eighteenth century reviews of *The Padlock* is their infrequent mention of Mungo's dialect. This is partly because regional and national nonstandard dialect (like Francized English of Dr. Caius in Shakespeare's *The Merry Wives of Windsor*) was an established staple of theatrical comedy. Additionally, as the letter in Figure 1.7 demonstrates, mimicry of African diasporic voices had already been circulating for decades by the time *The Padlock* is first performed. Mungo's dialect, therefore, is not *sui generis*, but is produced in a context of existing written parodies – and likely ephemeral, spoken ones, as well. The credit the play receives in the nineteenth century for its linguistic innovation (as well as its position as the initial text in the corpus) is at least partially the result of its fame and longevity.

While the Toby letter provides evidence for the existence of a specific kind of literary dialect, the "Bengali Baboo" letter (see Figure 1.8) illustrates the reiteration of particular ideological and linguistic debates. Its conceit is that it has been submitted by a reader (much like the Toby letter) in order to "enlighten many English, in and outside the administration of the country, as to the real value of the many thousands of Bengalis whom we have educated in our schools and Colleges throughout Bengal" (An Anglo-Indian, 1907). In other words, it is offered up as a critique of British

education in India, proof of the sycophantism and intellectual limitations of its subjects. A reply that is published in the following issue objects to that characterization. The writer asserts “that the teaching in Indian schools and Colleges is good, and that the proficiency of students in the subjects taught is remarkable” (Markby, 1907). Such contestation over the purpose and effectiveness of English language education in India harkens back to the debates surrounding the English Education Act of 1835 (which are discussed at greater length in chapter 6) and has a harbinger in an objection to William Neale’s representation of Indian vocal culture in his 1833 novel *The Port Admiral* (which is introduced in chapter 4).

Finally, all three letters suggest the importance of global, Anglophone circulation in propagating representations and figuring their social meanings. The “Bengali Baboo” letter ostensibly comes from an expatriate in India via a relative in Burma, and both the Toby letter and the beekeeping letter (see Figure 1.9) originally appear in North American publications, before being reprinted in Britain. In the case of the beekeeping letter, even the specific location of its purported origin in North America is salient. San Francisco was an important catalyst – as a site of both cultural contact and publication – for changing representations of Chinese speakers in the nineteenth century (a topic that is explored in chapter 7). This influence is clearly evident, for example, in a source work like *The Shadow of Quong Lung* (1900), which takes San Francisco’s Chinatown as its setting. *The Shadow of Quong Lung* also points to another aspect of global circulation: it applies to people as much as it does to texts. Its author, Charles William Doyle, traveled widely – like many of the authors whose works are included in the corpus. Doyle was born in India and educated in Britain, where he began a medical practice. Eventually, he moved to California. There, he was inspired by the American author and editor Ambrose Bierce, to whom he dedicated his novel.

The avenues of influence are hardly one-way, however; they flow as much outward from Britain and its writers as into them. *The Padlock*, for example, was widely performed in North America, most notably by the theater company run by the Hallam brothers, William and Lewis. Lewis Hallam’s performance of Mungo – the genesis of blackface minstrelsy in North America – was celebrated in the eighteenth century (Gibbs, 2014). A short item in the Boston newspaper the *Columbian Centinel* ("A Negro," 1792) reports that a slave, after seeing the play, is supposedly convinced that Mungo is a member of the Igbo tribe, that he is authentically African. “In truth,”

the story concludes, “it is impossible that the negro can be personated with more appropriate accent and gesture than by Mr. Hallam in that character.”

As these short examples illustrate, artifacts from the imperial archive can intersect with and inform our understandings of the corpus in a number of ways. Most obviously, perhaps, they are needed to show how social meanings accrue to features and how those meanings are continually reinscribed. As part of that reinscription, our three letters do more than provide metalinguistic commentary. Each participates in the iterative processes of dialect mimicry; they exemplify vocal cultures in addition to ascribing values to them. In concert with the computational analysis, therefore, they can help to explain why particular constellations of features emerge as literary conventions at particular historical moments, as well as why specific features (or groups of them) may rise and fall in popularity. These artifacts assist not only in tracking evolving language attitudes and ideologies across periods of imperial conflict and change, but also in probing the relationship among events, ideologies, and linguistic structure.

1.6 Patterns, their propagation, and their meaning

Before concluding with a chapter-by-chapter overview of the thesis, I want to briefly address an aspect of this study that may already be apparent, but one that I think is worth foregrounding. It is something that the story from the *Columbian Centinel* touches on, though the story itself is almost certainly apocryphal: the relative authenticity or accuracy of literary dialect. The evaluation of accuracy is an exceedingly significant and thorny question for historical linguists. After all, for much of language history, the only records we have are written ones. Thus, understanding spoken variation in the past requires the careful evaluation of the available evidence. But there is another way of approaching literary dialect. Rather than thinking about the partiality and artificiality of literary dialect as something that needs to be accounted for and possibly overcome, we can think of it as central to its structure and replication. The artificiality of literary dialect is emphasized by Blake (1981) in *Non-Standard Language in English Literature*, where he asserts that it is the power of literary dialect’s social signaling that is its defining characteristic.

From one point of view, then, establishing the authenticity of a text like the Toby letter would be critical. We would want to gauge its relationship to real-world speakers in order to make claims about the linguistic practices of speech communities

in colonial North America or to evaluate its significance as an aesthetic object. Is it a reductive stereotype or a carefully observed, naturalistic rendering? From the orientation of this study, however, such evaluations are irrelevant, which is not to say that that they are uninteresting or unimportant. Indeed, the data and results of this study may be of interest to scholars pursuing exactly those kinds of questions. Yet, as I hope the previous discussion has shown, what is salient to this study is not whether a text like the Toby letter is accurate. What is salient is that it exists at a time when literary dialect is emerging as a convention for racialized voices in drama and in fiction. It demonstrates that these representations are circulating particular sets of linguistic features around the Anglophone world, priming the conditions for their use upon the stage. Additionally, it instantiates demonstrably routinized evaluations of African diasporic vocal culture. Instead of its accuracy, the focus becomes literary dialect's indexicalities – the social signals it encodes and the apparatuses of its propagation. Returning to the metaphor of texts as nodes in a network of discourse, we might frame the questions thus: How do occurrences of literary dialect connect and form patterns? How can those patterns be described? And how can they be explained?

1.7 The organization of the thesis

Chapter 2 outlines past research in literary dialects. It begins by discussing the traditional emphasis in linguistics on evaluations of structural accuracy and contrasts that with the orientation of literary studies, which has tended to be more interested in the representational potential of literary dialect. The chapter goes on to describe other research that has combined these two approaches in the study of literary dialect and related forms of aesthetically performed or staged language. The chapter concludes with a discussion of corpus-assisted research that has a similar interest in representationality and of the study's position linking linguistics and literary studies in its approach to structure and social meaning.

Chapter 3 sets out the study's methodology, with a particular emphasis on the compilation and annotation of the corpus. The chapter describes how works were identified and the particular challenges that were faced in locating relevant works without biasing the data. It also explains the coding scheme, the logic of its development, and some of the complications in assigning codes to features or placing them within a taxonomy.

Chapter 4 provides an overview of quantitative results. That overview includes an introduction to some of the statistical techniques – like diversity indices and deviation of proportions – that are used in the analysis. The analysis itself begins with a discussion of feature frequency and dispersion for each superordinate category (lexical, morphosyntactic, orthographic, and phonological) in the corpus as a whole. The chapter continues with an examination of the ways in which literary dialect varies by speaker, over time, and across texts. That examination includes regression analysis in the diachronic trends for composite frequencies and diversity indices, as well as hierarchical cluster analysis for a subset of the corpus.

Chapters 5, 6, and 7 provide the results of the analysis for African diasporic, Indian, and Chinese dialogue respectively. Each chapter opens with a presentation of the overarching constituent patterns for lexical, morphosyntactic, orthographic, and phonological features, then continues with a discussion of diachronic trends. The analyses of diachronic trends in these chapters use the same computational techniques set forth in chapter 4. That is followed by an examination of resemblances using hierarchical cluster analysis, again conforming to the template established in chapter 4.

Chapter 8 concludes the thesis by considering some of its disciplinary interventions and potential implications.

Chapter 2
Literature Review

2.1 Introduction

Analysis of literary dialect has been typically approached from one of two orientations. The first assumes accuracy as its starting point and bases its evaluations on the relationship between the fictive orality of the text and the real-world speech of the community being ventriloquized. The second sees literary dialect as a symbolic resource that is contextualized by other meaning-making systems internal and external to a literary work. The first of these orientations has often resonated with linguists, and it is not difficult to understand why. Written records are the only source of information we have about historical varieties and separating the accurate from the spurious is a significant task. Thus, even when philological inquiry has not been the explicit purpose motivating an investigation into literary dialect, using accuracy as a benchmark has important analogues.

The second orientation has been more widely adopted in literary studies. Analyses of this type examine the ways in which dialect representations are implicated in themes, subjectivities, histories, and ideologies. While I am separating these orientations for the purposes of discussion, they are neither antagonistic nor incompatible. A number of studies have linked the two approaches – some arguing for linguistic analysis to incorporate literary theoretical frameworks, some arguing for literary scholarship to incorporate linguistic methodologies. Some of these have pointed to the work of sociolinguists as the bridge between disciplines, specifically work that examines the connection between style and the advertising of social identities.

This study aims to build from this small but growing body of interdisciplinary research by using the descriptive methods of sociolinguistics to facilitate an investigation of literary dialect's relationship to evolving conceptions of race and the British imperial project. In order to carry out the analysis, I conceptualize the literary dialects under study as a *representational system*. In other words, I look at lexical, morphosyntactic, and orthographic patterns used in the rendering of voices across texts and across time. The foundations for such a systemic approach come from theories of enregisterment, colonial discourse studies and theories of Orientalism, and

recent sociolinguistic research on staged linguistic performance. The concomitant identification and description of linguistic patterns are conducted using the techniques and tools of corpus linguistics.

In this chapter, I survey research in literary dialect beginning with Ives' germinal "A Theory of Literary Dialect." The survey is separated into three parts. The first frames the debate about realism and authenticity that began with Ives. The second focuses on research that has analyzed nonstandard voices from a more literary point of view, and the third on studies that have sought to bridge literary and linguistic concerns. Next, I describe the theoretical underpinnings for the systemic approach to literary dialect adopted for this study. Finally, I briefly introduce the corpus studies that inform my approach and explain how the articulation between theory and methodology situates this research relative to other disciplinary interests in literary dialect. A more thorough description of the specific methods is presented in the following chapter.

2.2 Literary Dialect

2.2.1 A question of accuracy

Ives' (1950) influential article on literary dialect lays out an important tension that has informed one, prevalent direction of research. He suggests that the creation of literary dialect results from both the keen observations of authors and from a tendency toward generalization and exaggeration:

> From the total linguistic material available, he [*sic*] selects those features that seem to be typical, to be most representative of the sort of person he is portraying. These features he generalizes so that the literary dialect is likely to be more regular in its variants than the actual speech which it represents. The character is likely to use initial [d] in every word in which an educated character would use initial [ð], in spite of the fact that it is by no means certain that the man in real life would do so. (146)

Furthermore, he argues that an author's ability to simulate real-world speech is inescapably limited by the inadequacies of orthography:

> No matter how conscientious an author is, and no matter how complete a representation of his [*sic*] character's speech he may wish to convey, he is limited in his accomplishment by the deficiencies of English spelling as a representation of English pronunciation. (148)

According to Ives' framework, the production of literary dialect is embedded in a series of conflicting impulses and constraints: a desire for verisimilitude, the tendency toward regularized patterning, and the shortcomings of standard orthography in rendering the complexities of phonology. He, thus, conceives of literary dialect as a selective and flawed representation of real voices and real speech communities;

nonetheless, it is simultaneously one that is motivated and contextualized by those real voices. A similar understanding of literary dialect is articulated by Page (1973, p. 52):

> And, though our concern is with literary conventions, it is worth saying that this [literary dialect] has a rough correspondence with the characteristics of real-life speech, since we constantly, on the evidence of the spoken language, both classify those we meet according to such loose categories as 'middle-class' and 'north-country', and, as we grow more familiar with their speech, acknowledge certain features as particular to the individual.

Such "rough correspondence," in Ives' (1950, p. 152) view, frames potential linguistic analysis: "Any literary dialect, therefore, will necessarily be a partial and somewhat artificial picture of the actual speech. It is the analyst's task to eliminate the spurious and interpret the genuine." In other words, the value of literary dialect as data resides in its ability to shed light on the linguistic practices of the speakers it presumably represents. This orientation naturally gives rise to a number of related questions: How accurate is the literary dialect? How consistent? How can we extract authentic phonological, lexical, and morphosyntactic information from data that is inherently imperfect and often messy? Answering such questions usually begins by comparing catalogues of features used in a speech community to the features used to represent speakers in a text or set of texts. In addition to the real-world practices of the community being represented, it has been further suggested that the author's own speech practices are key to understanding the relationship between written forms and their spoken counterparts (e.g., Kretzschmar, 2001). Studies approaching literary dialect in this way have used it in analyses of diverse English varieties including African American English (Barry, 2001; Burkette, 2001; Cooley, 1997; Ives, 1955; Minnick, 2001, 2007; Pederson, 1985; Tidwell, 1942), Southern American English (Carkeet, 1979; Ellis, 1994; Pederson, 1967), Appalachian English (Nickell, 1984), Illinois English (Fenno, 1983), East Anglian English (Poussa, 1999), Northern British English (Underwood, 1970), Southern British English (Melchers, 2010), Tyneside English (Beal, 2000), Scouse (Honeybone & Watson, 2013), Jamaican Creole (Schneider & Wagner, 2006), and Hiberno-English (Sullivan, 1980).

While inferring speech practices from written representations is complicated by issues of authorship, patterning, manipulation, and representativeness (Montgomery, 1999), literary dialect presents an attractive source of data, particularly in the absence of preserved historical forms of vernacular or nonstandard varieties (Wolfram, 2000). Moreover, as Minnick (2007) points out, the sum of our knowledge

regarding older forms, historical variation, and extinct varieties is based on writing. That knowledge partly comes from orthographic analysis that is analogous to the kinds of work described above. Thus, the history of philological inquiry provides methodological precedent and general frameworks for linguistic investigations into literary dialect. Nonetheless, critiques have been offered up from a variety of perspectives. First, there is the broad question of their reliability. Schneider's work on Englishes spoken by African diasporic communities in the Americas encapsulates this debate. On the one hand, he argues that transcribed ex-slave narratives – with all of the attendant problems involved in the processes of transcription (see, e.g., Preston, 1985) – are "the best evidence of an earlier stage of Black English that we have and are likely ever to get, and is clearly superior to literary dialect" (Schneider, 1993, p. 218). On the other, in his more recent study of the Jamaican Creole represented in Michael Thelwell's *The Harder They Come*, he concludes "that literary dialect does not necessarily have to be inaccurate or even invalid as linguistic data, which supports the view that literary representations of earlier stages of languages need not be ignored as sources of real-time data of language change" (Schneider & Wagner, 2006, p. 86). These positions are not necessarily contradictory. In the first, he articulates a widely shared skepticism of literary dialect, and a preference for other (though by no means unproblematic) sources of data. In the second, he argues for its potential utility, but with an important caveat: "[i]f the native-speaker status of the person who records the dialect and the breadth and quality of his/her intuitions can be proved" (Schneider & Wagner, 2006, p. 86). Together his statements express the field's general ambivalence toward literary dialect: it is an at once attractive but fraught source of data, useful only in as much as its accuracy can be demonstrated.

2.2.2 Literary dialect as representational resource

An alternative to historical linguistics' prevailing concern with accuracy comes from literary studies. From this orientation, dialect is a generative, semiotic resource – one that can reinforce the themes, subjectivities, and ideologies of an individual work or of a body of texts circulating within a time period or a particular culture. Literary dialect, therefore, can be understood as implicated in a variety of socially constituted phenomena: stereotype, humor, and satire, as well as ideas of self, community, region, nation, and other. The analysis of literary dialects can provide insights into the identities, individual or collective, that those dialects map, as well as

the ideological positions, the cultural histories, and the political economies that contextualize their production. Much work has been done, for example, in exploring the proliferation of "dialect literature" in nineteenth century America, and its connection to changing demographics, ideas of nationhood, and the power relations among groups variously claiming hegemonic dominance and asserting their right to difference (e.g., Jones, 1999; Kersten, 2000; Nettels, 1988; Pratt, 2002; Strand, 2009). Similarly, the political and economic tensions surrounding nineteenth century linguistic debates serve as a jumping off point for North's (1994) examination of the racial masquerades of transatlantic modernism during the twentieth century and how conventional, boundary-marking uses of dialect mimicry, in his words, became remapped "across national boundaries and also across boundaries between the practical and the decorative, the concrete and the ephemeral, motivated and conventional, dialect and standard" (194).

Some of the conflicts at play in a rapidly changing Anglophone world that are explored by North, Jones, and others – between an elusive "standard" and diverse vernaculars, between ideals for a polity that privileges unity and one that embraces plurality, between national and local identities – are also central to Blank's (1996) analysis of literary dialect in a very different context: Renaissance England (and, to a lesser extent continental Europe). Although the Renaissance period predates many of the institutional and popular apparatuses that uphold standard language ideologies in later centuries (government-run schools, widely available dictionaries, self-help manuals, etc.), she argues that the idea of the Renaissance era as being distinguished by laissez-faire attitudes toward language variation and change is an oversimplification. She suggests, rather, that the "'linguistic enthusiasm' of authors such as Shakespeare, Spenser, and Jonson needs to be re-examined as an expression of an age engaged in a struggle for possession of the vernacular, a struggle in which linguistic authority was just as much at issue as linguistic freedom" (6). Further, she views the political significance of such struggles as largely implicated in individual acts of authorial labor. The "politics of language," she argues, follow from "the effort of each individual writer to discriminate among versions of the language and to authorize preferred forms, to draw (and then, at times, deliberately to transgress) the borders that separate one dialect of English from another" (6). Thus, like North, she frames the choices of codes, their representations, and their juxtapositions as a process

of negotiating borders, a process that can reinforce existing political, social, and economic hierarchies, but one that may also subvert them.

Ideas of borders are similarly important in regionally focused, rather than period focused, studies of literary dialect. In emphasizing place, these studies connect linguistic boundaries to socio-spatial geographies and to concomitant performances of identities. In her analysis of Lancashire literary dialect during the Victorian period, for example, Hakala (2010, p. 407) explores writers' negotiations of regional and class identities in an industrial present through imaginings of "authentic" speakers in a rural past. Marshall (2011) similarly examines the work of Frederic Moorman, an early twentieth century academic, and his project of propagating an idealized "rustic" and "peasant" Yorkshire identity that is linked to local traditions of dialect poetry and miracle plays. Central to both Hakala's and Marshall's accounts are interconnected imaginings and mappings: past onto present, language onto geography, and authenticity onto class, all of which are bound up in historical antagonisms between the North and South in England. Notions of authenticity are also central to Gupta's (2000) study of Singapore English and its role in novels by Ming Cher and Rex Shelley. Gupta (2000, p. 163) argues that Shelly draws from a wide range of linguistic features representative of the repertoires available to Singapore English speakers in order to index local sociocultural divisions and subjectivities. Cher, by contrast, creates a stylized version that makes limited use of the lexical and morphosyntactic inventory of Singapore English. Despite these discrepancies, both authors, Gupta notes, were marketed as "authentic" voices of Singapore. This contradiction results from their intended audiences. Shelly's novel was produced for local consumption, while Cher's was produced for international consumption. Shelly's language, therefore, marks ethnic and class boundaries within Singapore itself. Cher's language, however, exoticizes Singapore for a Western audience, figuring a border between them and a "mysterious" East. Although Gupta focuses on a much more contemporary linguistic context, her study – like the others discussed in this section whether they emphasize time period or geography – demonstrates the indexical power and flexibility of literary dialect. Even what are presented as "authentic" representations of the same language variety can be constituted in different ways for different purposes. The question from a literary orientation is less the degree to which those representations are accurately modeled on real-world speech practices but more how they are used in the imagining of social, political, and economic relationships.

2.2.3 A combined approach

While linguistic and literary orientations may have differing emphases, they are by no means incompatible, and a number of studies have combined them to varying degrees. In her investigation of literary representations of African American English, for example, Minnick (2007, p. 33) proposes a connection between the language of speakers in the world and the ventriloquized language of fiction. She contends that an analysis of accuracy "can lead to insights about characterization and also about attitudes toward speech and speakers held by authors, and by various characters toward one another." Gupta's study discussed above is grounded in related principles. The lexicon and structure of non-literary Singapore English serve as the basis for her literary assessments of the authors' stylistic choices. Chinese Immigrant's Pidgin English occupies a similar role in Li's (2004) analysis of Maxine Hong Kingston's *Tripmaster Monkey: His Fake Book*. Kingston's manipulation of stereotypical features of Chinese Pidgin English phonology and morphosyntax, according to Li (2004, p. 280), advertise characters' humor, wit, and subversions of dominant ideologies. Li, for example, juxtaposes the language of Dr. Woo, a Chinatown doctor (1), with Witman Ah Sing's imitation of the doctor's speech (2):

(1) You hurt? You tired? Ah, tuckered out? Where you ache? This medicine for you. Ease you sprain, ease you pain. What you wish? You earn enough prosperity? Rub over here. Tired be gone. Hurt no more. Guarantee! Also protect against accidental bodily harm. And the law. Smell. Breathe in deep. Free whiff. Drop three drops – four too muchee, I warn you – into you lady's goblet, and she be you own lady. (Kingston, 1987, p. 14)

(2) Show the bok gwai that Chinese-Ah-mei-li-cans are human jess likee anybody elsoo, dancing, dressed civilized, telling jokes, getting boffo laffs. We got rhythm. We got humor. (Kingston, 1987, p. 15)

While many of the features in (1) (e.g., zero copula, null clausal subjects, and null auxiliaries in interrogatives) in Li's words (2004, p. 280) "ground Dr. Woo's speech in empirical reality," the *-ee* affix in *muchee* is associated with caricature. Li interprets the inclusion of this feature as a wry signal of the doctor's Chinese identity operating within a stereotypical situation. Witman's speech in (2) also realizes features that index caricatures of Chineseness like the *l*-for-*r* substitution in *Ah-mie-li-cans*. However, his speech additionally realizes features (e.g., *oo*-for-*o* substitution in *elsoo* and the phonologically unmotivated respelling of *laughs* as *laffs*) that appear to be satirical hyperbole. Much like Dr. Woo's, Witman's voice encodes resistance and sly subversion. The real, the stereotypical, and the satirical exist in a tension that

highlights the struggles of Chinese Americans in negotiating their linguistic and cultural identities.

Linguistic caricature and standard language ideology also figure prominently in Wales' (2006) analyses of Northern English varieties. She stakes out a position similar to Minnick's in arguing that literary sources are often the only available historical record documenting language variation and that writers were most certainly aware of high-frequency features and shibboleths that marked speech communities. However, in contrast to scholars like Minnick, Gupta, and Li, Wales does not use accuracy as a benchmark. She focuses her attention on evolving understandings of Northern English as both a lived and a literary language without measuring the latter against the former. While literary dialects are only a part of the expansive data she marshals, Wales demonstrates the crucial role they have played in the development and maintenance of symbolic and ideological boundaries that separate North from South. She shows how the voices of characters like Bob Cranky – a fictional nineteenth century Geordie depicted in the excerpt below from "Bob Cranky's Adieu" (Marshall, 1812) – become archetypes:

(3) Fareweel, fareweel, ma comely pet!
Aw's forc'd three weeks to leave thee;
Aw's doon for parm'ent duty set,
O dinna let it grieve thee!
Ma hinny! wipe them een sae breet,
That mine wi' love did dazzle;
When thy heart's sad, can mine be leet?
Come, ho'way get a jill o' beer,
Thy heart to cheer:
An' when thou sees me mairch away,
Whiles in, whiles out
O' step, nae doot,
"Bob Cranky's gane," thou'lt sobbing say,
"A sowgering to Newcassel!!"

These archetypal voices serve in the ratification of regional mythologies (Wales, 2006, p. 133). That ratification occurs not only in the North, as part of in-group identification. It also occurs in the South, where it figures in the ideological delineation between court and shire, between "metropolitan superiority" and "provincial inferiority" (Wales, 2006, p. 88). Thus, in indexing boundaries between North and South, literary dialect can reinforce standard language ideologies or challenge them. It plays a role in evolving conceptions of a "standard" English by offering a countervailing model of a "nonstandard." But it also serves to voice

resistance to the political, economic, and linguistic dominance of London's metropole.

Studies such as those by Wales, Li, Gupta, and Minnick are representative of approaches negotiating questions of accuracy and representationality from a point of view grounded primarily in linguistics. Others, however, have bridged linguistic and literary orientations with literary studies as their foundation. In their examination of literary dialect during the late Romantic period, Hodson and Broadhead (2013), for example, include in their analysis some discussion of the phonological and morphosyntactic features of Scottish, Irish, and Welsh English. However, that discussion is used to instantiate diachronic trends rather than to evaluate the authenticity of the dialect represented in particular texts. The authors defend their approach partly by citing the work of sociolinguists like Coupland (2003, 2007), Eckert (2003), and Bucholtz (2003), work that foregrounds the relationships between language style and identity. They argue that the "stylistic turn" in sociolinguistics and its concomitant interest in performance and perception invites new ways of conceptualizing the relationship between literary dialect and its participation "in the complex debates about correctness, education, artificiality, and linguistic virtue that were circulating at the time" (Hodson & Broadhead, 2013, pp. 316-317). These sociolinguistic frameworks are positioned as ones that shift the focus away from authenticity and that can, therefore, more fruitfully complement the interests of literary studies.

Hodson and Broadhead are not the only literary scholars to use the work of sociolinguists as a link between disciplines. Hakala, too, cites Coupland (2003) in arguing for "authenticity" as a contingent and contextualized identity that was performed by Lancashire dialect writers during the Victorian era. In his study of nineteenth century American literary dialect, Leigh (2011) similarly seeks to problematize the concept of "authenticity" by drawing on sociolinguistic research by Bucholtz (2000), Jaffe (2000), and Bailey *et al*. (2005). He further advances linguistic-literary connections in borrowing the term "literary sociolinguistics" from Mair (1992) as a frame for his analysis.

What Mair proposes is an interdisciplinary enterprise that "not only provides further, more empirical or objective corroboration for what literary scholars have known all along but that it deepens traditional understanding where, as in the question of 'style' in fictional prose, it is clearly deficient" (1992, pp. 111-112). As a model for

the analysis of literary dialect, "literary sociolinguistics" aims to marry the empirical methods of sociolinguistics with literary theories of the novel. The purpose of this interdisciplinarity is less to problematize notions of authenticity – as it is for scholars like Hakala and Hodson and Broadhead. Rather, it is to ground literary evaluations of nonstandard language in descriptions that are more robust than the traditional preoccupation with orthographic representations of phonology. For his part, Leigh (2011, p. 122) conceives of "literary sociolinguistics" as encompassing both a methodology that juxtaposes linguistic structure with literary themes, as well as a theoretical orientation that eschews what he calls "the wild goose chase of linguistic authenticity."

In a similar pairing of the linguistic and the literary, Ferguson (1998, p. 3) proposes the term "ficto-linguistics," which she compares to sociolinguistics:

> I will explore the *narrative consequences* of dialect use in fiction by looking at what be called the ficto-linguistics as opposed to the socio-linguistics of dialect in the novel. By ficto-linguistics I mean the systems of language that appear in novels and *both* deviate from accepted or expected socio-linguistic patterns *and* indicate identifiable patterns congruent to other aspects of the fictional world.

Ferguson's orientation is more literary than Mair's, but is very much in keeping with Leigh's adaptation of Mair's model. Like Mair, Ferguson seeks a marriage of linguistic methodology and literary analysis. However, where Mair casts an eye outward, on issues of "faithfulness" (though he is clear that such issues should make up only a part of the analysis), Ferguson's focus is inward, on literary dialect's function in "the fictional world." In arguing for the value of Ferguson's approach, Hodson (2014, p. 14) suggests that such an inward orientation opens the door for a different kind of linguistic analysis:

> The term 'ficto-linguistics' is valuable because it provides a way of talking about the patterns of language variety we find within fictional texts, and using terms and concepts borrowed from linguistics in order to so, while making it clear that language varieties do not function in the same way as language varieties in the real world. The term thus moves us beyond analysing language varieties in literary texts in order to rate them in terms of their real-world accuracy or consistency, which is what sometimes happens with linguists analyse literary texts, and instead enables us to see that they form an integral part of the fictional world within which they appear.

This study adopts an approach to literary dialect similar to Mair's and to Ferguson's in its combining of linguistic methodologies and literary interpretative frameworks. Additionally, it follows Ferguson's practice of analyzing literary dialect in terms of its representational functions, as opposed to its relationship to authentic speech. The study's focus, however, is not restricted to the fictional worlds of the source works. It aims to locate fictional representations within larger discoursal

patterns, following practices like those used to analyze imaginings of imperial subjects in studies of colonial discourse and the accrual of social values to language varieties in studies of enregisterment. Thus, it holds a more expansive notion of a representational system than a strictly novelistic one.

Some of what informs that notion of representationality is the subject of the following section, but before leaving the accuracy debate behind I briefly want to clarify an important point. While this study does not link fictive and non-fictive orality, I do not share the position – forwarded by Leigh (2011), Birnbaum (1991), and others – that questions of accuracy are chimerical. As has been repeatedly noted in this chapter, accuracy is of obvious importance to philological inquiry. Even in the pursuit of literary questions, studies like Li's (2004) and Gupta's (2000) demonstrate the potential value of measuring an author's manipulation of a code against its real-world counterpart. That said, neither do I hold with Mair (1992, p. 105) when he claims that "the indispensable first step in any analysis of a literary dialect should be the systematic comparison with the real thing in order to establish points of contact and points of deviation between life and art." Nor do I wholly agree with Minnick (2007, p. 33) when she similarly asserts, "To dismiss accuracy entirely as irrelevant ignores the artistic meanings generated precisely because of the author's choices about how to represent the dialect as well as about whose dialect to represent." Assessing accuracy is not a necessary step in the analysis of artistic, ideological, social, or political meanings, as is evident from the work of scholars like Jones (1999) and Leigh (2011). Rather than taking an absolutist position, I would argue, instead, that the decision to consider accuracy as part of an interpretative framework is entirely dependent on the purposes of the analysis. Because this study explores historical patterns of representation and the ways in which those patterns circulate, accuracy is not a material concern. Even so, the study's results may well be of interest researchers pursuing those very questions.

2.3 Literary dialect as a representational system

2.3.1 Orientalism and colonial discourse studies

The systemic approach to literary dialect adopted for this study has its theoretical roots in colonial discourse studies, theories of enregisterment, and recent sociolinguistic research on staged performance. The approach aims to describe

lexical, morphosyntactic, and orthographic patterns, to juxtapose those patterns against other novelistic elements, and to situate the representations of vocal culture in the shifting ideologies and political economies of empire. The emphasis on discoursal patterns and their political and ideological implications owes much to Foucault (1972, 1977) and to Said's (1994) specific application of Foucault in his foundational work on Orientalism. In setting out his theory of Orientalism, Said argues not only that Middle Eastern peoples and culture are essentialized though repeated renderings in the Western imagination, but also that those imaginings have upheld and continue to uphold political, economic, and military systems of dominance and exploitation. Part of his project, therefore, is to mine the imperial archive and to expose representational patterns. He explains:

> My analysis of the Orientalist text therefore places emphasis on the evidence, which is by no means invisible, for such representations *as representations*, not as "natural" depictions of the Orient. This evidence is found just as prominently in the so-called truthful text (histories, philological analyses, political treatises) as in the avowedly artistic (i.e., openly imaginative) text. The things to look at are style, figures of speech, setting, narrative devices, historical and social circumstances, not the correctness of the representation nor its fidelity to some great original. The exteriority of the representation is always governed by some version of the truism that if the Orient could represent itself, it would; since it cannot, the representation does the job, for the West, and *faute de mieux*, for the poor Orient. "Sie konnen sich nicht vertreten, sie miissen vertreten werden," as Marx wrote in *The Eighteenth Brumaire of Louis Bonaparte*.[3] (Said, 1994, p. 21)

Although Said's conception of representation and of discourse extends to visual images, the preponderance of his evidence comes from written texts. Despite this emphasis on language and its semiotic power, Bolton and Hutton (2000) argue that linguistics as a field has been slow to take up the mantle of Orientalism, particularly when compared to literary studies. Their critique comes in an introduction to a special issue of the journal *Interventions*, in which they set out to model linguistic engagement with Orientalism's debates. Hutton's (2000) article, for example, examines the construal of racial and linguistic identities in the nineteenth century writings of Sir Henry Sumner Maine and Edward Augustus Freeman. The writers both draw from the then emerging science of comparative philology in expressing their beliefs about the fictive kinship created though "blood and speech" – beliefs, according to Hutton, that reveal deep anxieties about the future of the British Empire, particularly in India.

Hutton's method of linking linguistics and Orientalism is to explore the discipline's historical role in representing the languages of subaltern people in ways

[3] "They cannot represent themselves; they must be represented."

that rationalized European dominance. Since the publication of *Orientalism and Linguistics*, a number of scholars have employed an approach similar to Hutton's (see, e.g., Errington, 2008), and that approach has relevance here. Circulating in greater Britain – sometimes concurrent with, sometimes predating the literary dialect of novels and plays – are the descriptions of voices from other sources including philological ones. Those sources provide important context, for, as Jones (1999) and Birnbaum (1991) point out, the contemporaneous emergence of the dialect novel and comparative philology in the nineteenth century was not coincidental. They drew from each other's enthusiasms and ideologies. That said, this study analyzes not just metalanguage related to vocal culture, but also the structure of the voices themselves. In other words, its scope is wider in seeking to explore relationships between lexical, morphosyntactic, and orthographic patterns and the social and ideological meanings that those patterns encode.

2.3.2 Enregisterment and staged linguistic performance

Connecting these multiple levels of representationality is facilitated through Agha's (2003, 2007) theory of enregisterment (§1.5). Enregisterment is the process through which social meanings accrue to repertoires of differentiable linguistic forms. Johnstone (2009), for example, shows how the practice of putting Pittsburghese onto t-shirts standardizes local speech and imbues that speech with particular local values. As Johnstone's study attests, the processes of enregisterment are carried out in a variety of communicative contexts, and a few scholars have examined enregisterment in relation to literary dialect. Honeybone and Watson (2013) posit that the social salience of Scouse spellings (e.g., *worra* for *what a*) in local, humorous dialect literature is linked to the enregisterment of Liverpool English. In another study of British regional dialect, Cooper (2013) analyzes nineteenth century Yorkshire English – comparing texts written from local and non-local audiences – and concludes that overlapping features were likely enregistered for a contemporaneous audience, often indexing an archetypal Yorkshire "character."

The work examining literary dialect and enregisterment is complemented by studies exploring the role of enregisterment in staged linguistic performance. Bell and Gibson (2011, p. 557) define staged linguistic performance as follows:

> Staged performance is the overt, scheduled identification and elevation (usually literally) of one or more people to perform, typically on a stage, or in a stage-like area such as the space in front of a camera or microphone. It normally involves a clearly visible and instantiated

> distinction between performer and audience. Prototypically, staged performance occurs through genres such as a play, concert or religious service, and in venues dedicated to such presentations – a theatre, concert hall or place of worship.

Staged performances are distinguished from everyday performances, which are spontaneous moments of conversational stylizing. Staged performances, by contrast, are sustained and planned, and thus emphasize the metalinguistic and poetic functions of language. In doing so, "[t]here is also a heightened awareness of the existing repertoire of cultural texts, and value is placed upon their skillful recontextualization" (Bell & Gibson, 2011, p. 558).

The salience of the metalinguistic and poetic functions of language in staged performance is made clear in another of Johnstone's studies of Pittsburghese. In the above-mentioned study, she examines the enregistering effects of commodification. In this one, she analyzes the staged comedic performances of Pittsburghese by two radio DJs (Johnstone, 2011). She finds that the enregistering effects of the staged performances oppose those of the t-shirts. Rather than narrowing and standardizing Pittsburghese through commodification, the performances "tend to open up new possibilities for the enregisterment of locally-hearable linguistic forms" (Johnstone, 2011, p. 676). This opening up, Johnstone argues, follows from staged performance's inherent reflexivity. As she says of staged performance, "it is about itself, sometimes as much as it is about what it denotes" (Johnstone, 2011, p. 676). Literary dialect is similarly reflexive. It, too, calls attention to itself, and invites the reader to place the imagined voice into interpretative schema beyond the purely denotational. Literary dialect shares other features with staged linguistic performance, as well, making enregisterment's role in staged performance particularly salient to this study.

The analogues between the literary dialect considered for this study and staged performance are perhaps most clear in Bucholtz and Lopez's (2011) analysis of white enactments of blackness in Hollywood films. First, based on a collection of 59 films, the authors describe a set of phonological, grammatical, and lexical features that have become emblematic of Mock African American English.[4] These include well-documented features of African American English like the zero copula (see, e.g., Green, 2002), as well as features like *-ass* as an intensifier that are representative of Hip Hop Nation Language (see, e.g., Alim, 2006). Next, the authors examine the

[4] Mock varieties are parodic imitations that are made up of a limited repertoire of iconized linguistic variables. Their performance often signals cross-racial or cross-ethnic othering, but can also subvert existing stereotypes (see, e.g., Chun, 2004; Hill, 1995).

realizations and indexicalities of those features in the performances of Steve Martin in *Bringing Down the House* and Warren Beatty in *Bulworth*. They describe the general semiotic terrain of the films as follows:

> Both *Bulworth* and *Bringing Down the House* open by highlighting the European American male protagonist's remoteness from African American language and culture. Just as hierarchical social differences are semiotically established between the white and black characters based on class and education, so too are hierarchical differences established between standard English and AAE. Neither film includes black characters who primarily speak standard English, while white characters – with the crucial exception of the European American male characters who engage in crossing into Mock AAE – consistently speak the standard. (Bucholtz & Lopez, 2011, p. 691)

The appropriation and mimicry of African American English affords the white protagonists access to the covert prestige that the variety indexes, and it is a catalyst for their self-actualization. However, the humor of their crossing is predicated on Martin's and Beatty's exaggerated whiteness, which emphasizes a divide between European American and African American vocal culture. The linguistic minstrelsy of the films, therefore, reinscribes stereotypes and "the deeply problematic dichotomy between rational middle-class whiteness and physical working-class blackness" (Bucholtz & Lopez, 2011, p. 702).

Staged linguistic performance as it is described and theorized in the above examples is oral and embodied. Thus, it includes the significations of posture, gesture, dress, adornment, etc. In that regard, it is clearly distinct from the purely textual voicings of literary dialect. That said, the cinematic linguistic minstrelsy analyzed by Bucholtz and Lopez and the literary dialects analyzed in this study share a variety of important characteristics: their strategic mimicry, their invocation of stereotype, as well as their reliance on and recirculation of historically marked repertoires of linguistic features. In light of these commonalities, we might view literary dialect as a kind of textually staged linguistic performance. As such, although it, like embodied staged performance, is outside the conventional purview of sociolinguistic study, it has a similar role in the circulation of linguistic forms and attitudes. Thus, its analysis sheds light on important sociolinguistic processes. Bell and Gibson (2011, p. 558) put it this way, "Performed language provides a window on the world of the creative and the self-conscious, the kind of language excluded from sociolinguistic work which targets 'natural, unselfconscious speech'."

2.4 The articulation between theory and method

Of the wide variety of studies on literary dialect and its analogues that have been discussed, only a few make use of corpus methods. Corpus methods are a fundamental component of this study, and although methodological specifics are reserved for the next chapter, I want to briefly discuss work that has integrated corpus linguistics and some of the theoretical orientations relevant to this study. In establishing the articulation between theory and method, the discussion provides some prefatory context for the more detailed chapter that follows. More importantly, it sheds light on how this study purports to explore links between structure and social meaning. In doing so, these discussions situate the study within the context of the various disciplinary and theoretical frameworks that have been presented in this chapter.

One specific area of social meaning that this study is keen to explore is the relationship between literary dialect and the imperial imaginary. As I noted previously, the latter is a central concern of colonial discourse studies. I also noted that there has been work like Hutton's (2000), which has sought to bridge linguistics and colonial discourse studies. Efforts to integrate corpus analysis with colonial discourse studies have been rarer, but some like Koteyko (2006, p. 145) have suggested that corpus methods can be leveraged in the study of historical social semiotics, the Foucauldian approach to discourse that Said draws upon in his germinal work on Orientalism:

> Nevertheless, there are common points which allow merging linguistic and "archaeological" methods of research in the corpus-driven approach to the study of discourse: 1) the view of language as a social construct 2) the emphasis on historical and cultural aspects of meaning production in discourse. From this perspective, the corpus-driven approach to discourse would be focused not on how meanings are constructed between sentences, which is characteristic of the abovementioned approach to discourse analysis in Applied Linguistics, but rather on how meanings come to be articulated at particular moments in history.

Though Koteyko does not mention colonial discourse studies explicitly, her emphasis on the historical articulation of meanings significantly overlaps with Said's interest in mapping patterns in Orientalist discourse. When Said (1994, p. 203) announces the need to understand "the Orient" as linguistically constituted because "the Orient was a word which later accrued to it a wide field of meanings, associations, and connotations, and that these did not necessarily refer to the real Orient but to the field surrounding the word," it is not difficult to imagine how computational methods might be mobilized in pursuit of such a project, just as Koteyko proposes.

While colonial discourse analysis and corpus methodologies may have been infrequently married, an approach with similar attention to discursive and ideological systems and those systems' role in power relations has been highly productive in applying corpus methods: Critical Discourse Analysis (CDA). CDA, in the view of Fairclough, Mulderrig, and Wodak (2011, p. 357), constitutes not a discipline but a "problem-oriented interdisciplinary research movement." It is united by an understanding of discourse as constitutive (i.e., both shaped by and shaping the contexts of its production) and an interest in the role of language in upholding and perpetuating power relationships. In arguing for the potential of corpus methods to support the goals of CDA, Mautner (2001, p. 123) forwards the following rationale:

- Corpus linguistics allows critical discourse analysts to work with much larger data volumes than they can when using purely manual techniques.
- In enabling critical discourse analysts to significantly broaden their empirical base, corpus linguistics can help reduce researcher bias, thus coping with a problem to which CDA is hardly more prone than other social sciences but for which it has come in for harsh and persistent criticism. (Widdowson, 1995, 2004)
- Corpus linguistics software offers both quantitative and qualitative perspectives on textual data, computing frequencies and measures of statistical significance, as well as presenting data extracts in such a way that the researcher can assess individual occurrences of search words, qualitatively examine their collocational environments, describe salient semantic patterns and identify discourse functions.

The promise that Mautner propounds has, in fact, been borne out in a wide variety of corpus-based studies (e.g., Baker et al., 2008; de Cillia, Reisigl, & Wodak, 1999; Mautner, 2005; Mulderrig, 2011, 2012; Wodak, de Cillia, Reisigl, & Liebhart, 2009). As the overlapping interests of colonial discourse studies and CDA would suggest, this study's theoretical concerns are very much in line with corpus-assisted models of discourse studies. Its application of corpus methodologies, however, is somewhat different. And it is with these methodologies that I want to begin situating this study within the context of the various disciplinary, theoretical, and methodological frameworks that have been presented in this chapter.

Typically, corpus-assisted discourse studies marshal the computational power of corpus techniques in order to identify patterns in the linguistic constituting of people and ideas – the kinds of patterns that would otherwise be assembled though human recognition and labor. The techniques employed in detecting and sorting such patterns might include collocations, keywords, or concordances. Caldas-Coulthard and Moon (2010), for example, examine the adjectival premodification of the nouns *man*, *woman*, *girl* and *boy* in order to explore the ways in which gender is constructed in a corpus of tabloids and broadsheets.

Studies like Caldas-Coulthard and Moon's require corpora designed to expose semantic relationships and variations. The corpus that is used in this study, by contrast, is designed to explain structural relationships and variations. As is described in detail in the following chapter, this study's corpus is comprised of dialogue that has been extracted from novels and plays and then coded for a variety of linguistic features. The computational focus, then, is placed on measuring those features along various dimensions and tracking changes over time. In this way, the study shares a lineage with more linguistically minded, corpus-based research of literary dialect like Burkette's (2001) study of *Uncle Tom's Cabin*, Minnick's (2007) study of African American dialogue, and Tamasi's (2001) study of *The Adventures of Tom Sawyer* and *The Adventures of Huckleberry Finn*, as well as with corpus-based research of fiction and drama that engages questions related to dialogue, though not necessarily literary dialect specifically (e.g., Biber & Burges, 2000; Culpeper, 2009; De Haan, 1996; Mahlberg, 2013; Murphy, 2015; Oostdijk, 1990).

At the same time, this study aims to examine the social values that accrue to features and to voices. In that respect, its interests overlap with those of literary scholars like Jones (1999) and North (1994) who have analyzed some of the ideological implications of dialect representation in the late nineteenth and twentieth centuries. It further has specific interest in the relationship between the representations of vocal cultures in fiction and evolving imaginings of empire. This puts the study in conversation with the work of scholars like Chakravarty (2005), Forman (2013), Netchman (2010), Nussbaum (2004), and Yang (2011), for whom questions of identity and the imperial imaginary are central, but whose research only tangentially engages literary dialect, if at all.

In light of these latter interests, it would be easy to conceive of using a corpus culled from the imperial archive and following corpus-assisted models of discourse analysis as a method of inquiry – emulating Said's efforts at exposing the associations and connotations of *orient* and *oriental* much like Caldas-Coulthard and Moon unpack the associations of *woman* and *man*, for example. However, as I have noted, the corpus that was built for this study is designed to maximize its potential for explaining structure, not meaning. For that reason, the study adopts an alternative approach. While the analysis of literary dialect structure is quantitative, the analysis of the ideological implications of those structures is qualitative. Thus, the study attempts to bridge both linguistic and literary methodologies and epistemic commitments,

similar to what Mair (1992) proposes. Moreover, it uses the results of statistical analyses to focus the qualitative readings of specific works, time periods, topics, or themes. The blending of the quantitative and the qualitative is common and, in the opinion of McEnery and Hardie (2012), best practice in corpus linguistics. The interface between quantitative and qualitative analysis in this study is somewhat unconventional, however. A common approach would be to measure a set of features and then examine their concordances – a corpus-internal set of processes. In this study, the statistical patterns motivate qualitative investigations not only into the source works, but also into archival sources more broadly – a corpus-internal process that drives a corpus-external one. Thus, some of the more immediately recognizable instruments of corpus research like concordances and collocations are not part of this study.

One might argue – and rightly so – that there is an obvious candidate for a second corpus that could be used to generate these types of outputs: the source works. In fact, as this project was developed, drafted, and revised, I experimented with using the source works precisely in this way: examining collocations around words like *accent* and *dialect*, using concordances to sort characterizations of speakers, even looking at semantic patterning in the descriptions of the social spaces that those speakers are imagined to inhabit. The results were interesting, I think, but ultimately stretched the study's boundaries too far. Consequently, they have been set aside as potential avenues for further inquiry. What is retained is still a mix of computational and discourse analysis, but the latter is "corpus-assisted" not in the sense that it is facilitated through the use of a tool like a concordancer. Rather, it is "corpus-assisted" in that the archival work is guided by quantitative descriptions of literary dialect structure.

2.5 Conclusion

This study seeks to unite a diverse range of theoretical and methodological threads. Most obviously, it aims to build on the rich body of scholarship investigating literary dialect that was pioneered by Ives. This study is somewhat unusual, however, in how it is positioned relative to linguistics and literary studies. On the one hand, the analysis aims to describe the structures of literary dialect, variations among structures, and the ways in which those structures change over time. These are typical of linguistic interests in literary dialect. The study also aims to triangulate those

descriptions with archival evidence related to language attitudes and with political, economic, and cultural events. While the latter approach is, of course, central to linguistically oriented research like social histories of language (e.g., Bailey, 1991; Errington, 2008; Wales, 2006), it is more common to literary studies when literary dialect is the primary object of inquiry. Thus, the study engages with the literatures from a number of disciplines and sub-disciplines, including, but not limited to, sociolinguistics and colonial discourse studies.

Critical to this mixed approach is the linking of the quantitative and qualitative analyses. Like the joining of the linguistic and the literary, such linking in this study is less unusual in its conception than its operationalization. The quantitative analysis uses computational methods to build models of structure. Those models point to trends and inflection points, alignments and discontinuities, which, in turn, focus discussions of the source works and direct archival research. Woven together, these data expose the relationships among changing literary dialect practices, the enregisterment of linguistic features, and a social landscape that is evolving in response to the developments of empire. In short, they shed light on the dynamics of a representational system.

Chapter 3
Research Design and Methods

3.1 Introduction

The corpus that is this study's foundation is the Voicing Imperial Subjects in British Literature (VISiBL) corpus. The methods that were used to build and parse it can be divided into four primary areas or stages: data collection, preparation, coding, and analysis. Data collection involved using digital archives to locate fictional works that contain an African diasporic, Indian, or Chinese character whose dialogue is rendered in literary dialect. That process required a determination of both what counts as literary dialect, as well as what counts as "British." Data preparation consisted of formatting and assigning metadata (e.g., dates and authors) to files. Data coding included the development of a taxonomy for identifying literary dialect features and protocols for assigning those codes. Finally, data analysis entailed the application of appropriate statistical methods for explicating quantitative patterns in the data.

The statistical methods are discussed at length in the following chapter. This chapter, therefore, focuses on the other three areas, each of which presented distinct challenges. Data collection necessitated procedures for finding requisite works that did not bias the data. Data preparation presented the fewest complications, though the assigning of metadata proved to be unexpectedly knotty for a few source works. By contrast, data coding raised a number of complex questions including: How should features be organized into a coherent taxonomy? When should certain codes be applied or withheld? What codes should be applied in particular environments? And how many codes should be applied?

The protocols for addressing such questions, which are explained in the latter half of this chapter, adhere to a few principles. In organizing codes, there is a preference for grouping systems (verbs, modifiers, etc.) together, as much as such groupings are rational. In assigning codes, there is a preference for conservatism. This means that one objective of the coding is to describe a feature with as few codes as possible (which sometimes means no code at all, as I will explain). It also means, for phonological and orthographic features, that there is an effort to minimize guessing at the correspondences between respellings and sounds they are intended to signal.

These procedures are partly a response to the realities of a corpus that includes authors with lives spanning more than a century-and-a-half (the author who was born first, David Garrick was born in 1717 and the latest-born, Sax Rohmer, in 1883), as well as authors from a variety of backgrounds and regions. "British" in this study is defined as a participant in the "composite Britannic culture" (a term that is defined more fully in the discussion of the data collection). The rather capacious boundaries that the phrase implies result in the inclusion of works from writers with roots not just in England, but also in Ireland and Scotland, in addition to those with itinerant upbringings. Some authors spent part of their lives in England and some of it elsewhere – in the United States or out in the empire. The diversity of the 116 different authors whose works populate the corpus creates the possibility, if not the likelihood, that the same respelling may not have the same phonological signification across multiple works.

A more important consideration than those conditions, however, is the study's goals. This study is designed to investigate the practices of representation, not their authenticity or their accuracy. Thus, inferring phonological salience is less central than it might be otherwise. In fact, the coding for phonological features is organized by spelling rather than phonetics. That does not imply that phonology is irrelevant; neither does it imply that phonological conjectures are eschewed altogether. There are times when such conjectures are necessary. There are times, too, when they are interesting – in trying to understand perceptions of accent or the take up and circulation of imitative conventions, for example.

Furthermore, as essential as phonological questions have been to literary dialect research, this study seeks a broader scope. The coding of the data accounts not only for respellings that are phonologically motivated, but phonologically unmotivated ones as well. It additionally identifies grammatical and lexical features. In total, the coding organizes 222 different features into four superordinate categories: lexical, morphosyntactic, orthographic, and phonological. There are 10 lexical features, 85 morphosyntactic features (grouped into 12 subcategories), 2 orthographic features, and 125 phonological features (grouped into 7 subcategories). Although the coding is extensive, neither is it intended to be nor should it be taken as a complete catalogue of every literary dialect feature present in the data. Indeed, the very notion of "completeness" is a problematic one. Many literary dialect features may be

obvious, but others can be frustratingly ambiguous – something that I grappled with in coding the data.

Such ambiguities, in combination with the arbitrariness of data availability and the complexities of data selection, place limitations on some of the study's findings. (These limitations are introduced in this chapter, but are explained more fully in the statistical overview.) The data, nevertheless, provide a rich account of literary dialect practices – which enables the quantitative descriptions that are developed in the chapters that follow. Importantly, too, they both focus and underscore the qualitative analysis that provides context for the quantitative patterns. The discussion of the methods that were used to produce the corpus and its attendant data begins with descriptions of the processes of collection and preparation (§3.2 and §3.3). That is followed by a brief overview of the corpus (§3.4) and an explanation of the coding (§3.5), which includes some discussion of the dilemmas involved in carrying out the latter.

3.2 Data collection

3.2.1 Corpus internal data

The source works include novels, novellas, short stories, and plays and were collected from the following digital archives:

- Eighteenth Century Collections Online (http://quod.lib.umich.edu/e/ecco/)
- Google Books (http://books.google.com)
- Internet Archive (https://archive.org/)
- Literature Online (http://literature.proquest.com)
- Nineteenth Century Collections Online (www.gale.cengage.com/ncco/)
- Project Gutenberg (http://www.gutenberg.org/)
- The Salamanca Corpus (http://salamancacorpus.usal.es/SC/index.html)
- The University of Oxford Text Archive (http://ota.ahds.ac.uk/)
- Women's Genre Fiction Project (http://womenwriters.library.emory.edu/)

The archives themselves were chosen for a number of reasons. Some like Project Gutenberg and the University of Oxford Text Archive have been used in previous historical, corpus-based research (e.g., De Smet, 2005). Others like Google Books and the Internet Archive are extremely large (containing billions of words) and thus provide access to an extraordinary number of works. Still others like the Salamanca Corpus and the Women's Genre Fiction Project provide access to specific types of works. The Salamanca Corpus contains works with literary dialect, though primarily

regional dialect, and the Women's Genre Fiction Project contains adventure novels that are important producers of literary dialect at the turn of the century.

For works to be included in the corpus, they must: 1) have literary dialect voiced by characters identified as African diasporic, Indian, or Chinese, and 2) be written by British authors. Of these, character identification was the most straightforward criterion to apply. One work, however, did present a borderline case. The character of Ling-Wong in Harry Collingwood's (1915) *A Chinese Command* is identified as Korean, but also as speaking "pidgin English." On that latter basis, his dialogue was included in the corpus. It serves as a test of resemblances and of claims that representations of Chinese speakers come to be generalized to imagined Asian identities more broadly (e.g., Chun, 2004). The results are discussed in chapter 7 (§7.4).

The other parts of the criteria raise questions that are more knotty: What counts as literary dialect? And what counts as British? In answer to the first, this study largely adopts Blake's (1981, p. 13) definition of literary dialect as signification through contrast. It is marked as different from the normative voices of other characters or the narration. Where this study diverges from Blake somewhat is in his emphasis on phonology. Blake (1981, p. 15) asserts that orthographic approximations of phonology are the principal form of marking. This study makes no such assumption. Differentiation can be realized lexically, morphosyntactically, orthographically, or any combination thereof.

One of the study's most formidable challenges was locating works without biasing the data. It would be tempting, for example, to use words that are common in literary dialect like *massa*, *dat*, or *hab* as search terms. While such an approach would surely yield works containing literary dialect, it would also lead to an over-representation of those terms. What would it mean that *massa* is the most frequent word, if that word was pre-determined? The protocols used for searching would dictate at least some of the corpus content.

In order to avoid that kind of selection bias, I adopted two approaches. First, I used reports from existing scholarship. For example, Mungo in Bickerstaff's *The Padlock* is a widely studied character, and his dialogue is recognized as one of the earliest examples of African diasporic literary dialect (e.g., Cooley, 1997). In addition to specific scholarship on literary dialect, I also relied on scholarship that analyzes performance and represenationality from the period without particular interest in

language (e.g., Forman, 2013; Muthiah, 2012; Nechtman, 2010; Waegner, 2014; Yang, 2011). My second approach was to search databases using signifiers like *CHINESE*, *CELESTIAL*, *NEGRO*, *INDIAN*, *HINDOO*, *GENTOO*, etc. This approach and the previous one allowed me to locate characters, but the identification of literary dialect then required physically scanning the works.

As for what counts as British, the answer is somewhat more complex than the question might first appear. During this period of empire, scholars have argued that British identities can be complex and fractured (Kumar, 2000; Magee & Thompson, 2010; Ward, 2001). This study takes a broad view of "Britishness." More specifically, it includes authors who participate in what Darwin (1999) calls "composite Britannic culture." Magee and Thompson (2010, p. 32) describe Darwin's conceptualization as a "British-centered system of global communications, transmitting news, ideas and values." From this perspective, British authors are those whose works circulate, in publication and in performance, in that British-centered system. Thus, the corpus includes a work like *The Octoroon,* which was written by Dion Boucicault. Boucicault was born in Ireland. He achieved success in London before leaving for the United States to live for six years. It is there that he wrote *The Octoroon,* and it was first performed in New York in 1859. Because of this, the play is sometimes found in anthologies of American theater (e.g., Hischak, 2012; Richards, 1997). However, after premiering in the U.S., Boucicault returned to London, and the play made its debut in London in 1861. In London, the dramatic climax of the play – the suicide of the protagonist – caused a controversy, which led Boucicault to publish an open letter in the press and then to revise the ending for British audiences (Meer, 2009).

In performance and in popular discourse, *The Octoroon* is clearly a part of "composite British culture." The example of Boucicault's play is additionally instructive in highlighting the interconnected nature of that composite culture. Among the authors in the corpus, Boucicault is not alone in participating in a wider Anglophone circulation of ideas, aesthetics, and discourse. Perhaps no author exemplifies the complexities of globalized networks, of movement and of influence, than Charles William Doyle (who was introduced in §1.5).

The son of a British soldier, Doyle was born in India in 1852. He studied medicine in England, graduating from the University of Aberdeen. In the late-1880s, he settled in California, where his writing began to appear in publications like the San Francisco-based *The Overland Monthly*. His first novel, *The Taming of the Jungle*,

was set in India and drew on his Anglo-Indian upbringing. It is a romanticized portrayal of the people of Terai, contrasting, in Doyle's (1899, p. 5) words, their "Arcadian simplicity" to the "monstrous Pantheism of the Brahmin" and "low-caste Hindoos of the plains." His second novel, *The Shadow of Quong Lung*, was set in San Francisco and was clearly influenced by the sinophobic discourse circulating in California. Both novels were published in Britain and in the U.S., and each responded to distinct national anxieties. One offers what the author saw as a corrective to negative British stereotypes of India; the other embraces and reinscribes American fears of the "yellow peril." Together Doyle's novels illustrate the transatlantic flow of ideas and the cross-pollination that influences representationality. What is British, therefore, also affects and is affected by wider Anglophone currents of cultural circulation, with some authors like Doyle and Boucicault being more obvious in their hybridity. As one reviewer noted of Doyle, he could be identified with either British or American letters, quipping that he was deserving of "a double [literary] citizenship" (Bowker, 1900, p. 194).

3.2.2 Corpus external data

In the introduction (§1.5), I asserted that this study's analysis begins with a machine identification of patterns, which, in turn, informs a human one. In other words, there is the corpus data, which is subjected to a series of computational transformations in order to identify patterns. Those patterns, in turn, are tested for their significance. Finally, the computational results are contextualized and explained using qualitative, archival data. The relationship between qualitative and quantitative data in this study, as it was framed in chapter 2 (§2.4), is unconventional for corpus-assisted research. Most corpus research that includes qualitative data draws that data from the corpus itself, and presents that data in conventional formats like concordance lines. In this study, the corpus is comprised of what is essentially decontextualized data – data that have been stripped out of a collection of works in order to computationally analyze them. Context, then, needs to be reconstructed, as it were, by looking outside of the corpus for relevant evidences that might help to explain what the computational analysis reveals.

Some of those evidences come from obvious places: the source works from which the corpus was extracted. However, as Baker (2014, p. 29) observes, "context" (or , as he defines it, "the constraints on a communicative situation that influence

language use") is wide-ranging. Accordingly, this study explores the contexts that shape quantitative trends by drawing evidences from a variety of sources. The approach partly seeks to identify patterns in the discourse surrounding the speakers and language varieties that are the subject of the study. Locating those patterns was partly accomplished by consulting physical archives at the United States Library of Congress and the Folger Shakespeare Library, but was largely facilitated by the availability of extensive digital collections of historical materials. These include some of the same archives that were used for the source works like Google Books, the Internet Archive, Eighteenth Century Collections Online, and Nineteenth Century Collections Online, as well as others like The Spectator Archive, The Burney Collection, and Early American Newspapers. Artifacts in these archives can be found, for example by searching linguistically relevant terms like *DIALECT*, *ACCENT*, and *ENGLISH* with co-occurring terms like *NEGRO*, *INDIAN*, and *CHINESE*. As helpful as these resources are, such searches do not yield results that can be sorted in the ways that corpus data can. Baker (2014, p. 29) notes that with contextual data "researchers need to make judicious decisions about what to include and how far to go." It is ultimately up to the analyst to continually evaluate what is relevant and explanatory. Of the thousands of archival documents that I read and annotated for this study, those that are included here are the product of the winnowing process that Baker describes. The qualitative data were selected for their potential to instantiate attitudes and ideological positions, to illustrate reactions to historical events, and to elucidate quantitative findings. Further, they intersect not only with the computational data, but also with more contemporary secondary scholarship that additionally explain historical context.

3.3 Data preparation

While some source works come from archives that are curated and have been checked for accuracy, others are generated through optical character recognition. The works from these latter sources had to be compared against either scanned or hard copies of the originals. In addition, when possible, files from curated archives were also compared against scanned or hard copy versions. Although rare, a few transcription errors were found and corrected.

Two versions of each work were saved for analysis: the complete works and the VISiBL corpus, which was created by extracting the segments of dialogue that were then coded. A separate text file was made for each group of speakers present in

each work. For example, *The Sword of Peace* (Starke, 1788) has two separate files in the literary dialect corpus: one for the African diasporic dialogue and one for the Indian dialogue. The individual files containing the dialogue and which make up the corpus are referred to in the discussion and analysis as "texts." The complete novel and plays from which the texts are extracted are referred to as "source works."

Yet another concern with data selection is the edition of a given work. Many of the works in the corpus went through multiple printings. This is clearly important if we are considering the links between discourse and the political economy of the empire. Due to the nature of the publication process, there is an unavoidable lag between material events and fictionalized reactions. In addition to introducing uneven gaps in chronology, different versions of works can also contain editorial changes that effect feature coding. Consider the examples below from *The Highland Reel* by John O'Keeffe. Excerpt 1 is from a version published in 1789, and (2) is from a version published as part of 1809 collection edited by the playwright Elizabeth Inchbald:

(1) M^{C}GIL. This black dog here disturb'd me in a speech which wou'd have done honor to Cicero, to announce *Shelty, the piper!*

BEN. Why, Massa, I did taut –

(2) *M^{c}Gil*. You villain! you shouldn't have interrupted me at study – No, not for the Lord Advocate of Scotland!

Benin. [*Crying*.] Why, Massere, I did tought –

There are a number of clear differences in the dialogue of Benin. First, the later edition has changed "Massa" to the Francized "Massere." Additionally, while both have the *t*/*d*-for-*th* substitution word-initially in *thought*, the first has also respelled the remainder of the word as "taut." Finally, McGilpin's line that prefaces Benin's is also different in the two versions. One significant change is the manner in which Benin is addressed. In the older version he is referred to in racialized terms, as a "black dog," while in the later one he is referred to more generically as "villain." Thus, not only is the content of Benin's line altered, but so too is the context in which it is uttered. In order to mitigate both chronological lag and potential editorial changes, the corpus contains the earliest edition of a work when multiple editions exist.

Of course, this is possible only when those earlier editions are available in digital formats. This is an issue that De Smet (2005) discusses in relation to the construction of his historical corpus, The Corpus of Late Modern English Texts (CLMET). De Smet (2005, p. 79) notes that public digital archives like Project

Gutenberg often contain later editions of works, particularly works originally published in the eighteenth or early nineteenth centuries. In such cases, his protocol is to use the year of first publication for dating purposes, even if the corpus contains an edition published sometimes decades later. As the excerpts from the *Highland Reel* suggest, this practice may introduce some error into diachronic corpora like CLMET and VISiBL. However, it is also an unavoidable consequence of working with data that others have already digitized, as De Smet observes, and in trying to track diachronic changes, we would create even more error by using the date of an edition rather than the date of first publication.[5] Thus, I follow De Smet's practice of dating all works by their year of first publication. This includes plays, which are dated by first publication, not first performance. In most instances, determining the date of first publication is reasonably straightforward (a triangulation among publication database records, the records of libraries like Oxford's, and extant scholarship). In a few, however, this proved to be more complicated.

Nearly identical versions of *Don Juan*, for example, are attributed to two different authors. One published by The John Dicks Press in the late nineteenth century identifies John Buckstone as the author and lists its debut as occurring at the Adelphi Theatre in 1828.[6] Another is published in 1837, but with Charles Milner as the author and listing the City of London Theatre as the performance site. The confusion likely results from two dramatic versions of the Don Juan story, both of which are submitted to the Lord Chamberlain in 1828 (*Lord Chamberlain's plays*, 1828). The first, a two-act burletta, is titled *A New Don Juan* and premiered at the Adelphi Theatre. The second, a two-act musical drama, is titled *Juan's Early Days* and premiered at Drury Lane. No authorship is provided for either original manuscript, but *A New Don Juan* is subsequently published by T. Richardson with Buckstone appearing as its author (Buckstone, 1828). *Juan's Early Days*, however, is

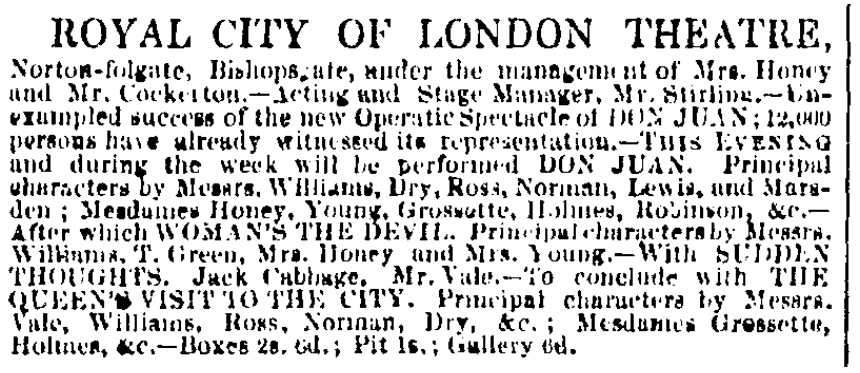
[5] In her corpus examination of *quoth*, for example, Moore (2015) observes an unexpectedly high frequency in fiction through mid-nineteenth century. The frequency, she argues, is an artifact of Google's n-gram being organized by date of publication (rather than date of first publication). Thus, it preserves older forms in reprintings of canonical works like Shakespeare's plays.

[6] The John Dicks Press was in the business of reprinting earlier works and selling them cheaply. (Copies of Dicks Standard Plays sold for a penny each.) According to a history of the press authored by one of Dicks' decedents, publication of the series began in 1874 (G. Dicks, 2006). A catalogue of Dicks Standard Plays printed a decade later lists 1072 titles (J. T. Dicks, 1884). As a result of its success and popularity, one of the press' legacies is that it has preserved works that might otherwise have been lost. And many of these have made their way into digital archives like the HathiTrust. For a discussion of Dicks' first printings of Shakespeare's works see Young (2012). An earlier article by Richard Altick (1958) has a more general discussion of the press and its importance to the print culture of the late nineteenth century.

the antecedent of the version that is included in the corpus. Though it premieres in 1828, it has a revival in 1837 at the City of London Theatre, with the theater's manager, Laura Honey, playing Don Juan to great acclaim.[7] The success of that revival is what apparently prompts the play's publication that same year.

To complicate matters even further, an 1828 publication of the songs from *Juan's Early Days* (Milner & Reeve, 1828) lists the author as Henry M. Milner (primarily known for his theatrical adaptation of Mary Shelley's *Frankenstein*), not Charles Milner as does the 1837 edition. A review of the debut is of little help, noting only that "the adaptation […] has been assigned to a Mr. Milner about whom we know no more than that he has produced several successful pieces at one of the minor theatres" ("Drury-Lane theatre," 1828). Neither does a review of the 1837 production clear things up in observing that the play is "dramatised by Milner and Stirling" (Phillips, 1837).

Figure 3.1: A notice advertising *Don Juan* from *The Morning Chronicle*, November 6, 1837.



ROYAL CITY OF LONDON THEATRE, Norton-folgate, Bishopsgate, under the management of Mrs. Honey and Mr. Cockerton.—Acting and Stage Manager, Mr. Stirling.—Unexampled success of the new Operatic Spectacle of DON JUAN; 12,000 persons have already witnessed its representation.—THIS EVENING and during the week will be performed DON JUAN. Principal characters by Messrs. Williams, Dry, Ross, Norman, Lewis, and Marsden; Mesdames Honey, Young, Grossette, Holmes, Robinson, &c.—After which WOMAN'S THE DEVIL. Principal characters by Messrs. Williams, T. Green, Mrs. Honey and Mrs. Young.—With SUDDEN THOUGHTS. Jack Cabbage, Mr. Vale.—To conclude with THE QUEEN'S VISIT TO THE CITY. Principal characters by Messrs. Vale, Williams, Ross, Norman, Dry, &c.; Mesdames Grossette, Holmes, &c.—Boxes 2s. 6d.; Pit 1s.; Gallery 6d.

A notice from *The Morning Chronicle* (see Figure 3.1), however, offers a clue. Though it makes no reference to Milner, it does mention Stirling as the stage manager and promotes another of Stirling's plays, *Woman's the Devil*, on the same bill. The notice also lists the same cast as the 1837 edition of the play. The Dicks Press edition lists that cast, as well, but with the notable substitution of Buckstone for Ross (whose name appears in both the Milner edition and the notice) in the role of Lambro. Based on that evidence, it seems more likely that Milner not Buckstone is the author of the version of the play in the corpus, and I assigned authorship accordingly.

The authorship question, while an interesting puzzle, is a decidedly secondary concern to the date of the work, however. Because the 1837 edition is the earliest published version of the play that I could locate, I use that to date the work. And this

[7] There are a number of surviving prints that depict her in the role of Don Juan.

gets back to what I was showing previously with the excerpts from *The Highland Reel*. Although much of the literary dialect in the 1837 edition is the same as literary dialect in the manuscript written a decade earlier, there are changes. The earlier version uses word-final *-a* insertions (e.g., *blacka*) that do not appear in the later version, for example.

Finally, there are two works that deviate from the normal protocols for dating works. The first is *Americans Abroad* by Richard Brinsley Peake. The only published edition that could be located was printed by The John Dicks Press in the late 1870s or early 1880s. The play, however, premiered in 1824. Its star, Charles Mathews, was a widely known performer, and the play's early nineteenth century run is well documented. Because of the gap between the premier and the available published edition, the play is dated by the copy that is included in *Lord Chamberlain's Plays* in 1824 (*Lord Chamberlain's plays*, 1824). That copy appears under the title *Jonathan in England*, but is otherwise largely identical to the Dicks Press edition. The other work that made the process of dating more difficult is *Obi, or, Three-Fingered Jack*. Like *Don Juan*, the play's central character (an escaped slave and Jamaican folk hero, Jack Mansong) and plot were popular in the early nineteenth century and recycled by a number of authors (see, e.g., Earle & Aravamudan, 2005; O'Rourke, 2006). One popular stage version of the story was a pantomime written by John Fawcett, which premiered in 1800. The version that is included in the corpus is a melodrama attributed to John Murray, but the edition is undated. To date the play, I rely on the scholarship of Rzepka (2002), who fixes its production date as circa 1830. Note, too, that Rzepka argues that the assignation of John Murray is incorrect and that the author is, in fact, a theater manager, William Murray.

3.4 The Voicing Imperial Subjects in British Literature (VISiBL) corpus

The resulting corpus is comprised of dialogue extracted from 126 novels, plays, and shorts story collections, beginning with Isaac Bickerstaff's *The Padlock* in 1768 and ending with Edgar Rice Burroughs' *The Monster Men* in 1929. It includes 35 plays and 91 novels or short story collections, and represents the works of 116 separate authors. Table 3.1 provides a summary of the corpus composition. For a full list of the individual works refer to Appendix A. As noted earlier, the fiction from which the dialogue was extracted is referred to as "the source works." The individual text files that are produced as the result of the extraction process are referred to as

"texts." The total number of literary dialect texts (n = 136) is greater than the total number of source works (n = 126). Ten of the works contain literary dialect from two different groups of speakers, while no work contains representations of all three. Thus, some source works produce two corpus texts – one for African diasporic dialogue and one for Indian dialogue, for example.

Table 3.1: Composition of the VISiBL corpus.

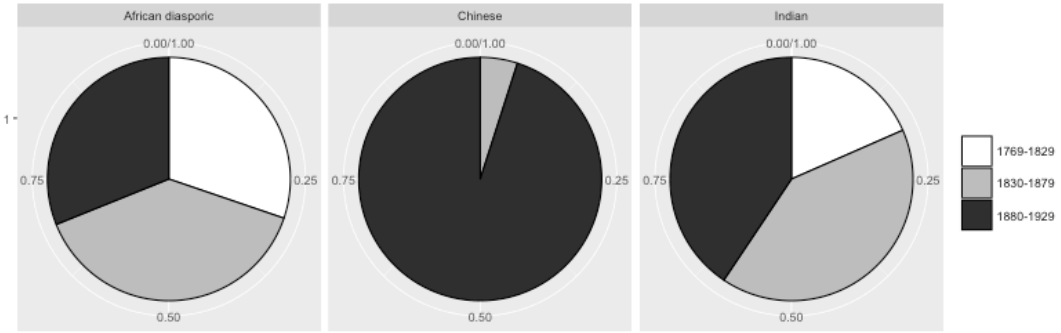
	Texts:	Words:
LITERARY DIALECT		
TOTAL	**136**	**51151**
African diasporic	60	26541
Chinese	39	7971
Indian	37	16639
SOURCE WORKS		
TOTAL	**126**	**7952399**

Note that the number of literary dialect tokens in comparison to the total number of words in the complete works is very small. The literary dialect corpus makes up less than 1% of the complete works. In many of the novels and plays, the characters whose dialogue is included in the corpus are minor ones. They are often servants, cooks, or shopkeepers with limited roles in the narrative and few lines. In the works where this is not the case, where relevant characters are more narratively central and have many more lines of dialogue, the number of words is capped at approximately 1000 by using the sample function in R to randomly select from all of a speaker's utterances. This process involves complete lines of dialogue. Utterances are sampled until the word count exceeds 1000; it does not cut off part of an utterance to achieve a precise 1000 word limit. In the few instances where I have multiple works from the same author, I have also limited that author's contributions to roughly 1000 words for any one group of speakers. Although this corpus is much smaller and more specialized, these methods follow those set forth in the construction of historical corpora like the Corpus of Late Early Modern English (De Smet, 2005). The aim is to get the largest possible range of authors and to statistically limit the influence of any single writer's practices.

In addition to making provisions for authorship, I similarly tried to account for publication date in constructing the corpus. The changing conventions in the voicing of different groups, however, make the kind of period balancing (e.g., having the same number of words for each decade) that is traditionally done in diachronic

corpora difficult. Rather than by decade, word counts were balanced across longer period divisions. Figure 3.2 shows the word counts for each speaker group (as a percentage of the total for that group) across three periods: early (1768-1829), middle (1830-1879), and late (1880-1929). As the chart makes clear, the data are roughly balanced across all three periods for African diasporic speakers. However, the word counts for Indian speakers increase after the early period, reflecting a greater availability of texts, and the word counts for Chinese speakers are concentrated in the later period. This is because Chinese literary dialect does not enter the corpus until Henry Addison's (1858) *Traits and Stories of Anglo-Indian Life*.

Figure 3.2: Pie charts showing percentages of word counts by period (1768-1829, 1830-1879, and 1880-1929) and controlling for speaker.



Although the separation of historical corpora into sub-periods is essentially arbitrary, as Hilpert and Gries (2009) observe, the divisions in this corpus reflect a number of statistical and theoretical considerations. First, the time period divisions need to be relatively large because the corpus is so small. Another consideration is the relative scarcity data in the early period. Because of changes in production and consumption, far more books are available in the middle and later periods (see, e.g., Erickson, 1996; Feather, 2006). In order to better balance word counts, the middle and late periods cover fifty years, while the early period extends for an additional twelve.

Finally, the partitions reflect significant historical inflection points that are recognized in other scholarship. 1830 marks the end of the Georgian era. More significantly, it is also the beginning of a period of tremendous imperial expansion and a time of increasing governmental involvement in the administration of the empire. These changes are accompanied by shifts in ideologies that scholars have mapped onto concomitant changes in literature (see, e.g., Brantlinger, 1988; Fulford & Kitson, 1998; Trumpener, 1997). If 1830 marks a point of rising imperial

enthusiasm, 1880 is sometimes recognized as a moment of rising imperial anxieties, a time when Britain's power was subject to increasing competition (see, e.g., Kennedy, 2002; Thompson, 2000). The third period covered by the corpus is also sometimes considered a transitional era in British literature (e.g., Lauterbach & Davis, 1973), and one marked by transformations in print culture that circulate Orientalist discourse through new currents of mass publication (Ardis & Collier, 2008; Long, 2014). Ending the corpus in 1929 gives the corpus a total span of approximately 160 years that centers on the nineteenth century.

3.5 Data coding

Once the corpus was compiled, the data were coded using the UAM Corpus Tool (O'Donnell, 2009), which allows for multiple layers of customized annotation. For this study, two layers of annotation were created. The first layer codes segments of dialogue by speaker – African diasporic, Chinese, or Indian. The second layer codes words or phrases within that dialogue by linguistic feature. These linguistic features are annotated according to a scheme developed "from the ground up." In other words, the scheme was not established prior to the coding process. Rather, codes were added as needed in order to describe contrastive features as they arose. Many of the codes come from extant scholarship on language variation and describe well-documented features like the zero copula.[8] Others describe features that are either rare or do not correspond to speech of communities in the real world.

In total, there are 222 different codes, which are separated into four main, superordinate categories (see Appendix B for the coding taxonomy and Appendix C for descriptions of the codes):

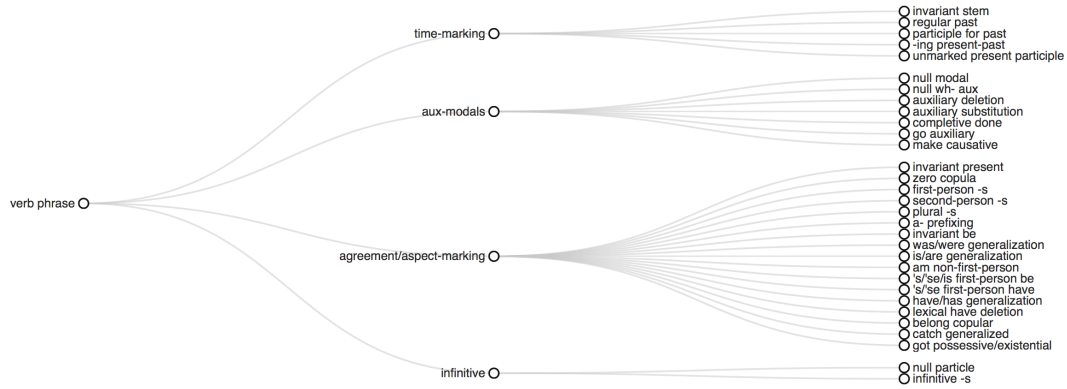
1. **Lexical**: word usage including general vocabulary, forms of address, inserts, words conventionally belonging to one part-of-speech being used as another, and code-mixing.
2. **Morphosyntactic**: word formation and grammatical patterns including the morphosyntax related to noun phrases, pronoun cases, verb tense marking, verb agreement, verb aspect, auxiliary verbs, adjectival and adverbial modification, and discourse organization.
3. **Orthographic**: unconventional spelling that approximates the spontaneous discourse of "standard" speakers but is used to mark difference (what is typically referred to as "eye-dialect").

[8] The codes corresponding to documented features were particularly informed by the work of Kortman and Szmrecsani (2004) on global Englishes, Kouwenberg and Singler (2008) on pidgins and creoles, and Bolton (2002) on Chinese Pidgin English.

4. **Phonological**: respellings used to approximate differences in phonology.

The protocols for creating and assigning codes adhere to just a few basic principles. The first principle is to group like with like when generating codes – for example, to group features related to verbs, as much as possible, under one category. The purpose is to be able to highlight systems that tend to be writers' focus of manipulation, and most of time this is a fairly straightforward process. However, lexical, syntactic, and phonological systems interact. Some features can straddle categorical divisions and could reasonably be classified in a number of different ways. Take the verb phrase subcategory that is illustrated in Figure 3.3. This is a morphosyntactic subcategory, and most of the features fit comfortably under that umbrella. However, others (like generalized *catch*, for example) are not prototypical. There are a variety of possible strategies for identifying such a feature: perhaps creating part-of-speech subdivisions in the lexical category or creating a fifth superordinate category that attempts to capture the blurry middle ground between lexicon and grammar. The preference here is for grouping features that belong to similar subsystems (like the verb phrase) together whenever that is plausible. There are, of course, drawbacks and benefits to such an approach, some of which are discussed later in this chapter and some of which will become apparent in the subsequent analysis.

Figure 3.3: A diagram showing the taxonomy of the verb phrase subcategory.



Another principle is to be conservative when applying codes, which plays out in a couple of different ways. One way that conservatism applies to the coding protocols is that as few codes as possible are attached to features in order to sufficiently describe such features. This consideration primarily affects respellings

and their relationships to the phonological and orthographic categories, which I discuss in detail shortly. A second is that ambiguous structures (i.e., structures which may or may not be marking difference) are not coded. The determination of "difference" is based on internal patterns and can be admittedly tricky and subjective. The problem is particularly acute with features like reduced clause structures (e.g., null clausal subjects), as those occur widely in the dialogue of all speakers. Therefore, an effort was made to apply codes only to those features that demonstrate systematic dissimilarities within a source work.

Consider the clause in bold from Edward Howard's (1836) nautical adventure *Rattlin, the Reefer*:

(3) Eh! Massa Ralph, suppose **no marry me to-day** – what for you say no yes to dat?

The line is spoken by Miss Belinda Bellarosa, whom the narrator describes as "a nice, matronly, free mulatto, who was a mother to me." The clause in question is a complement of the verb *suppose*. The verb in that clause (*marry*) has a subject (*you*) that is implied but not explicitly expressed. This structure can be compared with other occurrences of *suppose* complements in the dialogue of other characters and in the narration (e.g., "I suppose **that <u>you</u> have some favour to ask**" and "we may suppose **<u>the wrathful lioness</u> springs upon the buffalo**"). It turns out that all other examples of *suppose* complements have explicitly expressed clausal subjects (which are underlined in the examples). The instance from excerpt 3, therefore, is coded as a null subject. An important implication of this conservative approach is that the study makes no claims to being a complete catalogue of every literary dialect feature present in the dialogue under consideration. It is, instead, a robust accounting of that dialogue – one that enables meaningful statistical comparisons between representations and across time.

To conclude the discussion of coding, the next two subsections explore dilemmas that were faced when developing the coding scheme and applying codes to features. These discussions are intended to illustrate the reasoning behind some of the more difficult decisions. They are not meant to suggest that the ultimate solutions are the only ones or even the "best" ones in some objective sense. Their purpose is to make explicit some of the difficulties the data present and the ways in which those difficulties were addressed. In doing so, these discussions help to frame both the strengths and limitations of the data analysis that is presented in the following chapters.

3.5.1 Lexical versus morphosyntactic features

It is important to recognize that the four main categories represent linguistic classifications that are complex and connected, not separate and distinct. It is difficult, for example, to separate lexicon from syntax as words and grammar work together to shape structure and meaning. The conception of lexicon and syntax as existing along a cline without a clear division is particularly central to systemic functional linguistics and its theorization of "lexicogrammar" (e.g., Halliday, 1961; Hasan, 1987). It is also a concern of corpus approaches to syntagmatic analysis. Sinclair (2004), for example, argues for what he terms "lexical grammar."

In order to illustrate some of the questions surrounding categorization, let us consider *heap* as an intensifier. It appears, for example, in the dialogue of Quong, a Chinese servant in Beatrice Harraden's (1908) novel *Interplay*:

(3) Yes, Mr. Stilling upstairs. He **heap** hungly to-day. Eaten cucumber big as clocodile.

One way of categorizing the feature would be to identify such uses as instances of class shifting, a lexical-type feature. The argument for this would be that *heap* is ostensibly a noun that is used in (3) as an adverb modifying *HUNGRY*. This, I think, would be a perfectly defensible code to assign to the feature.

There is, however, an alternative way of analyzing the feature. *Heap* can be looked at as a constituent of the phrasal quantifier *a heap of*. In the early nineteenth century, there is evidence of a reduced form (losing its preposition and equating roughly to "a lot") and which is particularly associated with regional variation in American English. Upon overhearing fellow his passengers aboard a Mississippi steamboat, Amos Parker (1835, p. 88) complains in his travelogue about "these western people" (by which he means "Kentuckians, Tennesseans, Mississippians, &c.") and their linguistic habits. One of his peeves concerns this use of *heap*:

(4) [T]he word *heap* has too much by far *heaped* upon its shoulders. "A *heap* better," "a *heap* easier," and "a *heap* of ladies," are phrases often heard. I may be a little sensitive, but the word *heap* is very disagreeable, and I wish it was expunged from the English vocabulary.

From this form appears to emerge the further reduced form (absent the determiner) that is evident in (3). That form is most commonly linked to representations of Native Americans, not just in fiction but also in film. Meek (2006) terms these representations "Hollywood Injun English" and identifies *heap* as an identifier of manufactured "Indianness." Similarly, Cutler (1994) observes the presence of *heap* in early nineteenth century depictions of Native American vocal culture and notes its

highly marked inclusion in *Roughing It* (1872), where Mark Twain claims "that 'heap' is 'Injun-English' for 'very much.'" More than twenty years earlier, the British writer George Ruxton (1849, p. 19) makes the almost identical assertion in his travelogue of the American West. In a footnote, he writes, "An Indian is always a 'heap' hungry or thirsty – loves a 'heap' – is a 'heap' brave – in fact, 'heap' is tantamount to very much."

The linguistic processes contextualizing *heap* as an intensifier, therefore, could be distinguished, for example, from those surrounding *joy* in the following excerpt from William Dimond's (1820) *The Lady and the Devil*:

(5) Missey touch the string so sweetly, Oh! it **joy** my heart to hear.

In (5), *joy* is a noun being used as a verb, rather than a constituent of a larger structure that takes on new functions as that structure changes. Importantly, too, it is a nonce formation, rather than part of a documentable process. By the latter, I do not mean a process that takes place within real speech communities. It is debatable whether *heap* as an intensifier was actually used by nineteenth century Native American speakers or whether its placement in the mouths of Native Americans was an invention of white writers. But even as a partially imagined feature, its evolution is traceable in works like those authored by Parker, Twain, and Ruxton.

Ultimately, then, the coding of a borderline feature like *heap* depends upon what one wants to emphasize. For my purposes, I chose to group *heap* with other features related adverbial and adjectival systems – systems of modification – which places it within the morphosyntactic category. And there are other features, like generalized *catch*, for which I have followed a similar line of reasoning. The upshot is that in the statistical analyses, the lexical category may be slightly under-represented. The benefit, however, is that systems of similar features (like systems of modification or systems of verbs) are largely grouped, instead of being spread across multiple superordinate categories.

3.5.2 Orthographic versus phonological features

The determination of whether a feature falls within the orthographic or the phonological category requires distinguishing between what is a phonologically motivated respelling and what is a phonologically unmotivated one. Making such a determination requires a kind of conjecture that, for the most part, this study tries to avoid: inferring the phonological intent of an author. The authors that are included in

this study came from a variety of places, spoke with a variety of accents, and lived in a variety of eras. Thus, pinning down the phonological import of a respelling like *countree* for *country* is difficult, to say the least. It is for this very reason that coding for the phonological category is organized by spelling and not by phonetics.

Yet, scholars like Kretzschmar (2001) argue that the distinction between what is phonologically motivated and unmotivated in literary dialect is an important one because both contribute to characterization, though often in very different ways. Moreover, although this study is not designed to address questions of accuracy, trends in phonological representations provide critical information regarding perceptions of accents and their enregisterment. As useful as this distinction may be, making it can sometimes prove more complicated than it may seem, however. Consider Kretzschmar's paradigmatic example: *wuz*. Certainly, the substitution of *z* for *s* is straightforwardly eye dialect. But what of the vowel change? To throw the matter into relief, what if the variant were *wus* rather than *wuz*? Is that spelling still clearly phonologically unmotivated? Or is it attempting to signal a change in the vowel quality? A fronting of the vowel? A raising? Or maybe even marking the devoicing of the final sibilant?

Although the variant does not occur in the corpus, it does appear in the source works, and it is useful in showing how a single change in orthography can complicate the interpretation of features, as well as in highlighting the particular challenges of vowels in isolation. In *Stand By!* by Henry Taprell Dorling (1916), *wus* appears in the dialogue of a sailor named Smith:

> (6) "Yes, sir," said the seaman, bursting with merriment. "'Cos the sick bay, and it weren't none too large, was all but filled up wi' six 'efty great casks, wi' flagstaffs and sinkers complete. They **wus** the buoys Number One 'ad bin talkin' abart all along."

Smith is described as "a massive, rotund, bull-necked individual, with a face the colour of a ripe tomato," but no information is given as to where he is from or what accent he is meant to be speaking. In the early twentieth century, this variant of *was* does appear in representations of Hiberno-English (e.g., St. John Greer Ervine's play *Mixed Marriage*), as well as representations of cockney, like in this excerpt from Randall Parrish's (1918) *Wolves of the Sea*:

> (7) "That **wus** part o' the luck, Tom," he acknowledged, his accent that of a cockney. "Did yer git eyes on thet new feller Manuel Estevan brought back with him in the boat?"

In light of some of the other features present in Smith's dialogue like word-initial *h*-deletions, it seems likely that Dorling is encoding a sociolect that is meant to signal Smith's class status, something like cockney though perhaps not cockney specifically.

That information, however, gets us little closer to figuring out what to make of the vowel change in *wus*, but there are some additional clues. There are a variety of other vowel substitutions in Smith's dialogue (e.g., *bin* for *been*, *git* for *get*, *cos* for *cause*, *fur* for *for*). Also, vowel pronunciation plays a central role in Smith's comic purpose: the telling of a story involving the confusion between *buoys* and *boys*. Those facts would seem to support an interpretation of *wus* as phonologically motivated. At the same time, there are respellings that appear to be more certainly eye dialect in Smith's dialogue (*sez* for *says*, *bizness* for *business*). It also seems arguable that *wus* is an instantiation of that phonologically unmotivated pattern.

If I had needed to, I would have coded *wus* as a *u*-for-*a* substitution – a phonological rather than an orthographic feature – for this particular context. That is where the preponderance of evidence seems to fall for me. However, as this scenario makes clear, this is not a cut and dried judgment. Alternatively, if the variant had been rendered *wuz* rather than *wus*, I would have coded it as an eye dialect feature. That estimation would have been based on its closer parallels to *sez* and *bizness*, as well as the conventionality of the spelling, which fossilizes in the late nineteenth century. But that, too, is hardly an unassailable position. One could argue that the spelling indicates a coloring of the vowel that is more North American than British. Ultimately, that ambiguity is the point of this discussion: there are few clear-cut cases, particularly when it comes to vowels.

There are other important considerations in coding eye dialect, too. For one, even if there are multiple alterations to a word (as in *bizness*, which changes the vowel and the sibilant, and shortens the word in an approximation of allegro speech) only one code is applied. For another, sometimes eye dialect features co-occur with phonological features in combinations that only receive a phonological code. I did not assign an orthographic-type code to features that realize combinations of phonological and orthographic features when the orthographic changes are necessary for disambiguation. The respelling of *love* as *lub*, for example, was only assigned a phonological-type code as the vowel change is needed to distinguish the word from *lob* or *lobe*.

3.6 Conclusion

These extended discussions of coding ambiguities are partly an effort to clarify, what I think, are some theoretical complexities that are inherent in working with written representations of speech – problems that are compounded by a data set that is spread across time and place. They have also been an effort to make clear my procedures for mitigating these complexities. These procedures reflect the study's emphasis on describing representational practices rather than evaluating their accuracy. They also underline the study's limitations, particularly that the coding is not a comprehensive record of every manipulation and every possible literary dialect feature.

As much as these caveats are important in framing the analysis that follows, they also need to be placed within the larger context of the study. Most of the features that are ambiguous and challenging to code are also relatively infrequent. Vowel substitutions, for example, comprise only 4% of the coded features. Of those, 60% co-occur with another phonological feature (like a consonant substitution), making them more interpretable. The feature I would have assigned to *wus*, *u*-for-*a* substitution, accounts for less than one-half of one percent of the coded features.

Additionally, I have emphasized the conservatism of the coding protocols. Even with that approach, the outcome is a data matrix with 29,240 points and 18,186 separate coded features. These data provide a robust foundation for the statistical analysis. Moreover, although the categorical placements of some features may be arguable, those placements affect superordinate category balances, but have no impact on calculations like analysis of variance, distribution, diversity indices, and hierarchical clustering, which are blind to taxonomical structure. These and other measures are the focus the next chapter, which introduces them and explores their application and explanatory potential. It also demonstrates how these measures can serve to highlight continuities and anomalies, which, in turn, can direct qualitative readings of textual evidence – evidence from the source works and other artifacts, literary and nonliterary, from the historical record.

Chapter 4
Statistical Overview

4.1 Introduction

This chapter outlines the statistical approaches used in the study, presents a summary of the quantitative data, and traces the contours of the analysis that is elaborated in the chapters that follow. Although the emphasis here is on quantitative data, the discussion includes qualitative data as well, in so much as that data is helpful in contextualizing quantitative patterns. That said, most of the qualitative data related to language ideologies and enregisterment – narration and description from the novels and plays, quotations expressing language attitudes from the imperial archive, etc. – come later. The focus for this chapter is on laying out the statistical techniques, patterns, and trajectories that are the study's foundation.

The chapter is divided into two main sections, preceded by some brief comments regarding a few computational tools and techniques (§4.2). The first main section (§4.3) looks at variation among the coded features. In total, the corpus was coded for 222 features divided among four main categories – lexical, morphosyntactic, orthographic, and phonological. Because there are so many features, the data tables presented in this chapter are selective. For the complete data tables, refer to Appendices D-G. The second section (§4.4) looks at variation among representations of groups of speakers. The statistical methods used in this part of the chapter engage three fundamental questions: 1) How many features are being used? 2) How many different kinds of features are being used? 3) How similar is the distribution of features in one text to the distribution of features in another? The composite frequency (the combined frequencies of the coded features in a text) addresses the first question. The second is answered using what is called a diversity index. A diversity index accounts for not only the number of different types of features that an author uses in a particular representation, but also how evenly those features are distributed. That means a respelling that occurs only once idiosyncratically is accorded less weight than one that occurs repeatedly and systematically. The third question is addressed using hierarchical cluster analysis. This is a method that has been used by Moretti (2005), Jockers (2013), and others in studies of literary genre and style.

These measures provide a detailed snapshot of literary dialect along three dimensions: 1) frequency, 2) diversity, and 3) similarity. That snapshot, then, is used to identify patterns of variation, which is the first step in addressing the first three research sub-questions (§1.3). One type of variation this section analyzes is variation within representations of vocal cultures. For example, how frequently do literary dialect features appear in representations of Chinese voices? How uniformly are those voices rendered and with what complexity? How do the complexities and consistencies of Chinese representations compare to those of African diasporic representations? This section also analyzes variation over time. For example, are there changes in the patterns of representing Indian voices? Changes in the amount features used to voice characters? Changes in the types of features?

What emerge from the data are representational patterns. These patterns throw into relief ideologies and logics related to race, language, and empire. They suggest through lines that bind texts to traditions and conventions, but they also highlight changes, tensions, and exceptions. Thus, the analysis here sets the stage for the discussions of enregisterment and indexicality, of language ideology and imperial anxieties, which follow in the ensuing chapters.

4.2 A few comments about the statistical analysis

Note that all statistical calculations were carried out in R, a computer language and programming environment (R Core Team, 2013), as were all data visualizations – the latter were created specifically using ggplot2 (Wickham, 2009). Calculations were produced using code that was written expressly for this project, with one exception: the code for dispersion measures, which is discussed in the next section (§4.2.1). Some of that code takes advantage of packages (like ggplot2), which are pre-compiled functions that can be downloaded and called in R. Where those are employed, they are cited accordingly.

Many of the statistical methods used in this and future chapters are common practice in corpus research. Normalized frequencies and log-likelihood comparisons are immediately recognizable to anyone who has engaged in the quantitative analysis of corpora. The former are produced by popular online corpora like the ones hosted by Brigham Young University, and the latter are part of the built-in toolkits of concordancers like WordSmith and AntConc. Even hierarchical clustering, which is more unusual, has an established track record in both the digital humanities and

corpus linguistics. There are, however, two statistical measures that are likely to be unfamiliar. I want to introduce these and my rationale for using them, before proceeding to the data.

4.2.1 Dispersion

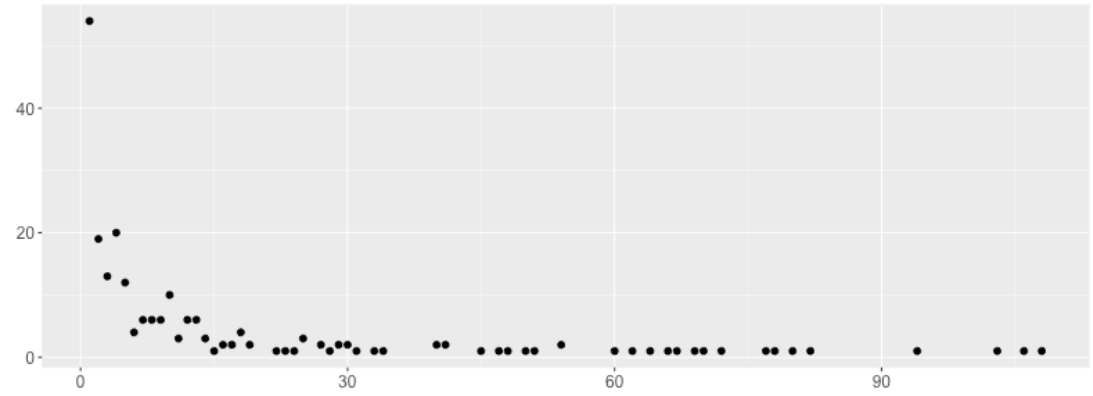
An obvious limitation of simple frequencies is that it is unclear whether a high frequency is driven by a large number of tokens in just a few texts or a similar number of tokens in many texts. Dispersion measures mitigate this problem by describing how a token is distributed throughout a corpus. The simplest method for doing this is to calculate the percentage of texts in which a token occurs. So if we were looking at the letter *a* in a corpus of two texts that each contain the letter *a*, the dispersion would be 100%. Again, this a calculation that is often a part of the tool-kits provided by concordancers. The weakness of a percentage calculation is that it reveals nothing about the potential variation within each text. Suppose one text has a single occurrence of *a* and the other has one hundred. The percentage would seem to elide important information about how that token is dispersed.

Thus, a number of dispersion measures like Juilland's *D* (Juilland, Brodin, & Davidovitch, 1970; Lyne, 1986) have been developed for corpus analysis. The VISiBL corpus, however, has an additional complication. Many dispersion measures work best when the components of a corpus are equal in size. When analyzing texts like complete novels, this is not a problem. Texts could be divided into equally sized chunks for the calculations. Unfortunately, the texts in VISiBL not only vary in size, but also are too small to chunk. The analysis, therefore, requires a dispersion measure that accounts for texts of differing lengths. Consider, again, our hypothetical example. What if the first text has only ten words and the second a thousand? That would affect our understanding of how *a* is dispersed.

For these reasons, dispersion is measured using deviation of proportions (DP). The measure was developed by Gries (2008; 2010), and as the name implies, is designed to interpret data from a corpus with constituent texts of varying lengths. The calculation of DP, as Gries describes it, is a three step process: 1) the size of each corpus part is determined and normalized against the total corpus size, 2) the frequencies of a token are determined in each of the corpus parts (for an observed frequencies) and normalized against the total number of occurrences of that token (for an expected frequency), and 3) the pairwise absolute differences between observed

and expected are calculated, summed, and divided by two. The result is a measure that varies along a 0 to 1 scale, making them easily interpretable. The only quirk is that the scale is inverted so that a result closer to 0 indicates a highly dispersed token and a result close to 1 indicates one that is minimally dispersed. The calculations were carried out using a modified version of the R code that Gries references in his article. Deviation of proportions is provided only for the 222 features, not for categories and subcategories in which they are embedded (e.g., there is a DP for *l*-for-*r* substitution but not for the consonant substitution or phonological categories). When applicable, deviation of proportions measures are included in the features tables. Additionally, a value of DP ≤ 0.80 is frequently used as a threshold for a feature's inclusion on the data tables in this and subsequent chapters. Note that this threshold is arbitrary. It focuses the analysis on approximately 20% of the 222 coded features. The maximum value excludes features that occur in roughly fewer than 10% of the texts in corpus (though deviation of proportions is more than a measure of range, so that percentage is not absolute).

Figure 4.1: Scatter plot showing the number of features (the y-axis) appearing within a given range of texts (the x-axis).



To get a broad sense of how coded features are distributed in the corpus, consider Figure 4.1. The scatter plot shows the number of features that appear in a given range of texts. 54 (almost a quarter) of the features appear in only one text, represented the highest point close to the y-axis. The analysis, however, concentrates on the more distributed features – those that are represented by the points extending along the x-axis where x > ~15. Importantly, deviation of proportions is calculated for four separate data sets: all dialogue, African diasporic dialogue, Indian dialogue, and Chinese dialogue. Thus, the 0.80 threshold for dispersion does not mean that the same

20% of features are analyzed in this and each of the subsequent chapters. Additionally, the complete data tables are available in Appendices D-G.

4.2.2 Diversity

In addition to explaining features, the analysis endeavors to explain texts, individually and collectively. Just as frequency measures are a fundamental tool for the former purpose, so too are they for the latter. Frequencies can indicate how marked an example of literary dialect is overall, and they can show how a text or a group of texts is structured. What are its constituents and how does it compare to others? As much as we can learn from those calculations, there are other things that we might want to know. We might, for example, be interested in the breadth of features that representations incorporate. From such information, we could make evaluations about the relative complexity of texts or examine whether or not complexity changes over time. One way of doing this would simply be to add up the number of different features that occur in a text, giving us a range. One text, say, has five features and another seven. We could conclude that the second text is more complex. However, like percentage as a measure of dispersion, a range as a measure of complexity has a number of shortcomings. For one, it does not capture how balanced or frequent features are. Does a text include just one idiosyncratic occurrence of a feature? Or is there a more sustained pattern of usage? What if all five of the features in the first text are repeated and highly frequent, while of the seven in the second, only one occurs more than a few times? Which would we consider more complex?

One measure that accounts for both the number of distinct elements and their distribution within a system is a diversity index. Diversity indices are typically used in ecology in order to measure biodiversity within a community or ecosystem (see, e.g., Magurran, 1988). Jarvis (2013) makes the case such measures might find productive application in linguistics – in calculations of lexical diversity, for example. One of the specific measures he focuses on is Shannon's diversity index (also sometimes called the Shannon-Wiener index), which has its origins in information theory. It was first developed by Claude Shannon (1948) as a measure of the unpredictability of an information signal or string of text, and in information theory is referred to as entropy. The insight in ecology was that high entropy (or high unpredictability) corresponds to greater diversity (whether textual or ecological). Further, such diversity is affected by

"richness" (the number of different types in a system) and "evenness" (their distribution). Although, as Jarvis observes, Shannon's diversity index is not widely used in linguistics, it has been applied in a number of studies measuring linguistic complexity along various dimensions (e.g., Juola, 2013; Utsumi, 2005; Utsumi, 2007).

In this study, Shannon's diversity indices are used alongside feature frequencies in order to provide complementary views of literary dialect. One tells us how many features are being used and the other how those features are being arrayed. Is frequency balanced or unbalanced among feature types? Concentrated or distributed? This allows us to pose new questions of the data. Do, for example, representations of African diasporic speakers get more complex over time? Or how does the complexity of African diasporic dialogue compare to that of Indian dialogue? All of the diversity calculations were carried out using the R package *vegan* (Oksanen et al., 2008).

4.2.3 Significance

When I first began this project, I consulted with a colleague who has extensive expertise in statistical methods for corpus analysis. His advice was to eschew p-values altogether. P-values, of course, are measures of significance, which indicate the probability of an observed effect being at least as extreme if the null hypothesis were true. His point was not that significance measures would have no meaning in the context of the study. Rather, he pointed to the variables that make extrapolating from the data difficult (e.g., the study is reliant on works that have already been digitized, sampling literary dialect from those digital archives is unpredictable, etc.).

Originally, I had planned on following his advice. As I began generating calculations, however, I changed my mind and decided to include them. They provide useful and recognizable ways of sorting results and modulating claims. Without them, it is difficult to indicate how salient particular features are within texts, portions of the corpus, or the corpus as a whole. That said, I think that his larger point is important to foreground. Measures of significance and their interpretation are bounded by the limitations of the data.

Another somewhat similar issue relates to coefficients of determination or r-squared values. These are generated as part of regression analysis and describe the goodness-of-fit of a regression model. Typically, they show how much variance a

linear model explains. Although they are not measures of significance (and I want to be careful not to conflate the two), they are used to demonstrate the robustness of a model. As such, they are important statistical indicators and, like p-values, need to be contextualized within the parameters of this study.

Many fields like bioinformatics prefer high r-squared values (often greater than 0.75, sometimes greater than 0.90) as evidence of a model's adequacy. This is because in such fields, one of the primary purposes of regression analysis is to produce a model that accurately predicts future iterations of the variable under consideration. The regression analysis in this study is not predictive. It would be impossible to forecast with any precision the frequency of a feature or feature category from a given novel published beyond the horizons of the study. The data are too noisy. They are produced by people with motivations, conscious and unconscious, that are influenced by a variety of cultural, economic, political, artistic, and material forces.

Rather than being predictive, the regression analysis in this study is descriptive. Some of the questions that this study engages relate to changes over time, and the r-squared values provide a measure of the explanatory power of models describing those changes. If, for example, there is a rising trend in the use of a feature or feature category, an r-squared indicates how robust that trend is. Although those values are lower than they would be for some other kinds of data (less than 0.50), those values are expected.

4.3 Feature variation

Table 4.1: Frequencies of superordinate feature types for all categories of speakers. *N* is the raw number of occurrences; *% Global* is the percentage a feature or category contributes to all coded features; and *Freq.* is the normalized frequency of a feature or category (per 1000 words).

Feature	N	% Global	Freq.
FEATURES-TYPE			
TOTAL	**18186**		**355.54**
lexical	3524	19.38%	68.89
morphosyntactic	7432	40.87%	145.30
orthographic	190	1.04%	3.71
phonological	7040	38.71%	137.63

As discussed in the previous chapter, literary dialect features are coded into a taxonomy. This section provides quantitative descriptions of each of the four main

classifications of that taxonomy (lexical, morphosyntactic, orthographic, and phonological). Before proceeding, however, I want to comment briefly on the distributions of the main classifications, themselves (see Table 4.1), as these distributions suggest some generalized patterns. For one, they indicate the importance of phonologically motivated respellings in the corpus. Their relative presence in the corpus (making up 39% of all coded features)[9] at least partially accounts for the historical emphasis on phonology in literary dialect studies. That said, the frequency of phonological features is equaled by the frequency of morphological features (which make up 41% of all coded features). This fact supports the argument suggested by other scholars, that the emphasis on phonology has sometimes obscured the salience of other kinds of features in the rendering of literary dialect (see, e.g., Traugott, 1981).

In contrast to phonologically motivated respellings, unmotivated respellings and ambiguous respellings (which fall within the orthographic category) are far less common in the corpus (accounting for only 1% of coded features). Their relative rarity can be partly explained by the conservative approach that I have taken in coding the features. I noted in my methodology (§3.5.2) that the interpretation of vowels is particularly thorny given the size of the corpus and the range of authors it includes. Because of that, I chose to resist speculating on their phonological salience as much as possible, but instead coded them by their respellings. Thus, there are features that might reasonably be recoded as orthographic rather than phonological. Even with that potential, a few highly marked (and enregistered) phonologically motivated features far surpass unmotivated ones, as we shall see.

Finally, though lexical features trail morphosyntactic and phonological features in their frequency, they still comprise a robust 20%. Lexical features are also some of the most widely dispersed items in the corpus. Moreover, their frequency only hints at their larger indexical role in signaling relative status, power, and identity.

[9] Figures are rounded up to the nearest percentage.

4.3.1 Lexical features

Table 4.2: Frequencies of lexical features for all speakers, where DP ≤ 0.80. *DP* is the deviation of proportions (a dispersion measure).

Feature	N	% Global	Freq.	DP
LEXICAL-TYPE				
TOTAL	**3524**	**19.38%**	**68.89**	
address	1984	10.91%	38.79	0.33
self address	329	1.81%	6.43	0.48
general vocabulary	494	2.72%	9.66	0.54
lexical substitution	126	0.69%	2.46	0.58
inserts	260	1.43%	5.08	0.66
class shifting	38	0.21%	0.74	0.71
code-mixing	206	1.13%	4.03	0.74
wh- word	33	0.18%	0.65	0.80

Of the lexical features, forms of address are the most frequent and most dispersed (see Table 4.2). They account for 11% of all of the coded features. While their high frequency suggests that forms of address are common particularly in comparison to other kinds of lexical items, they only hint at the importance of address in the marking of subaltern voices during this period. One complicating factor has to do with potential opportunities for address as opposed to a feature like *t*/*d*-for-*th* substitution. Consider the excerpt below from Banks' (1882) *Through the Night*:

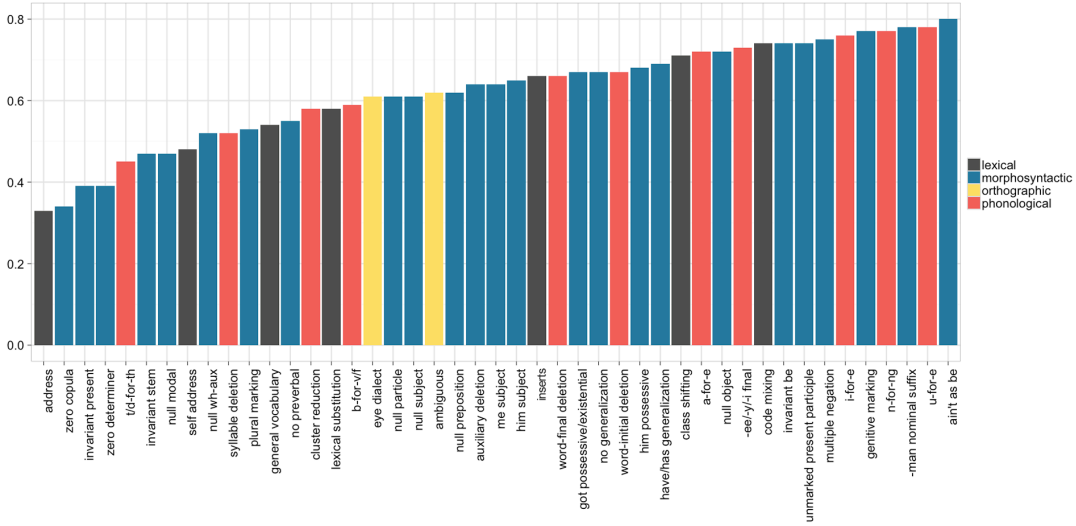
(1) **Massa** Walcot, **de** 'mighty God above send Cuffy to warn you. **Dere** am doom on **dis** house till Corbyn heir be found, and **de** first thun'erbolt fell last night. For own sake, **Massa** Walcot…

In the excerpt, the speaker, Cuffy, addresses Walcot twice as "Massa." The same excerpt has four examples of *t*/*d*-for-*th* substitution. Because the latter is often realized in common function words like *the*, *there*, *this*, and *that*, there are more opportunities for the phonological feature to appear. Thus, their relative frequencies are influenced by structural constraints.

In order to mitigate these constraints, frequencies are complemented by dispersions calculated by deviation of proportions. Providing some context for the dispersions of lexical features, Figure 4.2 includes the features from all four superordinate categories where DP ≤ 0.80. The plot shows address-type features to be not only the most dispersed lexical feature, but also the most dispersed feature from any category. The most frequent form of address is *massa* (n = 741). Other common forms include *sahib* (n = 348), *sah* (n=165), *missy* (n = 124), and *sar* (n = 109). Self-address is the second most dispersed lexical feature and the eighth most dispersed

feature overall. Self-address is most commonly realized as characters referring to themselves by their own names (e.g., *Shedallah*, *Ching*, *Smutta*, *Snowball*), in phrases headed by the modifier *poor* (e.g., *poor negro*, *poor black man*, *poor slave*, *poor black*), or in combinations of the two (e.g., *poor Wowski*, *poor Gangica*, *poor Snowball*, *poor old Boule-de-neige*). Address and self-address often work together in positioning nonstandard-speaking characters in relation to their standard-speaker counterparts.

Figure 4.2: Bar plot showing the deviation of proportions for features in all dialogue with DP ≤ 0.80 and color-coded by category.



The other lexical feature that I would like to touch on briefly is general vocabulary, which is the third most dispersed lexical feature and the twelfth most dispersed feature overall. Blake (1981, p. 15) argues that vocabulary has a less prominent role in literary dialect because of its potentially ambiguous social signaling – a word or phrase may be informal or colloquial without being stigmatized or being strongly associated with a stigmatized community of speakers. The relatively lower frequency of the lexical category in the corpus would seem to confirm that claim, at least partly. However, specific lexical-type features like address are clearly salient in marking certain kinds of literary dialect. Moreover, although general vocabulary is less dispersed than seven morphosyntactic features and two phonological features, it is still among the more dispersed features in the corpus.

The general vocabulary category includes 118 different lemmatized token types. The most common tokens are variants of *SAVVY* (n = 59, DP = 0.79), *PICANINNY* (n = 46, DP = 0.87), and *BUCKRA* (n = 33, DP = 0.89). As the deviation of proportions

suggest, not even the most frequent tokens are highly dispersed. For context, *SAVVY* appears in 23 texts, *PICANINNY* in 12, and *BUCKRA* in 8. Dispersions can be affected by speaker restrictions. For example, variants of *BUCKRA* occur only in African diasporic dialogue. There are similarly restricted tokens for other groups of speakers like *CHOP CHOP*, *CHOW CHOW*, and *CHIN CHIN*, which occur only in Chinese dialogue. Tokens can also be chronologically restricted. The texts in which *BUCKRA* appears are published in the 1820s and 1830s, with the exception of *Pirate of the Carribees*, which is published in 1898. Variants of *PICANINNY* occur in texts published largely in the middle to late nineteenth century. By contrast, *SAVVY* is one of the few tokens (*BOBBERY* being another example) that is specific neither to time period nor speaker.

4.3.2 Morphosyntactic features

Table 4.3: Frequencies of morphosyntactic subcategories for all speakers.

Feature	N	% Global	Freq.
MORPHOSYNTACTIC-TYPE			
TOTAL	**7432**	**40.87%**	**145.30**
verb phrase	3636	19.99%	71.08
noun phrase	1462	8.04%	28.58
pronoun	1116	6.14%	21.82
discourse organization	556	3.06%	10.87
negation	459	2.52%	8.97
adjective-adverb	172	0.95%	3.36
complementation	31	0.17%	0.61

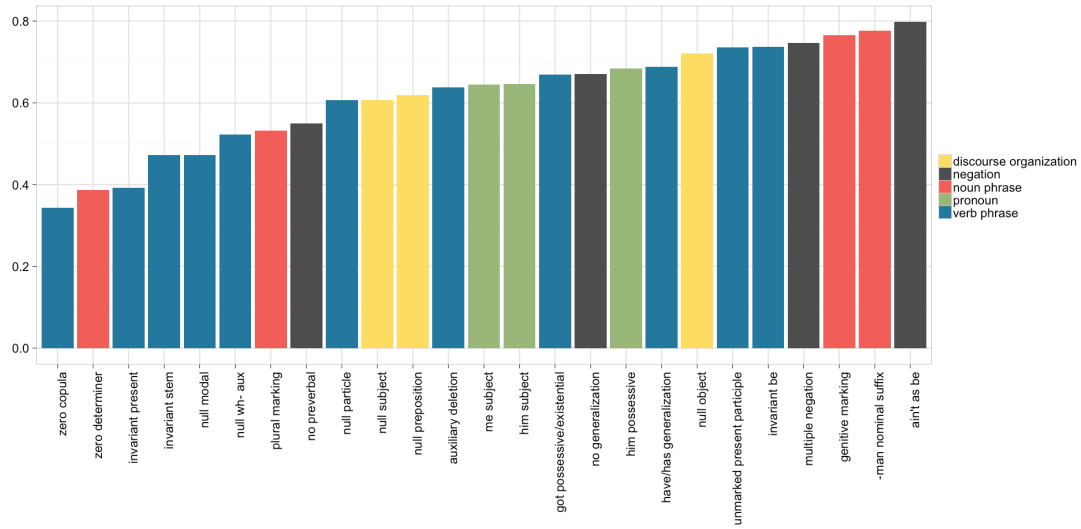
Much like lexical features, morphosyntactic features are salient – if sometimes overlooked – constituents of literary dialect. As I pointed out at the beginning of this chapter, morphosyntactic features are on par with phonological features in their overall frequency. Additionally, as Figure 4.1 shows, morphosyntactic features make up a large number of the most dispersed features in the corpus. Morphosyntactic marking occurs most often in the verb phrase (see Table 4.5). The highest frequency of that marking is related to agreement and aspect, or in other words the structuring of grammatical relationships between verbs and their subjects and between verbs and the flow of time. Agreement/aspect features make up 11% of all literary dialect features. Secondarily, time-marking-type features and auxiliary/modal verb features each account for approximately 4% of the coded features.

Table 4.4: The ten most dispersed morphosyntactic features for all speakers.

Rank	Cat.	Feature	N	% Global	Freq.	DP
1	VP	zero copula	977	5.37%	19.10	0.34
2	NP	zero determiner	1112	6.11%	21.74	0.39
3	VP	invariant present	592	3.26%	11.57	0.39
4	VP	invariant stem	607	3.34%	11.87	0.47
5	VP	null modal	350	1.92%	6.84	0.47
6	VP	null *wh*- aux	120	0.66%	2.35	0.52
7	NP	plural marking	202	1.11%	3.95	0.53
8	NEG	*no* preverbal	323	1.78%	6.31	0.55
9	VP	null particle	168	0.92%	3.28	0.61
10	DO	null subject	257	1.41%	5.02	0.61

The most dispersed of the morphosyntactic features (and the second most dispersed feature overall) is the zero copula (see Table 4.4). Its dispersion is, perhaps, even more surprising than its frequency. Although its presence has been mentioned in relation to literary representations of varieties like Appalachian English (Nickell, 1984) and Yorkshire English (García-Bermejo Giner & Montgomery, 2001), it has been most widely discussed in the context of African American English (e.g., Green, 2002; Rickford & Rickford, 2000), Caribbean Creoles (e.g., Holm, 1984; Rickford, 1998), and literary representations of those speech communities (e.g., Buzelin & Winer, 2008; Troike, 2010). The corpus data confirms that the zero copula has a long history as an index of African diasporic speech. However, it is also a frequently used resource in representing Indian and Chinese voices, as one might guess from its high dispersion. In fact, it is fairly evenly dispersed across groups of speakers. This is also true of other highly dispersed morphosyntactic features. Of the ten features presented in Table 4.4, only four (null modal, *no* preverbal, null particle, and null subject) have significantly skewed distributions toward one group of speakers or another. This phenomenon is covered more extensively in the discussion of analysis of variance and in subsequent chapters, but the zero copula is our first indication of an important pattern: a constellation of morphosyntactic features that is commonly used in the construction of a generically racialized, nonstandard literary dialect. In other words, they are features that mark a boundary between imagined white and non-white speakers, but not among African diasporic, Chinese, and Indian speakers.

Figure 4.3: Bar plot showing the deviation of proportions for morphosyntactic features in all dialogue with DP ≤ 0.80 and color-coded by subcategory.



In addition to the zero copula, Table 4.4 contains five other verb-phrase-type features. In order to more clearly highlight the categorical breakdown of morphosyntactic features, Figure 4.3 presents those features where DP ≤ 0.80 (i.e., the same morphosyntactic features that are in Figure 4.1) color-coded by subcategory. In addition to the range of highly dispersed verb-phrase-type features, the chart underscores a number of other patterns. First, although pronoun features are the third most frequent morphosyntactic subcategory and although that subcategory is populated by 25 different types, only three have an even moderate dispersion: *me* as subject, *him* as subject, and *him* as possessive. These facts point to pronominal case paradigms as being an area with a few conventionalized features, but also one that is prone to idiosyncratic manipulation. Second, while the most dispersed morphosyntactic features are predominately located in the verb phrase, there are two highly dispersed noun-phrase-type features: the zero determiner and plural marking. The zero determiner is particularly intriguing both because of its high frequency and dispersion and because it is a feature that is not typically included in discussions of literary dialect.

4.3.3 Orthographic features

Table 4.5: Frequencies of orthographic subcategories for all speakers.

Feature	N	% Global	Freq.	DP
ORTHOGRAPHIC-TYPE				
TOTAL	**190**	**1.04%**	**3.71**	
eye dialect	99	0.54%	1.94	0.61
ambiguous	91	0.50%	1.78	0.62

Orthographic-type features are fairly evenly divided between the two different subcategories (see Table 4.5). Both subcategories also contain diverse sets of tokens. The eye dialect subcategory (§3.5.2) contains 61 different lemmatized types. The ambiguous subcategory, which is used to code respellings with unclear phonological salience, contains 52 different lemmatized types. The most common eye-dialect-type tokens are *pore* (for *poor*, n = 4), *troo* (for *true*, n = 3), and *sez* (for *says*, n = 3). The most common ambiguous-type tokens are *b'long* (for *belong*, n = 8), *p'rhaps* (for *perhaps*, n = 7), and *s'pose* (for *suppose*, n = 7). Of the ambiguous tokens, 58% follow the template illustrated by the three, most common ones. They contain at least one apostrophe that stands in for an elided vowel, but what phonological quality that proxy is meant to capture and how that quality is different from generic allegro speech (if, in fact, it is imagined as different at all) is difficult to determine. The other ambiguous tokens realize respellings like letter order changes (e.g., *littel* for *little*) that may be indicative of variant pronunciation, but like the elided vowels, what variation is being signaled is unclear.

The orthographic category as a whole is not particularly frequent. Orthographic features are nearly 20 times less frequent than lexical features and nearly 40 times less frequent than phonological and morphosyntactic features. Nonetheless, they have moderately high rates of dispersion, and they occur in the dialogue of all groups of speakers. In general, they serve to magnify the differences between standard and nonstandard dialogue and to encode caricatures of non-normative identities, as Preston (1985) suggests is typical of eye dialect. The following example is an excerpt from *Three Men on a Bummel* by Jerome K. Jerome (1900), with eye dialect in bold:

(2) Yes, sar, dat's what I'se **cumming** to. It **wuz** ver' late 'fore I left Massa Jordan's, an' den I **sez** ter mysel', **sez** I, now yer jest step out with yer best leg foremost, Ulysses, case yer gets into trouble wid de ole woman. Ver' talkative woman she is, sar, very –

The excerpt is from a short anecdote that describes an African diasporic man, Ulysses, testifying before a magistrate, having been charged with trying to steal a chicken from a deacon's poultry-yard. Ulysses continually eludes the magistrate's questions by straying from the topic at hand: his presence in the poultry-yard at midnight. The anecdote is not connected to the main narrative, but prefaces a chapter as a metaphor for the detour that the protagonists take in their travels and the digression that the narration is thus obliged to follow. Much of the anecdote is comprised of Ulysses' dialogue, and its supposed comedy is predicated on the contrast between the nonstandard voice of Ulysses and standard voice of the narrator. That contrast is not only advertised by an array of lexical, morphosyntactic, and phonological features, but also exaggerated by the use of eye dialect. The aggregation of features illustrates how eye dialect can work in concert with other features to amplify stereotypical effects, without necessarily having to occur in high frequencies.

4.3.4 Phonological features

Table 4.6: Frequencies of phonological subcategories for all speakers. Note that some rows have no deviation of proportions because those rows are for subcategories. Dispersion was calculated only for features.

Feature	N	% Global	Freq.	DP
PHONOLOGICAL-TYPE				
TOTAL	**7040**	**38.71%**	**137.63**	
consonant substitution	3957	21.76%	77.36	
insertion	1045	5.75%	20.43	
consonant deletion	837	4.60%	16.36	
vowel substitution	731	4.02%	14.29	
syllable deletion	370	2.03%	7.23	0.52
exaggerated	42	0.23%	0.82	0.85
metathesis	13	0.07%	0.25	0.88
doubling	45	0.25%	0.88	0.89

Phonological features closely follow morphosyntactic features in their frequency, occurring 137.63 times per 1000 words and accounting for 39% of the coded features. From one perspective, the salience of other feature types calls into question some more categorical claims like Blake's (1981, p. 15) suggestion that "[t]he *most* important aspect of non-standard language in literature is the use of spelling to suggest a deviant pronunciation" (emphasis mine). Frequencies and dispersions suggest that phonologically motivated respellings are certainly important, but not disproportionately so. There are, however, alternative ways of parsing the

data, ways that better support claims like Blake's that assert the unique salience of phonological features.

Table 4.7: Frequencies of phonological features for all speakers, where DP ≤ 0.80.

Feature	N	% Global	Freq.	DP
PHONOLOGICAL-TYPE				
t/*d*-for-*th*	2227	12.25%	43.54	0.45
syllable deletion	370	2.03%	7.23	0.52
cluster reduction	457	2.51%	8.93	0.58
b-for-*v*/*f*	654	3.60%	12.79	0.59
word-final deletion	149	0.82%	2.91	0.66
word-initial deletion	225	1.24%	4.40	0.67
a-for-*e*	90	0.49%	1.76	0.72
-*ee*/-*y*/-*i* final	760	4.18%	14.86	0.73
i-for-*e*	109	0.60%	2.13	0.76
n-for-*ng*	107	0.59%	2.09	0.77
u-for-*e*	47	0.26%	0.92	0.78

We will begin looking at those methods in the section on speaker variation (§4.4). As a brief primer to the category, let us first start as we have done with other categories: by examining the category's most dispersed features. Three of features included in Table 4.7 are of the consonant-substitution-type, and one of those (*t*/*d*-for-*th* substitution) is the most dispersed phonological feature. It appears in the dialogue of all speaker groups, though it is primarily associated with African diasporic vocal culture. This is a pattern that is shared by *b*-for-*v*/*f* substitution, as well.

The other consonant-substitution-type feature in Table 4.7 is *n*-for-*ng* substitution (or what is popularly called "*g* dropping"). In the late nineteenth century, the feature is associated with a number of English varieties in Britain, as well as in North America. A description of "the Anglo-Irish dialect" notes its "clipping of *g* in present participles" (Burke, 1896, p. 698). An 1891 article from *The Pall Mall Gazette* reporting on the Tranby Croft affair – a scandal in which a lieutenant in the Scotch Guard, Sir William Gordon-Cumming, was accused of cheating at baccarat – ascribes the feature to cockney. The writer of the article takes note of Gordon-Cumming's pronunciation of *owing* as *owin'*. "Having is English; Avin is Low Cockney; and Havin is High Cockney," the writer claims. In a lecture published after his death, George R. Kingdon (1895, p. 155), the Prefect of Studies first at Stonyhurst College and later at Beaumont, does not associate the feature with any particular community, but suggests that it is "dreadful" and "has a very vulgar and repulsive

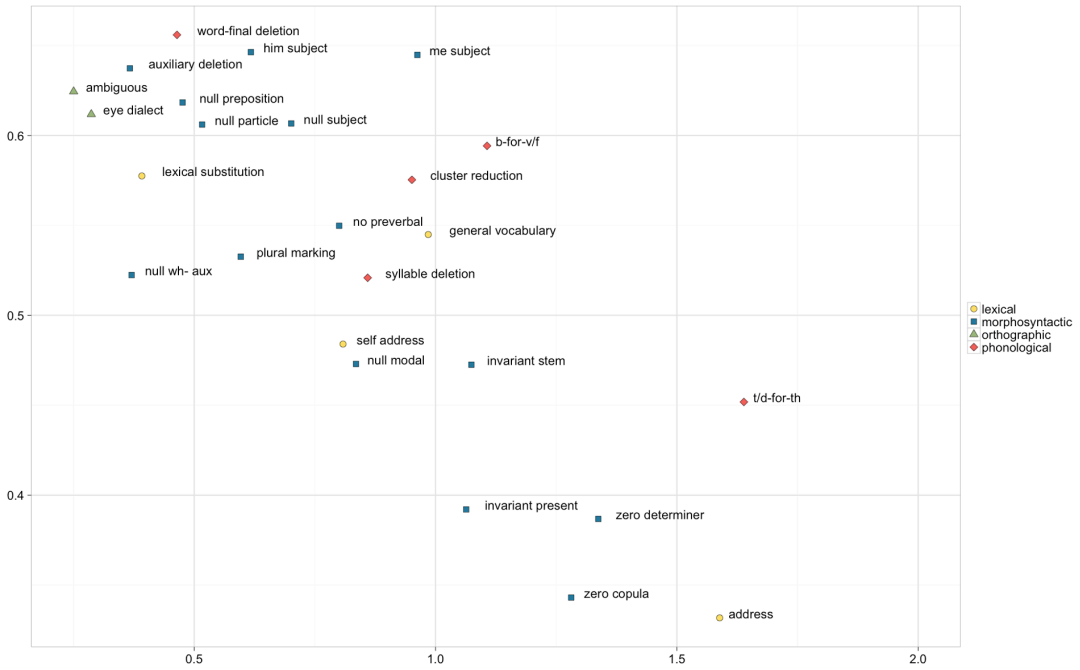
sound." An editorial in a United States education magazine echoes the lament: "but where is the cultivated American society where one may not hear people dropping their *g*'s" (Mowry, 1887, p. 290).

The first appearance of *n*-for-*ng* substitution in the corpus occurs in Charles Milner's *Don Juan*, which is published in 1837. It is only a single occurrence, and one that is perhaps suspect, as it is an outlier both in its form and its chronology. In the play, it appears in the form of *youn* (for *young*). This is the only *n*-for-*ng* substitution in the corpus that is not a respelling of word-final *-ing* as *-in*. It also appears early. The next occurrences are found in Dion Boucicault's *The Octoroon*, which is published more than two decades later. Boucicault's play is a more predictable adopter of the feature. It is published at a time when there is increasing discussion of the feature, as the above quotations suggest. The play is also set and first performed in North America at a time when the feature is gaining an association with "Yankee" speech. This is not to say that the feature does not circulate in literature prior to the mid-nineteenth century, or to imply that that it is used as an exclusively American index. Charles Dickens (1837) uses it in voicing the cockney Sam Weller in *The Pickwick Papers*, as does John Banim (1825) in his imagining of Hiberno-English accents in *Tales by the O'Hara Family*. As a feature of racialized literary dialect, however, it appears to be popularized by antebellum North American works like Harriet Beecher Stowe's *Uncle Tom's Cabin* – a novel, perhaps not coincidentally, from which Boucicault drew inspiration (Brody, 1998; Burkette, 2001). Also of note is the fact that the feature is entirely restricted to African diasporic dialogue in the corpus.

An important characteristic of the consonant substitution subcategory is that it is dominated by a small set of features. Only three (*t*/*d*-for-*th*, *b*-for-*v*/*f*, and *l*-for-*r*) account for 88% of occurrences, and six (including *n*-for-*ng*, *v*-for-*w*, and *f*-for-*th*) account for 94% of occurrences. From those substitutions, it may have been noted that *l*-for-*r* substitution is not among the most dispersed features listed in Table 4.7. Its absence points to a weakness when looking at dispersions across the entire corpus. In order to illustrate that weakness, let us consider two plots. The first is a scatter plot of the type that Gries advocates in his discussion of dispersion statistics. It takes the twenty-five most dispersed features and plots the deviation of proportions along the y-axis and the base 10 logarithm of the frequency along the x-axis (see Figure 4.4). The result is exactly what Gries describes: a downward trending pattern from the top left

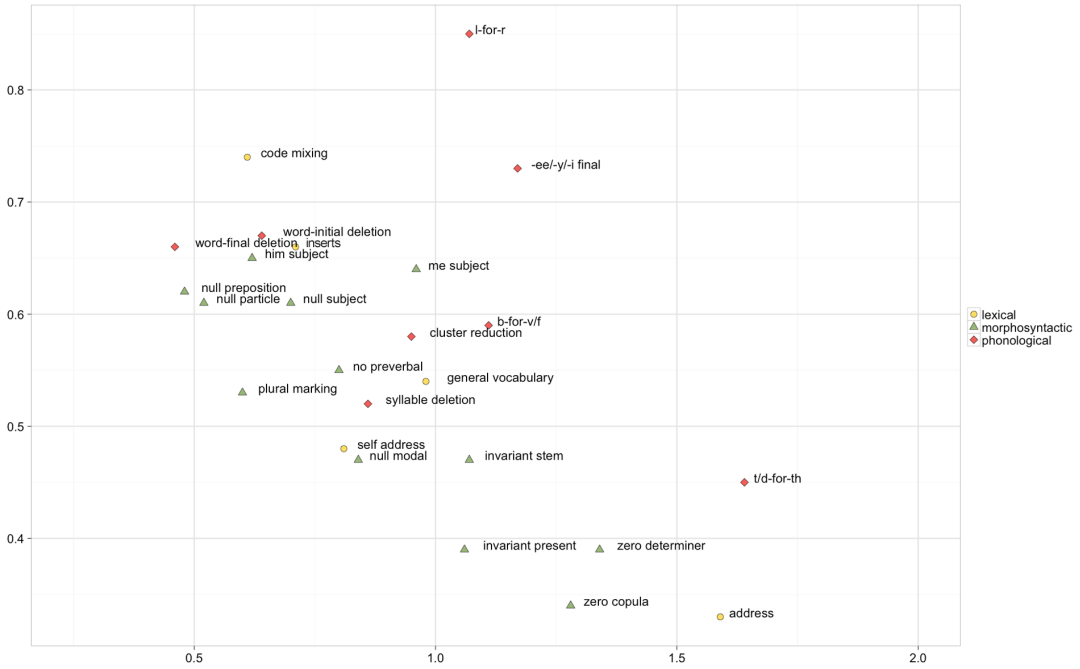
of the graph to the bottom right, a movement that tracks increasing frequency (left to right) and increasing dispersion (top to bottom). The upper edge of the pattern shows features with higher frequencies relative to their dispersions. Features like *t*/*d*-for-*th* substitution, *b*-for-*v*/*f* substitution, and *me* as subject have comparatively high frequencies in relation to their dispersions, reflecting their greater opportunity for realization, which was illustrated in (1). Features at the bottom edge of the pattern – features like null *wh*- auxiliary, lexical substitution, and eye dialect – instantiate the inverse. They have comparatively high dispersions relative to their frequencies.

Figure 4.4: Base 10 logarithms of frequencies (x-axis) plotted against deviation of proportions (y-axis) for the twenty-five most dispersed features.



Now, consider the same plot, but one that charts the twenty-five most frequent features rather than the twenty-five most dispersed (see Figure 4.5). The plot looks largely as expected, but with two clear outliers: word final -*ee*/-*y*/-*i* insertion and *l*-for-*r* substitution. For both features, their dispersions lag behind what we would expect based on their frequencies. The positioning of these two features is the result of the imbalance of the corpus. Because Chinese literary dialect practices emerge rather late in comparison to either African diasporic or Indian practices, Chinese dialogue is under-represented relative to the other groups. And both of these features are concentrated in and highly indexical of Chinese dialogue. That imbalance is addressed in future chapters, as each group of speakers is analyzed separately.

Figure 4.5: Base 10 logarithms of frequencies (x-axis) plotted against deviation of proportions (y-axis) for the twenty-five most frequent features.



Unlike consonant substitutions, vowel substitutions are not dominated by a small set of features. The most frequent feature, *i*-for-*e* substitution, makes up only 15% of the category and occurs only 2.13 times per 1000 words. The diffuse nature of the category suggests that vowel substitutions are not enregistered in the same way that other phonologically motivated respellings are (a claim that is borne out in the analysis of variance). The patterns governing vowel substitutions can be illustrated by looking at the two most dispersed types, *a*-for-*e* and *i*-for-*e*, in more detail.

The second most dispersed vowel substitution-type is *i*-for-*e* substitution. 74% of those substitutions occur in variants of YES: *iss*, *yis*, and *is*. This is a feature that Waters (2009, p. 85), for example, notes in the speech of Gus, a black plantation slave in the mid-eighteenth century play *The Staff of Diamonds* by Colin Hazelwood. Waters describes Gus's dialogue as having "all the markers of comic 'black' speech," which she argues is used to index his lack of intelligence, his treachery, and his suitability for servitude. The first linguistic marker that Waters lists is Gus's use of *iss* for *yes*. In this case, the frequency of the vowel substitution is tied to a specific word form that is used to encode stereotypes of African diasporic vocal culture.

The most dispersed vowel-substitution-type is *a*-for-*e* substitution. In contrast to the lexical restrictedness of *i*-for-*e* substitution, *a*-for-*e* substitution is restricted to particular orthographic environments. Its realization is largely linked to its relationship to *r* either as a pre-rhotic vowel or as the vowel indicating word-final

r-lessness. Pre-rhotically, the vowel is used in a respelling like *sarve* (for *serve*). Word-finally, the vowel is used in a respelling like *ansah* (for *answer*). Of the instances of *a*-for-*e* substitution, 64% are pre-rhotic and 17% are word-final.

4.4 Speaker variation

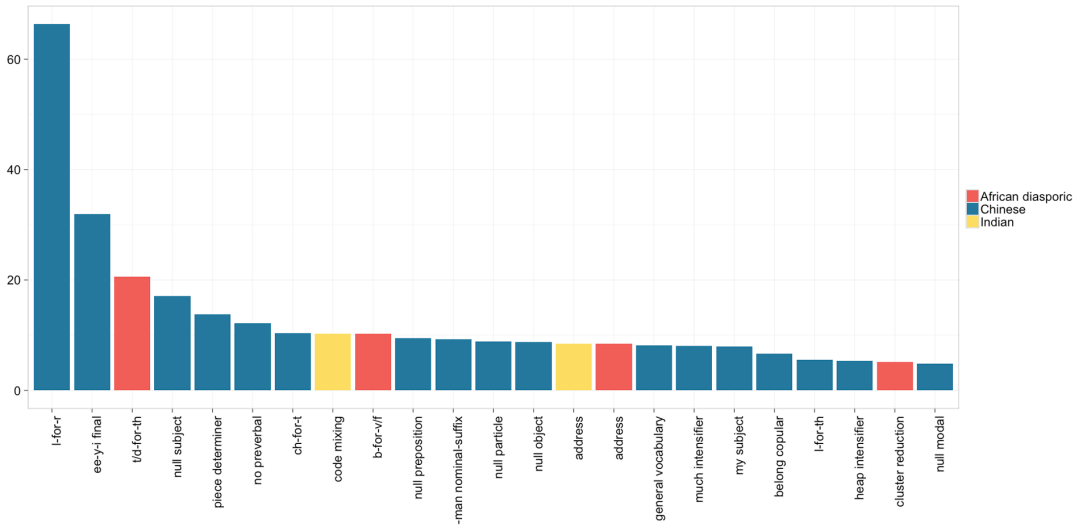
Feature data provides an interesting look at the broad contours of literary dialect. In order to add detail to those contours, this section begins the process of analyzing the ways in which literary dialect varies by speaker, over time, and across texts. This analysis is elaborated in the succeeding chapters both quantitatively and qualitatively, but this section frames some of patterns that I take up later. This initial framing is accomplished using a number of statistical methods. First, I examine composite frequencies (i.e., the total frequencies of coded features). Second, I apply diversity indices in a test of its usefulness as a measure of complexity (§4.2.2). Finally, I use hierarchical clustering, which is a more established technique in computational and corpus linguistics (see, e.g., Stefan Th. Gries & Hilpert, 2008; McMahon & Smith, 1996). Together these measures describe important and sometimes surprising patterns in the variation and change of literary dialect.

4.4.1 Analysis of variance

Analysis of variance (or ANOVA) comprises a suite of methods for examining the statistical relations among categorical and continuous variables by measuring how much variance in a data set is attributable to a category (or independent variable). For this study, ANOVA can help to explain fluctuations in the continuous variables (feature frequencies) in a couple of ways. First, ANOVA can sort out which coded features have variations that are due to speaker category and measure how significant those relationships are. Thus, it provides a quantitative account of which features accrue disproportionately to representations of which speakers. Such accounting bears on questions of indexicality and the distinctions between specific and generic literary dialect patterning – the kinds of questions that were raised in previous sections. Is, for example, the variation in the zero determiner distinguishable by speaker? If not, might the feature be considered generically nonstandard? Or given that *t*/*d*-for-*th* substitution occurs in the dialogue of all speakers, are claims that it is stereotypical of African diasporic speech qualitative and impressionistic? Or are occurrences in non-African diasporic dialogue statistical outliers? A second application of ANOVA

pertains to cluster analysis and unraveling the complex ways in which features are distributed in groupings of texts that are structurally similar. This second application is elaborated later in the discussion of cluster analysis.

Figure 4.6: F-values as determined by ANOVA for features with variations that are significantly attributable to speaker ($p < 0.01$).



The discussion here focuses on analysis of variance as it applies to speakers. The calculations involve a two-stage process. In the first, independent one-way ANOVA is calculated for the features, which generates F-values and p-values. Those values determine which features have variations affected by speaker and how significant those variations are. In order to establish how those variations are realized (whether they significantly differentiate African diasporic from Chinese dialogue, for example), a second calculation is needed. This second calculation is a post-hoc Tukey test. Figure 4.6 combines both of these measures. It shows the F-values for those features where $p < 0.01$. It also breaks out the features according their distributions based on a Tukey test. For example, *l*-for-*r* substitution is color-coded for Chinese dialogue because it significantly distinguishes Chinese from both African diasporic and Indian dialogue. Address, by contrast, is color-coded for both the African diasporic and Indian dialogue because it distinguishes African diasporic and Indian dialogue from Chinese dialogue, but not African diasporic and Indian dialogue from each other. Thus, the chart suggests which features significantly differentiate the three speaker groups.

The chart contextualizes a number of features that have been introduced already and ones that will be further explored in later chapters. For example, the three features with the highest F-values (*l*-for-*r* substitution, word-final *-ee*/*-y*/*-i*, and *t*/*d*-

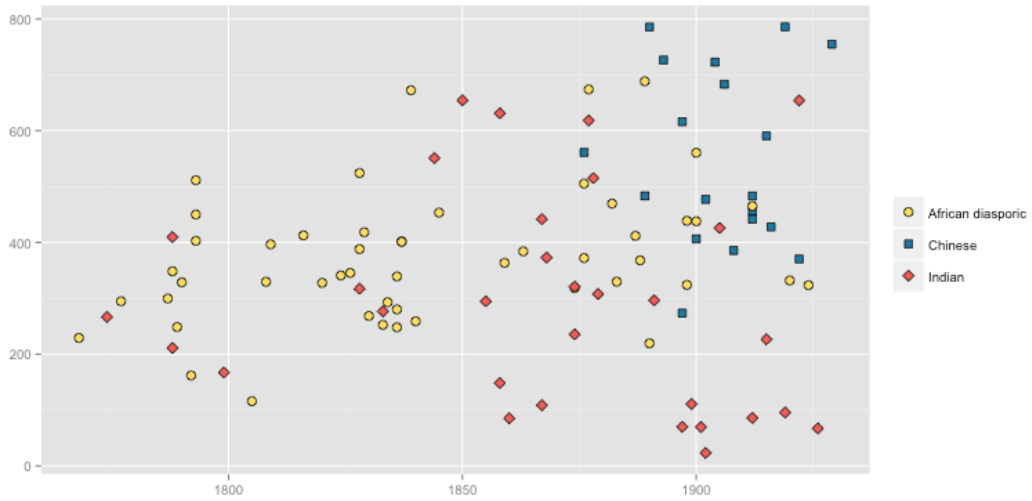
for-*th* substitution) are all phonological features. Their significance gives some support to Blake's insistence on the primacy of phonology in literary dialect, an assertion that was somewhat undercut by the dispersion measures. The chart also suggests the importance of lexical marking through code-mixing and address in Indian dialogue, but an otherwise lack of distinctive features.

Alternatively, Chinese dialogue has a relatively large number of distinguishing features. Aside from their quantity, one of their notable shared characteristics is that many of them (*piece* as a determiner, *-man* as a nominal suffix, *much* as an intensifier, *belong* as a copular verb, and *heap* as an intensifier) are lexicogrammatical (i.e., they function at the intersection of lexicon and syntax.) In the discussion of methods, I noted that one of the difficulties in assigning codes involved drawing just these kinds of distinctions. What features should get a separate category and what should be subsumed under a more expansive category like general vocabulary? While these results do not provide anything close to a definitive answer, they do underscore some of the implications of those choices. For one, breaking them out into separate categories highlights the importance of lexicogrammar in Chinese dialogue, a pattern that might have been more difficult to tease out if the features had been coded more generically (§3.5.1).

4.4.2 Composite frequencies

One method of approaching the ways in which different authors represent different speakers is to look at the composite frequencies of coded features. The rate at which literary dialect features appear in dialogue provides a gauge for the markedness of that dialogue. In other words, the greater the frequency of coded features, the more an author distances the speaker's voice from an imagined standard. It is important to note that "distance" is a statistical designation not an indexical one. The latter may depend on the specific features used to ventriloquize a character, as well as extra-linguistic factors like the ways in which a character is described. With that caveat in mind, we can see clear variation in literary dialect by plotting composite frequencies against the date of publication and controlling for speaker (see Figure 4.7).

Figure 4.7: A scatter plot showing the normalized composite frequencies (the y-axis) over time (the x-axis) of dialect features for texts with a minimum of 95 words. The texts are color-coded for speaker.

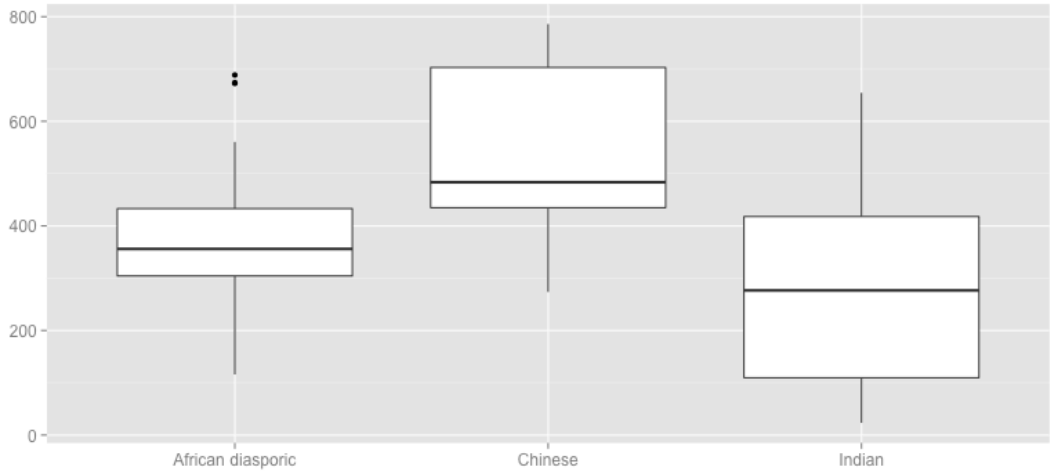


One pattern that the scatter plot reveals is the increasing range of frequencies starting in the mid-nineteenth century. In the late eighteenth and even more so in the early nineteenth centuries, the frequencies are relatively clustered between roughly 250 and 400 features. Beginning in the mid-1840s, the range of frequencies spreads out dramatically. Part of that increased range is due to emerging conventions of representing Chinese speakers. Five of the highest frequencies are found in representations of Chinese speakers. These are denoted by the blue squares in the upper-right quadrant of the plot. Conversely, nine of the lowest frequencies also appear after the mid-1840s and are found in representations of Indian speakers. These are denoted by the red diamonds in the lower-right quadrant.

The groupings of higher and lower frequencies after the mid-1840s suggest another important pattern. In addition to the frequencies associated with Indian speakers near or below 100 features, there are representations of Indian speakers with frequencies above 600 features. By contrast, the lowest frequency associated with Chinese speakers is just below 300 features. Variations in the frequencies used to voice different groups of speakers can be illustrated by differences in how those frequencies are distributed using box plots (see Figure 4.8). The plots show that representations of Indian speakers have the lowest median but the highest interquartile range. In other words, while representations of Indian voices may be more standard-like on average, those representations also display far more variance. There are, I

think, reasons for this, which are related to changes in Britain's role in India and contradictory racial logics – issues that I explore in detail in chapter 7.

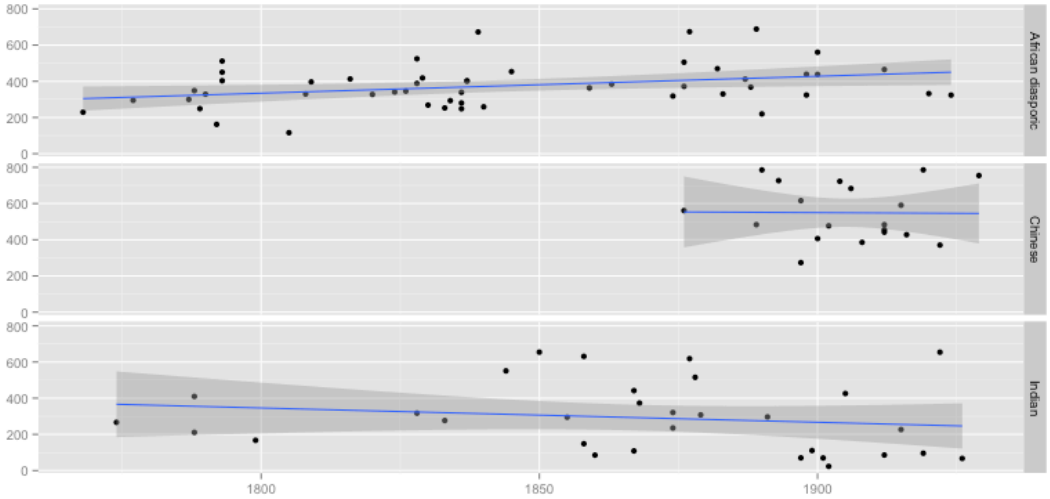
Figure 4.8: Box plots of composite feature frequencies by speaker.



In contrast to the relatively wide-ranging frequencies linked to Indian voices, the frequencies related to African diasporic speech have the lowest interquartile range and are thus the most consistently represented, at least in terms of feature frequency. Their median is also higher than in the voicings of Indian characters, indicating a greater degree of differentiation from an imagined standard. The feature frequencies associated with Chinese voices have the highest median, however, suggesting that those voices are even more distanced from the standard. As is the case with representations of Indian speakers, I argue in subsequent chapters that behind these numbers lie attitudes circulated in colonial discourse: logics of race, ideas of empire, competition and conflict with China, and the rise of sinophobia.

Finally, changes over time in composite frequencies can be modeled using regression analysis. Figure 4.9 takes the same data from Figure 4.6, breaks it out by speaker, and adds trend lines and confidence intervals. The first thing to note about each trend is its slope. The trend line for African diasporic dialogue has an upward slope (β = 0.94). For Chinese dialogue the trend line is essentially flat (β = -0.16), and for Indian dialogue, the trend line has a downward slope. (β = -0.78). These slopes appear to show that African diasporic dialogue tends to realize higher frequencies over time, Indian dialogue fewer, and Chinese dialogue about the same. However, Chinese dialogue enters the corpus later, as is clear from the chart.

Figure 4.9: A plot showing the trend lines for composite frequencies.



The question, then, is how explanatory these trends are. The data are clearly noisy and this is reflected in their coefficients of determination. African diasporic dialogue has the most robust r-squared value ($r^2 = 0.11$, $p < 0.05$). The r-squared values for Indian ($r^2 = 0.03$, $p > 0.1$) and Chinese dialogue ($r^2 = 0.00$, $p > 0.1$), however, are less so. For Chinese dialogue in particular, these results are unsurprising. Because the conventions of Chinese literary dialect develop later, the data is sparser. Also, as the slope of a trend line approaches zero (as the one for Chinese dialogue does), its r-squared will approach zero.

The r-squared for Indian dialogue is potentially more troublesome. Before acknowledging that there is no discernible, defensible trend, however, we have a variety of alternative approaches and tools at our disposal for explicating diachronic changes. Frequencies can be separated into component categories, for example, to see if those categories exhibit more plausible trajectories. Also, there are other models, linear (e.g., segmented regression, quantile regression) and nonlinear (e.g., generalized additive models), that can be applied in an effort to produce more explanatory descriptions. In future chapters, these approaches and tools are explored in detail.

4.4.3 Diversity indices

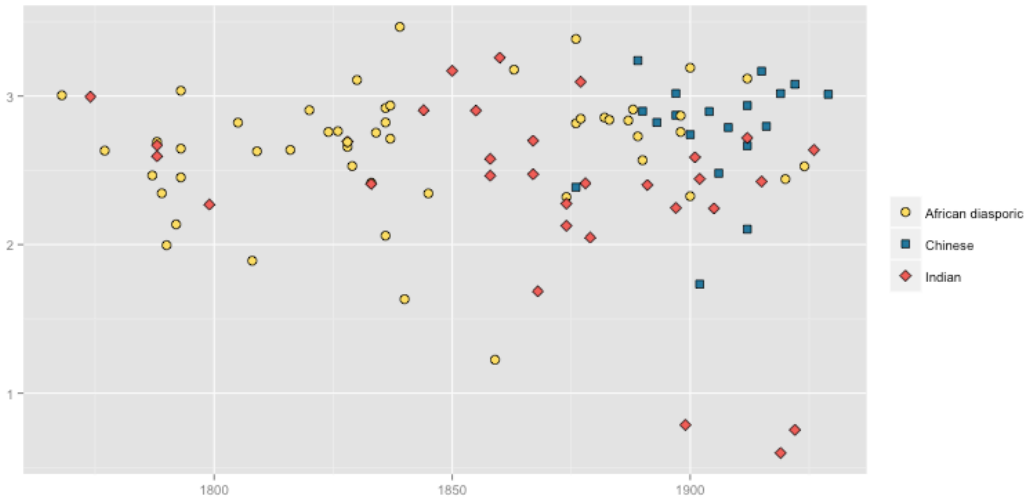
Composite frequencies are a useful measure for illustrating the density of features that authors use in voicing characters, and that density tells us something about how authors distinguish different voices. Frequencies, however, elide other kinds of information that is of potential interest. They do not capture, for example, the

breadth of features that a representation incorporates. Neither do they capture how distributed features are. Consider, for example, the text with the highest composite frequency, John C. Hutcheson's *Afloat at Last* (1890). In its rendering of the voice of a Chinese cook, Ching Wang, the literary dialect realizes 41 different features. This is well above the average of 28.83 different features for the texts with more than 95 words. However, it is also well below the 72 different features that Hutcheson (1889) uses in the dialogue of Sam Jedfood in *The Black Man's Ghost*. Moreover, in Ching Wang's dialogue, a fewer number of features are statistically dominant. For example, *-ee*/*-y*/*-i*-final insertions account for 28% of all features, a prevalence that is evident in the following excerpt:

(3) And dis one manee you tellee Ching Wang cocky-fightee one piecee -- hi?

In fact, among the texts with more than 95 words, the only one in which *-ee*/*-y*/*-i*-final insertions make up a larger percentage (a remarkable 44 %) is Nesbit's (1904) *New Treasure Seekers*. As this brief example demonstrates, composite frequencies provide important information about the density of features, but they tell us little about other kinds of measures related to the complexity of literary dialect.

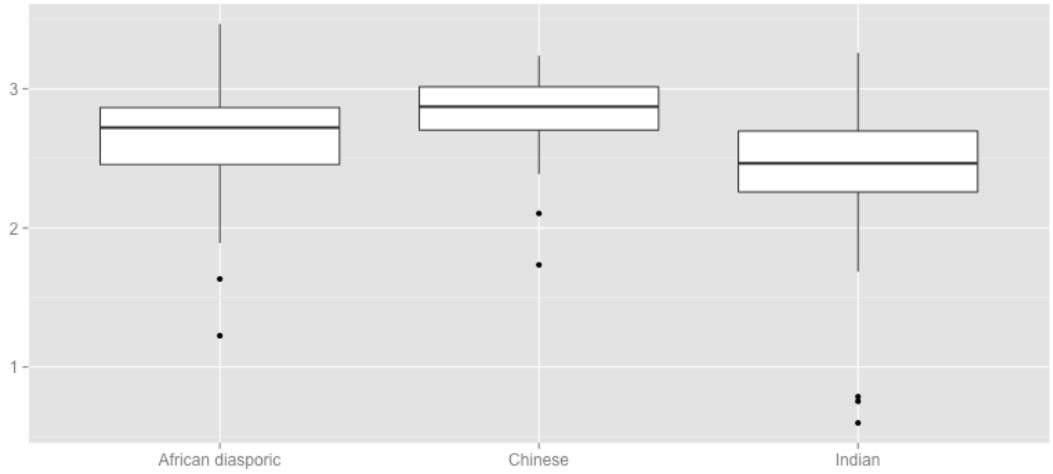
Figure 4.10: A scatter plot showing the diversity indices for texts with a minimum of 95 words. The texts are color-coded by speaker.



It is here, therefore, that we turn to Shannon's diversity index as a measure of literary dialect complexity. Just as with composite frequencies, the diversity indices of each text can be plotted by year of publication. The results reveal some interesting intersections in the modeling of complexity versus the modeling of density (see Figure 4.10). First, the chart reinforces some patterns that are present in Figure 4.7. In

particular, it shows a cluster of Indian speech representations in the lower-right quadrant.

Figure 4.11: Box plots of diversity indices by speaker.



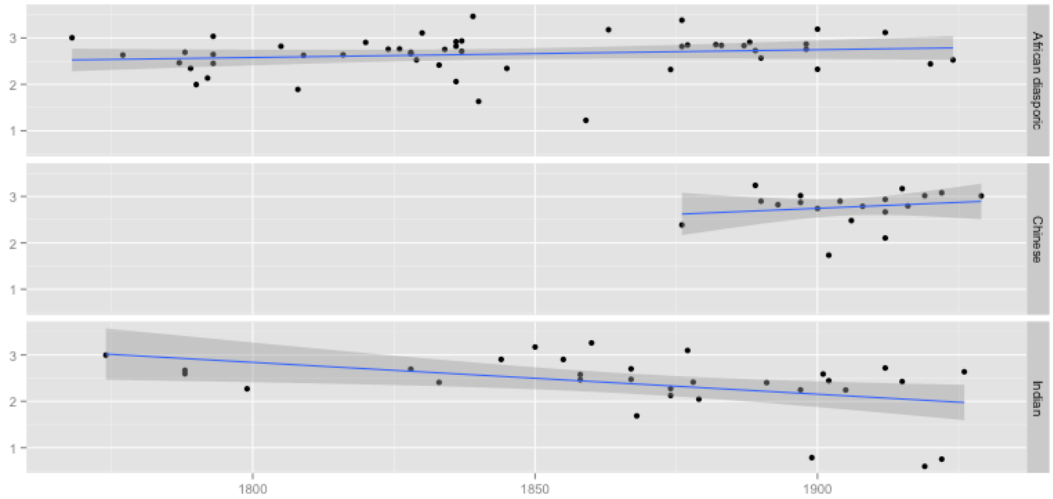
There are other similarities, too. Box plots of diversity indices show the medians and interquartile ranges for the different groups and in many respects mirror those in Figure 4.8. Representations of Indian speakers have the lowest median and the highest interquartile range; representations of Chinese speakers have the highest median. These similarities again confirm previous claims based on composite frequencies: that Indian voicings, on the whole, are the most standard-like and Chinese voicings the least; also that Indian voicings evidence the greatest variation.

While much of the data show parallels between the frequency and diversity of features, there remain a few important differences. For one, the composite frequencies of Chinese dialogue have an interquartile range that falls between that of Indian and African diasporic dialogue. The diversity indices of Chinese dialogue, however, have the lowest interquartile range. In other words, when compared to the representations of African diasporic features, representations of Chinese speakers show greater disparity in the rates at which literary dialect features occur, but less in the variety of feature types.

The parallels between composite frequencies and diversity indices also extend to their diachronic trends – although, as with the box plots, there are a few important differences, as well (see Figure 4.12). For African diasporic dialogue, the trend line in diversity indices has an upward slope ($\beta = 0.002$), as it did in composite frequencies. However, the r-squared value is a little less robust ($r^2 = 0.07$, $p < 0.1$). For Indian

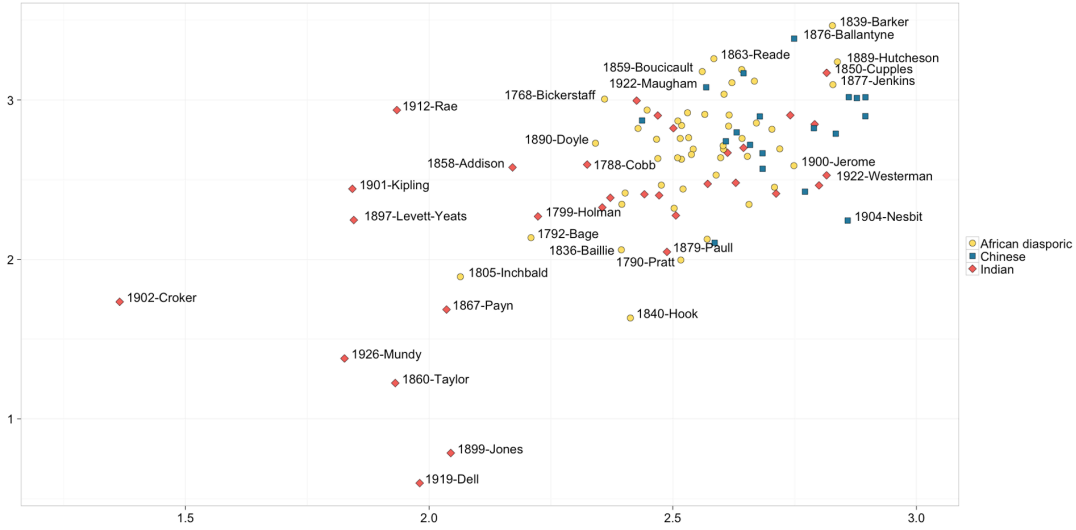
dialogue, the trends in the two measures also run in similar directions. Again, the data reveals a downward slope (β = -0.006), and the r-squared is a little more convincing this time ($r^2 = 0.18$, $p < 0.05$). For Chinese dialogue, there is a rising ($\beta = 0.005$), as opposed to a flat trajectory. As was the case with the composite frequencies, however, that trend does not seem particularly explanatory ($r^2 = 0.03$, $p > 0.1$).

Figure 4.12: A plot showing the trend lines for diversity indices.



Based on the combined regression analyses from both composite frequencies and diversity indices, some patterns become evident, though claims about those patterns must be taken as tentative and provisional. First, Indian dialogue appears to become more standard-like over time. By contrast, African diasporic dialogue appears to become less so. What is perhaps even more surprising than its change in frequency is the change in diversity indices. We might expect that as representations circulate, features emerge and fossilize into shibboleths. Those shibboleths, then, become distilled markers for vocal cultures. Thus, over time, we might expect diversity indices to decline as writers increasingly rely on extant stereotypes. Yet, the diversity indices for African diasporic representations increase. In addition to using features more frequently, writers also use a greater range of features in rendering African diasporic voices.

Figure 4.13: Base 10 logarithms of composite frequencies (x-axis) plotted against diversity indices (y-axis).



As a final look at the frequency and diversity data, let us consider a less conventional visualization. This one applies the same principles that were used in generating Figures 4.4 and 4.5. Those plots show the relationship between dispersion and frequency for features. Figure 4.13 adopts their principles in order to shed light on the relationship between diversity and frequency for texts. The figure plots the base 10 logarithms of composite frequencies along the x-axis against diversity indices along the y-axis. As we would expect, the plot is a mirror image of dispersion versus frequency (since diversity measures move from low to high rather than from high to low). As we move rightward along the x-axis and texts' frequencies increase, diversity indices tend to correspondingly increase upward along the y-axis.

Most of the texts aggregate into groupings where diversity is greater than 2 and the logarithm of the frequency is greater than 2.25. At the lower end of the spectrum, however, nine texts with Indian dialogue and one with African diasporic dialogue (Elizabeth Inchbald's *To Marry or Not to Marry*) are scattered vertically near the line where the logarithm of the frequency is equal to 2.0. These are texts with similar frequencies, but very different diversity indices. For example, Milne Rae's *A Bottle in the Smoke* and Tom Taylor's *Up at the Hills* have very similar composite frequencies (85.88 versus 85.23). However, Rae's novel has a diversity index of 2.94 (accounting for a range of 30 different features) and Taylor's play a diversity index of 1.22 (accounting for a range of only 16 different features). Taylor uses a few features more often, and Rae more features less often. That these texts contain primarily

Indian dialogue fills out the picture established by earlier plots. Not only does Indian dialogue exhibit the greatest range of composite frequencies and diversity indices, as the box plots indicate, but it also exhibits a high range of diversity indices at the lower end of the frequency range.

The upper quadrant of the plot is equally interesting. Although many of the texts with the highest frequencies contain Chinese dialogue, the clustering of texts in the upper right quadrant with high composite frequencies and high diversity indices is a more varied grouping. It contains the African diasporic dialogue of Matthew Barker's (1839) *Hamilton King* and Dion Boucicault's (1859) *The Octoroon*, the Chinese dialogue of Robert M. Ballantyne's (1876) *Under the Waves*, and the Indian dialogue of George Cupple's (1850) *The Green Hand*. Partly driving these differences are authors' attempts at imitating specific, regional varieties. In Matthew Barker's *Hamilton King*, for example, Quaco is voiced in an imitation of Caribbean Creole. His dialogue realizes phonological features like word-final *-a* (in, for example, *wharra*), as well as lexical features like *buckra* that appear in other source works that are from the same period and are set in the Caribbean like *Marly* (Anonymous, 1828).

Similarly, all of the voices in *The Octoroon*, not just the African diasporic ones, are elements of Dion Boucicault's self-professed verisimilitude. In a letter to *The Times* after the play's London premier, he asserts his purpose in writing the play was to "faithfully" depict slave society based on his "long residence in the Southern States of America" and his "every facility for observation" (Boucicault, 1861a). In a separate, unpublished note he wrote that same year, he affirms his intention to render "American homes, American scenery, and manners without either exaggeration or prejudice" (Boucicault, 1861b). "Scenery," certainly, is visual – the staging of a slave market and the elaborate recreation of a riverboat are commented on frequently in the British press after the play's opening at the Adelphi – but it is also vocal. The success of the simulacrum depends as much on how it sounds as how it looks. One reviewer observes positively, "The people dress, act, and talk very much as Southern Americans really do act and talk" (Townsend & Hutton, 1861).

The Southern American acoustic landscape of *The Octoroon*, which the British reviewer finds convincing, is comprised of diverse voices including the standard speech of the protagonist Zoe (4), the Yankee speech of Salem Scudder (5), the Native American speech of Wahnotee, which is characterized as "a mash up of Indian, French, and 'Merican" (6), and the African diasporic speech of Pete (7):

(4) And our mother, she, who from infancy treated me with such fondness, she who, as you said, had most reason to spurn me, can she forget what I am? Will she gladly see you wedded to the child of her husband's slave?

(5) Job had none of them critters on his plantation, else he'd never ha' stood through so many chapters. Well, that has come out clear, ain't it?

(6) Paul wunce – Paul pangeuk.

(7) It's dem black trash, Mas'r George; dis ere property wants claring – dem's getting too numerous round; when I gets time, I'll kill some on 'em, sure!

In attempting to evoke intersections of geography and race, Boucicault creates both text-internal and text-external patterns of differentiation. Text-internally, Pete's voice has shared features with Scudder's: they both use *ain't* as a form of *to be*, they both use demonstrative *them*, and they both realize consonant deletions. These overlaps encode not only geography, but also class. Although a "good" overseer, Scudder is an unsuccessful businessman. Thus, in spite of his love for Zoe, his voice announces that economically he is not a suitable match. Notwithstanding these overlaps, Scudder's dialogue and Pete's dialogue are clearly delineated. Pete's voice realizes features that are conventionally linked to African diasporic identities like *t*/*d*-for-*th* substitution and forms of address. These distinctions map sound onto complexion and form elements in the play's racial logic – a logic that finds slaves like Pete to be, as one reviewer puts it, "of a lively and cheerful disposition, attached to their homes and masters, [and] endowed with strong sympathy for the white man in his hour of need" (J. V. P., 1861, p. 52), but finds injustice in the constraints placed on the freedom and romantic fulfillment of the titular, standard-speaking character, Zoe.

In addition to being differentiated text-internally, Pete's voice is differentiated text-externally. While his dialogue contains conventional markers of African diasporic identity, it also manifests features that are far less dispersed like the *a*-for-*ea* substitution in *claring* (which occurs in only 7% of the texts in the corpus) and the first-person singular *-s* in *gets* (which also occurs in 7% of the texts). These less common features figure a distinctiveness and a semblance of authenticity that British audiences apparently found believable. One reviewer writing under the pseudonym "Old Footlights" (1861, p. 77) takes issue with actors' imitations of "Yankee" accents, while extolling their minstrelsy:

(8) Several performers, appearing for the first time as Americans, struggled manfully with the "down-east" pronunciation; but, very now and then they seemed to be haunted with some recollections of a stage-countryman's dialect and to think that

"One touch of *Yorkshire* makes the world his."

It is almost needless to say that it does nothing of the kind.

The author, however, goes on to praise the representations of "negro" characters as "so true to life in their talk as to be almost unintelligible to English hearers." In another review, the author singles out the performance of the American actor George Jamieson,[10] who played Pete in blackface for the British premiere:

(9) Mr. Jamison's portraiture of the aged negro, *Pete*, is thoroughly life-like, though at times his illustration of the indistinct articulation peculiar to the race is a trifle too real. (J. V. P., 1861, p. 53)

Nearly four years later in a review of Henry Thorton Craven's *One Tree Hill*, the critic recalls Jamison's performance in his assessment of James Stoyle's blackface portrayal of Dick White:

(10) His negro, like Mr. Jamison's Pete in the "Octoroon," is so natural that many of his speeches are almost incomprehensible. (Mackay, 1865, p. 431)

Of course, there is more going on in these reviews than acknowledgments of naturalism. Equations of acoustic "truth" with unintelligibility and incomprehensibility reinforce the plays' racial logics and encode asymmetries of power. These are all vitally important issues that are explored in the following chapters.

Finally, regional imitation may help to explain the positioning of some texts like *The Octoroon*, but it does not shed much light on others like *Hamilton King* or *Under the Waves*. Barker's and Ballantyne's novels are among texts with high diversity indices, high composite frequencies, or both that are adventures – and most of those are specifically nautical adventures. High frequencies and diverse ranges of features arise in this genre, first, because they often focus on themes of voyage and discovery and in so doing trade in the exoticism of non-European peoples. In evoking exoticized voices, authors tend to use more features more often. Second, ships are commonly imagined as sites of cultural, racial, and linguistic contact. Thus, they often present a panoply of voices – European (Irish, Scots, German, etc.) and non-European (African diasporic, Chinese, Indian, etc.). In some cases, such contact and resulting variation is distilled within a single voice. The dialogue of the African diasporic character Mephistopheles in Frederick Marryat's novel *Mr. Midshipman Easy*, for

[10] George Jamieson was a well-known impersonator and blackface performer. In his memoirs, one of his fellow actors claimed, "I never knew a more perfect 'chameleon'" (Leman, 1886, p. 181). Another nineteenth century account described him as "one of the best impersonators of the negro ever seen on the dramatic stage" (T. A. Brown, 1870, p. 193). Jamieson also gained notoriety for his connection to the divorce of actor Edwin Forrest and the singer Catherine Sinclair, a case that was widely reported on and followed at the time.

example, is described in a review as a transnational and transracial comic amalgam: “This character is admirably worked out, and his phraseology is a most laughable farrago of negro English, diversified by occasional Yankeeisms, and rendered poetical and impressive by a sprinkling of those vehement expletives and peculiar modes of speech by the natives of the Emerald Isle are supposed to be distinguished” ("Literature," 1836).

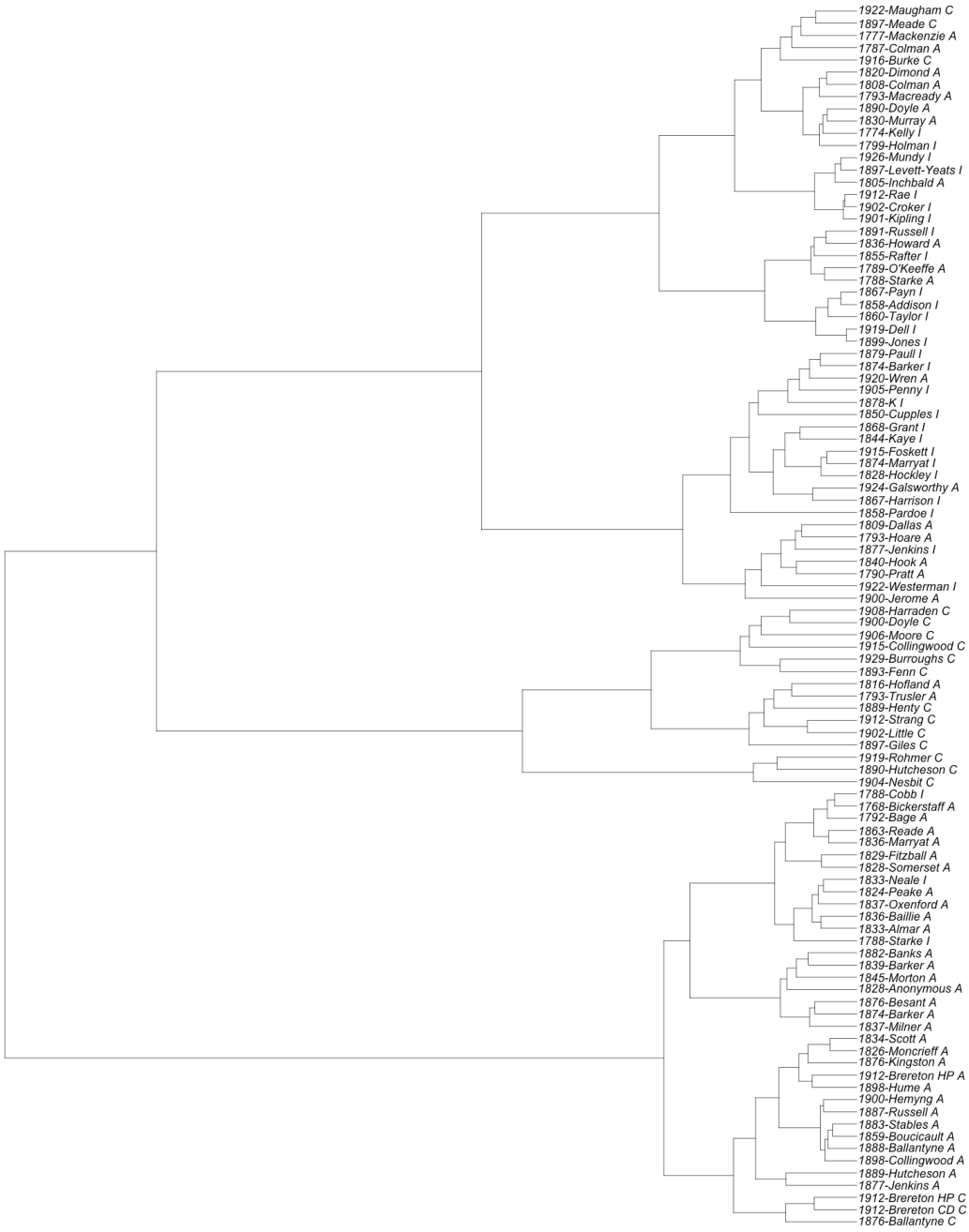
4.4.4 Cluster analysis

In addition to abundances and diversity indices, the third and final statistical technique that I want to introduce in this chapter is hierarchical cluster analysis. Clustering describes a number of statistical methods that group objects according their shared attributes, such that the objects in one group or cluster are more alike than those in another. They are, therefore, methods for reducing the complexity of multivariate data and locating patterns of similarity. Hierarchical clustering, as the name suggests, builds clusters according to a multilevel hierarchy or cluster tree (see, e.g., Kaufman & Rousseeuw, 2005; Murtagh, 1983).

The hierarchical clusters were calculated using the *APE* package (Paradis, Claude, & Strimmer, 2004) according to the Ward’s method.[11] The results appear in Figure 4.14. The terminus of each branch of the tree structure (or leaf) represents a text in the corpus. The arrangement of the branches (or clades) tells us which texts are the most similar to each other based on the coding data. Texts that are paired at the lowest level of structure are the most alike. The lengths of the clades tell us how alike the texts are. A short clade means that the texts are very similar; a longer clade means that there is less similarity. Thus, the dendrogram produces a picture of scaled resemblances. As we move up the tree structure (or to the left in the rotated version in Figure 4.13), we get larger and larger groupings, but also less and less similarity.

[11] Ward’s method is agglomerative, meaning that it builds clusters from the bottom up. Pairs of clusters are merged at each step based on the smallest increase in the sums of squares. It is a commonly used technique and is the one that Griffiths et al. recommend for document classification (1984). For a further explanation of different clustering methods, see, for example, Willett (1988).

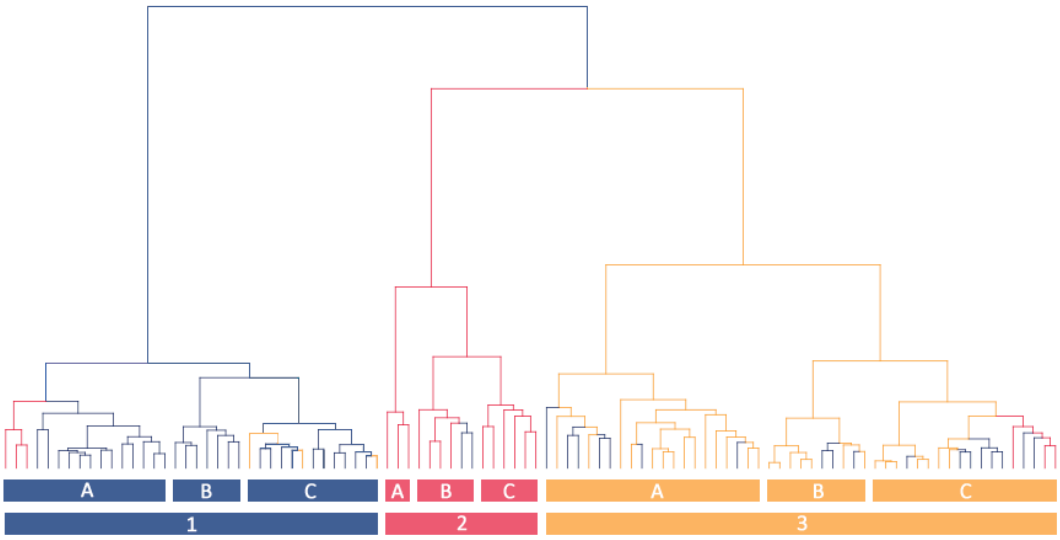
Figure 4.14: A dendrogram showing the hierarchical clusters for texts with a minimum of 95 words. The leaves show the year of the publication, the author, and the speaker category.



To provide a more global view of clustering arrangements, Figure 4.15 is color-coded for speaker. What that coding reveals are patterns that are central to the analysis going forward. First, it shows that the dendrogram can be cut into three clusters that broadly align by speaker. Cluster 1 is primarily made up of African diasporic dialogue; cluster 2, Chinese dialogue; and cluster 3, Indian dialogue. This result is at least somewhat expected. It simply confirms that there are conventional ways of representing groups of speakers in literary dialect. The dendrogram provides

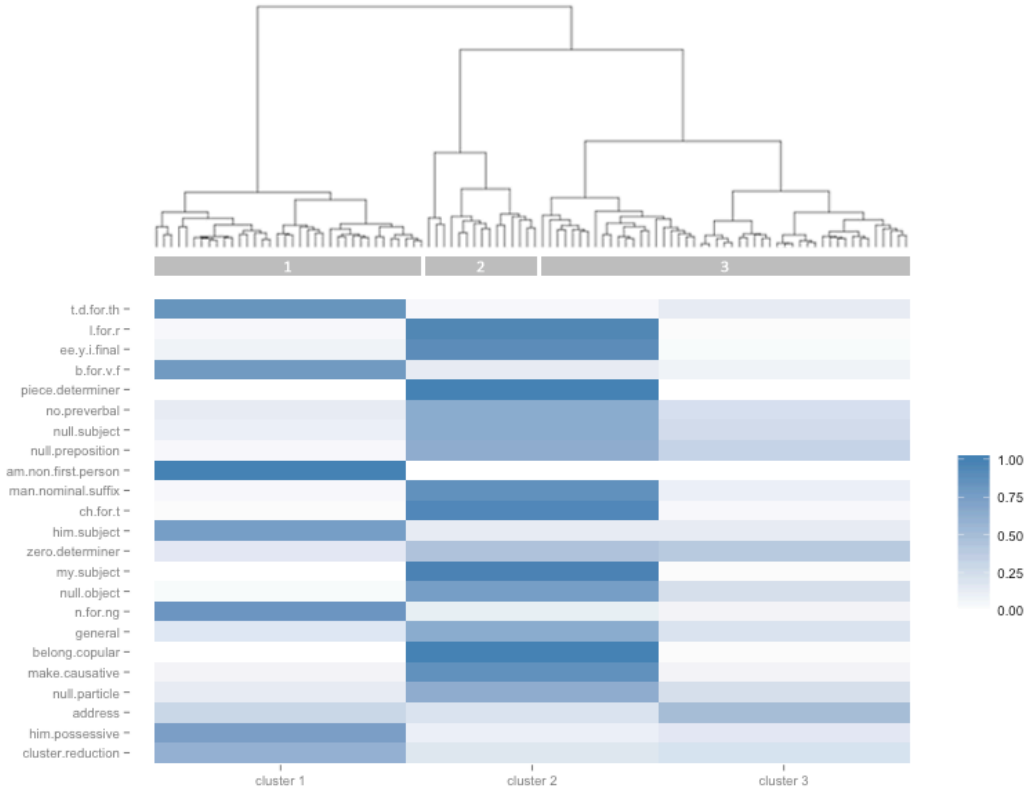
further evidence that those conventions are measurable. The first indication of this was the analysis of variance, which demonstrated that there are statistically significant differences in the distributions of some features according to speaker. The cluster analysis shows that the coded features in combination create what Jockers (2013) terms a "signal" – a constellation of measures that enable a statistical classifier to categorize texts (by author, gender, nationality, etc.). We could say, then, that there is a "speaker signal."

Figure 4.15: A dendrogram cut into three clusters and color-coded by speaker (blue for African diasporic, red for Chinese, and gold for Indian).



That signal appears the strongest for African diasporic and Chinese dialogue, as both clusters 1 and 2 exhibit a fairly high degree of consistency. The signal for Indian dialogue appears weaker. Visualizing how those signals form into clusters brings us to the second application of analysis of variance. Figure 4.16 is a heat map showing the weighted means of features in each cluster as shaded blocks (such that a darker blue block indicates a higher mean). The features included on the heat map were selected by ANOVA, so only features with significant distributions by cluster ($p < 0.01$) are present. The features are arranged by F-value. Thus, *t*/*d*-for-*th* substitution has the highest F-value and by implication is the feature with the most significant distribution by cluster.

Figure 4.16: A heat map showing the weighted mean frequencies for three clusters, with features determined by ANOVA and arranged by F-value.

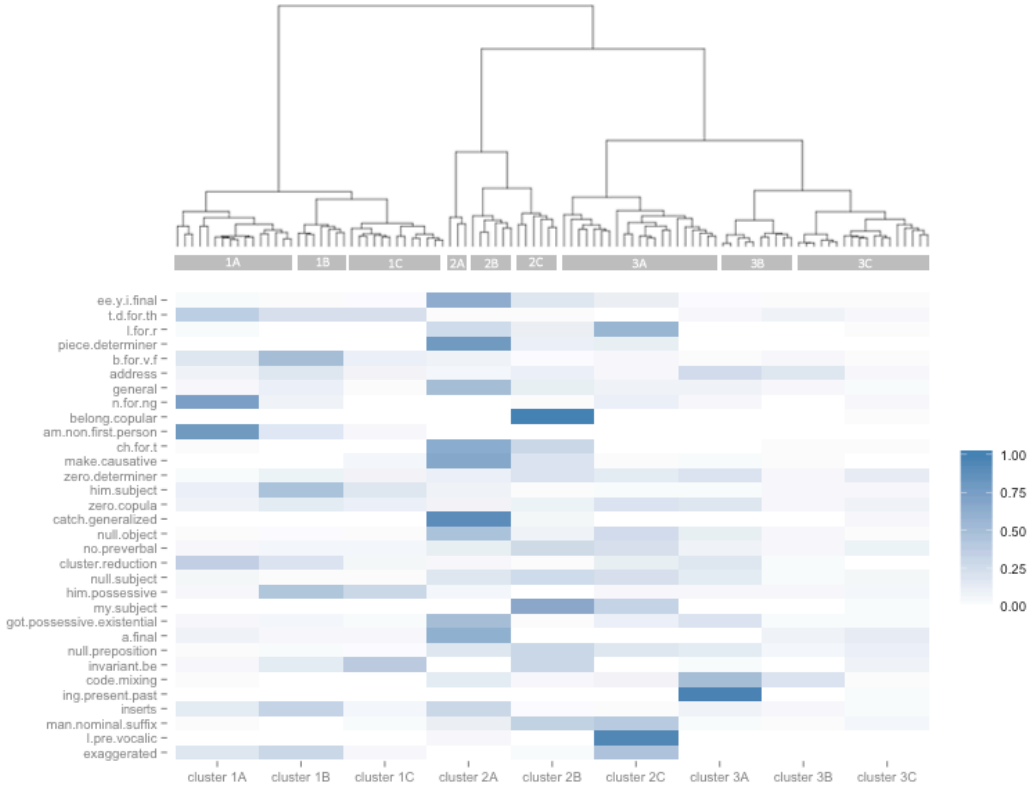


The heat map throws into relief patterns of co-occurrence. The pattern is particularly dense in cluster 2, somewhat less dense in cluster 1, and sparsest in cluster 3. Most of the means are heavily weighted toward one cluster, though a few, mostly morphosyntactic features, have means that are nearly balanced across two. Any substantial overlap generally occurs between clusters 2 and 3 or clusters 1 and 3. Clusters 1 and 2 exhibit little commonality.

This picture can be elaborated more fully by dividing the dendrogram further. Each cluster (1, 2, and 3) can be separated into three additional sub-clusters (A, B, and C). The same techniques that were used to produce the heat map for three clusters can be applied to generate a detailed look at variation across nine clusters. The resulting plot illustrates how the frequencies of features can fluctuate within the larger clusters depicted in Figure 4.17. For example, *belong* as a copular verb and generalized *catch*, both of which appear almost exclusively in cluster 2, actually tend to occur in distinct sub-clusters – *belong* in cluster 2B and *catch* in cluster 2A. Similarly, Figure 4.15 shows that *t*/*d*-for-*th* substitution is heavily concentrated in cluster 1. Figure 4.16, however, reveals that the highest frequencies are located in

cluster 1A. Additionally, those higher frequencies co-occur with higher frequencies of other features like *n*-for-*ng* substitution, *am* as a non-first-person verb, and cluster reduction. As we will see in future chapters, these kinds of variations in both frequency and co-occurrence have implications not only for the identities that authors imagine for speakers, but also for changes over time. The texts in cluster 1A are largely published later and those in cluster 1C earlier, for instance. These clusters, therefore, appear to align with the rising trends that were illustrated in the scatter plots for African diasporic dialogue (see Figures 4.8 and 4.11) and are a tool that can help to explicate those trends.

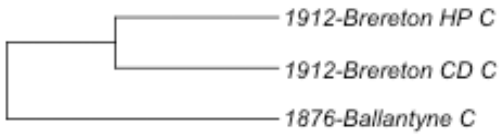
Figure 4.17: A heat map showing the weighted mean frequencies for nine clusters, with features determined by ANOVA and arranged by F-value.



The overview of cluster analysis concludes by zeroing in on a couple of specific groupings. The first is a trifoliate grouping of Chinese dialogue that appears in cluster 1A (see Figure 4.18). The grouping includes dialogue from two source works by Frederick Brereton (*The Hero of Panama* and *Under the Chinese Dragon*) and one by Robert Ballantyne (*Under the Waves*). The purpose in singling out this cluster is simple. Given that the Chinese dialogue from *The Hero of Panama* and *Under the Chinese Dragon* is both by the same author and from works published in the same year, we would expect the texts to cluster together. Thus, the pairing is

evidence in support of the overall approach. Beyond this relatively straightforward observation, the grouping poses other, more difficult questions. Why does this grouping appear in cluster 1 rather than in cluster 2, with the majority of Chinese dialogue? And what does Ballantyne's dialogue have in common with Brereton's that makes them aligned, but also outliers?

Figure 4.18: A trifoliate grouping from cluster 1A.



These more complicated questions are addressed in chapter 7, but they speak to another, important way that cluster analysis can direct and inform the examination of literary dialect. As much as points of similarity can provide avenues for discovery and investigation, so too can points of difference and discontinuity. As an example of how this process can work, consider the pentafoliate grouping pictured in Figure 4.19, which is located in cluster 1C. The grouping is remarkably consistent in its time period, containing texts from the 1820s and 1830s. It also contains only representations of African diasporic speech with the exception of the Indian dialogue in William Neale's *The Port Admiral*.

Figure 4.19: A pentafoliate grouping from cluster 1C.

What makes this alignment particularly interesting is that at least one commentator at the time complained vigorously about Neale's use of literary dialect in voicing Indian characters. The commentator specifically takes issue with Neale's use of *massa*, declaring that "we never heard the said '*Massa*' in those regions [India], except by the stray negroes who might be found here and there in the ships" (Colburn, 1833). While that particular lexical choice is clearly marked, the clustering indicates that Neale ventriloquized his Indian characters in accordance with patterns of African diasporic representation that were circulating at the time – lexical, morphosyntactic, orthographic, and phonological patterns independent of *massa* as a shibboleth. That alignment suggests a conflation of African diasporic and Indian identities that further

implicates Neale's voicings in racial ideologies and logics of complexion that, as the commentator's complaint attests, were contested at a time when Britain's role in India was changing.

4.5 Conclusion

Because of the substantial amount of data presented in this section, I want to use the space here to briefly synthesize some of the key results before proceeding to discussion of the quantitative data's relationship to other kinds of discourse that circulates in the imperial archive. This chapter has presented methods that measure literary dialect along three dimensions: 1) frequency, 2) diversity, and 3) similarity. Using those measures, in terms of speaker variation, I argue the following:

- **African diasporic**: African diasporic representations exhibit the lowest variation in their composite frequencies. They also are generally uniform in their clustering. In these ways, they appear highly conventionalized. According to ANOVA, that conventionality seems to be particularly shaped by three phonological features (*t*/*d*-for-*th* substitution, *b*-for-*v*/*f* substitution, and cluster reduction) and one lexical feature (address). That conventionality, however, does not appear entirely stable, as composite frequencies and diversity indices increase over time. Thus, over the course of the nineteenth century and into the twentieth, the imaginings of African diasporic voices become increasingly differentiated from an imagined standard. These trends produce a relatively robust r-squared value for frequencies, but a less convincing one for diversity indices. Determining whether or not those apparent changes are meaningful, therefore, requires further analysis.
- **Indian**: Indian representations have the lowest median composite frequency and diversity index. However, they also have the highest interquartile range. Thus, although they are the least differentiated from an imagined standard in some ways, they also realize the greatest amount of variation. The variation in Indian dialogue is further evidenced in its clustering pattern, which is the least consistent of the three. That inconsistency accords with the analysis of variance. The ANOVA results show only two features (code-mixing and address) that significantly distinguish Indian dialogue. When applied to the clusters, ANOVA also shows lower mean frequencies across most of the relevant features in cluster 3, where most of the Indian dialogue is located. Indian representations also appear to demonstrate changes over time. Both their composite frequencies and diversity indices decline, indicating that they become less differentiated from an imagined standard. However, the r-squared values produced by those trends are the inverse of those for African diasporic dialogue: the r-squared for frequencies is low, but the one for diversity indices is higher. As with diachronic trends in African diasporic dialogue, therefore, these demand further investigation.

- **Chinese**: Chinese representations have the highest average composite frequency and the highest average diversity index. By these measures, Chinese voices are the most differentiated from an imagined standard. Even more so than African diasporic representations, Chinese representations cluster relatively consistently. That uniformity results from the considerable set of features that mark Chinese dialogue. ANOVA shows that two phonological features (*l*-or-*r* substitution and word-final *-ee*/*-y*/*-i* insertion) accrue most significantly to Chinese dialogue, but there are also a number of morphosyntactic features (*piece* as a determiner, *-man* as a nominal suffix, *much* as an intensifier, *belong* as a copular verb, and *heap* as an intensifier) that have statistically significant distributions. However, the heat map demonstrates that these morphosyntactic features are not spread equally across all representations, but rather co-occur in distinct constellations. Finally, the regression analysis produces the least satisfactory results for Chinese dialogue. Contributing to the low r-squared values for both frequencies and diversity indices is the smaller window covered by the study for Chinese dialogue. Because its conventions in fiction emerge later in nineteenth century, there is simply less available data. In light of that fact, the analysis of Chinese dialogue will take an alternative tack. While the analysis of African diasporic and Indian dialogue will partly focus on questions of change, the analysis of Chinese dialogue will focus on questions of emergence.

The subsequent chapters follow the sequence set out above: African diasporic dialogue is analyzed first, followed by Indian dialogue, and concluding with Chinese dialogue. African diasporic dialogue is a logical starting point as it is the earliest data in the corpus and the most robust. Indian literary dialect is present in the corpus almost as long as African diasporic literary dialect. Additionally, the potentially contrasting trajectories in Indian and African diasporic dialogue suggested by Figures 4.9 and 4.12 make for a compelling juxtaposition. Finally, Chinese literary dialect is the outlier both in its relative consistency and in the time of its appearance. Its story highlights some of the evolving conditions of the empire at the turn of the century, as well as changes in print culture that reshape how texts are produced and consumed.

Chapter 5
Imagining African Diasporic Voices

5.1 Introduction

Nineteenth century philology, according to Foucault (1971, p. 281), constituted a radical change:

> (1) [T]he isolation of the Indo-European languages, the constitution of a comparative grammar, the study of inflections, the formulation of the laws of vowel gradation and consonantal changes – in short, the whole body of philological work accomplished by Grimm, Schlegel, Rask, and Bopp, has remained on the fringes of our historical awareness, as though it had merely provided the basis for a somewhat lateral and esoteric discipline – as though, in fact, it was not the whole mode of being of language (and of our own language) that had been modified through it.

In altering the way language was understood, philology, Foucault argues, sparked an epistemological and social crisis, calling into question the foundations of knowledge and the established social order. The old taxonomies separating barbarous and civilized tongues were questioned, and the forces of language change were located outside of conscious human control, "for language is neither an instrument nor a product" Foucault (1971, p. 290) contends, "but a ceaseless activity – an energeia." Concomitantly, Foucault notes that philology engendered a new acoustic awareness. Variations in sound, rather than spelling, were theorized as one of the fundamental building blocks of language, and the spoken word, not the written, was positioned as language's most essential expression.

In concert with philology's rise, there emerged a literary movement that took up philology's enthusiasms, advancing the phonetic variation of regional speech as an aesthetic and democratic force. This "cult of the vernacular," as Jones (1999, p. 39) refers to it, is usually framed as a North American phenomenon, a product of postbellum anxieties and social upheavals in the United States. Even so, many of these works had British editions and received wide attention in England – sometimes for better, sometimes for worse. An article in *The Bristol Mercury and Daily Post* (1881), for example, praises *The Uncle Remus* stories of Joel Chandler Harris as "[o]ne of the most entertaining works published for some time in the United States." The reviewer, however, concedes that the "phonetic reproduction" might present some difficulty for British readers: "English people can hardly be expected to know, for example, that 'might have just' is pronounced 'mouter des.'" A review printed in *The Graphic* (1881) complains even more bitterly about George Washington Cable's

use of literary dialect in "Madame Dauphine," opining that "the Creole *patois* grates so upon English ears." Regardless of their critical receptions, these works and their progenitors, like *Uncle Tom's Cabin*, circulate evolving literary dialect conventions transatlantically.

These material and aesthetic conditions are relevant to the analysis of African diasporic literary dialect because they coincide with important changes in the corpus. In the previous chapter, we saw how the overall frequencies of coded features increase in African diasporic dialogue over time. We also saw a similar increase in the diversity of features. In this chapter, I will show that those increases are driven by changes in the phonological category. What is causing these changes is, of course, a complicated question. The developing acoustic awareness that Foucault argues for and the concomitant rise of the "cult of the vernacular" are only pieces of the puzzle. The propagation of those texts and the ideas they embody are abetted by widening global networks of circulation and influence. Moreover, in the middle of the nineteenth century, the figuring of African diasporic voices is happening within the context of rancorous debates about abolition and legislative efforts to limit the slave trade, a context epitomized by two watershed events: the passage of the Slavery Abolition Act in Britain and the Civil War in America. It is during this period of ideological tumult and material transformation that phonological marking in African diasporic dialogue takes on new salience.

The chapter is divided into three main subsections. The first examines the constituent structure of African diasporic dialogue in terms of features and feature categories (§5.2). The second explores diachronic trends in African diasporic dialogue (§5.3), and the third resemblances in African diasporic dialogue (§5.4). Each of these sections addresses, in turn, the first three subsidiary research questions set forth in the first chapter (§1.3). The fourth and final question is addressed throughout the chapter in the discussions of the patterns, trends, and resemblances that are identified through the computational analysis.

5.2 Constituents of African diasporic dialogue

Of the three types of dialogue in the corpus, African diasporic dialogue has the most robust data: 26,541 words from 60 texts. It contains dialogue from the earliest source work in the corpus, *The Padlock*, which debuted in 1768, and is relatively balanced across the three periods, 1768-1829, 1830-1879, and 1880-1929 (§3.4). The

analysis uses the measures and techniques described and carried out in the statistical overview. These include: deviation of proportions, which is a dispersion measure (§4.2.1); diversity indices, which is used as a measure of complexity (§4.2.2); regression analysis to model changes over time (§4.4.2 and §4.4.3); and hierarchical cluster analysis to model resemblances (§4.4.4). In addition to the computational techniques listed above, the analysis also includes the first application of log-likelihood comparisons. Log-likelihood tests statistical significance, and, in corpus linguistics, are used to compare observed versus expected frequencies (Oakes, 1998, p. 42). Typically, such comparisons are done for token frequencies in keyword analysis. Here, log-likelihood tests are applied to the frequencies of coded features and feature categories, and, in this chapter, they are used specifically in time-period comparisons. As in the statistical overview, the discussion begins with an examination of how features and feature categories are distributed in the sub-corpus.

Table 5.1: Frequencies of the four superordinate categories in African diasporic dialogue. *N* is the raw number of occurrences; *% Global* is the percentage a feature or category contributes to all coded features; and *Freq.* is the normalized frequency of a feature or category (per 1000 words).

Feature	N	% Global	Freq.
FEATURES-TYPE			
TOTAL	**10110**		**380.92**
lexical	1847	18.27%	69.59
morphosyntactic	3531	34.93%	133.04
orthographic	133	1.32%	5.01
phonological	4599	45.49%	173.28

In African diasporic dialogue, features belonging to the phonological category are the most frequent, comprising just less than half (45%) of all literary dialect features. As was discussed in the previous chapter, comparing frequencies across categories can be tricky, since there can be greater opportunities for features like consonant substitutions than for some lexical or morphosyntactic features. That said, as a category, phonological features are arguably the most salient indexes of African diasporic literary dialect. The ANOVA data from the statistical overview (§4.4.2) provided some evidence of this. Among the four features for which there is a statistically significant F-value ($p < 0.01$) and for which African diasporic dialogue has the highest frequency, three are phonological: *t*/*d*-for-*th* substitution, *b*-for-*v*/*f* substitution, and cluster reduction (see Figure 4.6). Among those, *t*/*d*-for-*th* substitution has the third highest F-value overall. These are indicators of the central

role phonological marking plays in the conventions of representing African diasporic voices. As important as a subset of phonological features is in distinguishing African diasporic from other literary dialects, other measures reveal the wide range of features that are deployed in ventriloquizing African diasporic characters, even if those features are not uniquely indexical.

Figure 5.1: Chart showing the deviation of proportions for features in African diasporic dialogue with DP < 0.80 and color-coded by category.

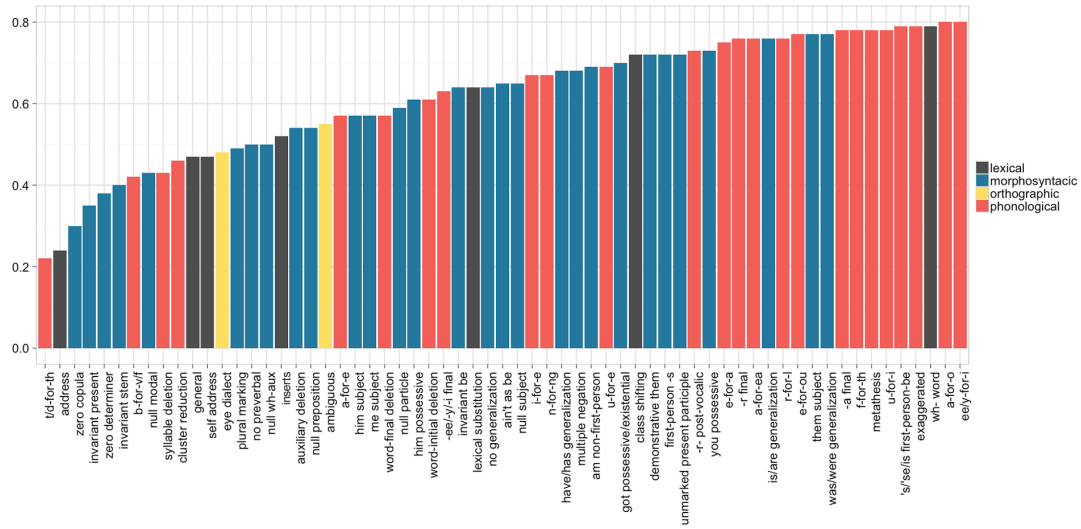


Figure 5.1, for example, shows the deviation of proportions (§4.2.1) for features in African diasporic dialogue (where DP < 0.80). They are arranged from the lowest DP (or the most dispersed) to the highest (the least dispersed) and are color-coded by superordinate category. While the chart reinforces the salience of the phonological category, it also highlights the diversity of features across the four superordinate categories. It contains 7 lexical features (70% of the features in category), 29 morphosyntactic features (34% of the category), 24 phonological features (19% of the category), and both of the orthographic features. The most dispersed feature is *t*/*d*-for-*th* substitution, reinforcing the significance indicated by the ANOVA data. The second-most dispersed feature is address. In fact, three additional lexical features are among those with DP < 0.60: general vocabulary, self-address, and inserts.

In addition to being the second-most dispersed feature overall, address is the most frequent lexical feature, occurring 43.63 times per 1000 words (see Table 5.2). Address appears in African diasporic dialogue in primarily three forms: *massa* (or the variant *massah*), which makes up 57% of the coded features, variants of *sir* (e.g., *sah*, *sar*, *sa*), which make up 19%, and variants of *missy* (e.g., *missee*, *missie*, *missey*),

which make up 7%. Other realizations tend to be work-specific or idiosyncratic (i.e., *masser*, *mass'*, *mas'r*, *missah*). As the numbers suggest, address is an important constituent of African diasporic literary dialect, in general, and *massa* is particularly indexical. It is used to signal the subordinate position of African diasporic characters within a social hierarchy. Put into the mouths of African diasporic characters, it indicates an imagined consciousness that cheerfully accepts its own subservience. This may seem obvious in works where African diasporic characters are rendered in racist, comic stereotype. The feature's indexical potential, however, may be even more remarkable in works that attempt to resist or undermine those stereotypes.

Table 5.2: Frequencies of lexical features in African diasporic dialogue, where DP < 0.80.

Feature	N	% Global	Freq.	DP
LEXICAL-TYPE				
TOTAL	**1847**	**18.27%**	**69.59**	
address	1158	11.45%	43.63	0.24
general vocabulary	201	1.99%	7.57	0.47
self-address	130	1.29%	4.90	0.47
inserts	215	2.13%	8.10	0.52
lexical substitution	59	0.58%	2.22	0.64
class shifting	21	0.21%	0.79	0.72
wh- word	11	0.11%	0.41	0.79

Consider Mariana Starke's (1788) *The Sword of Peace*. Though set in India and exploring themes related to changes in British colonial rule in the late eighteenth century, the play also contains an abolitionist subplot. Caesar, a slave, has his freedom purchased by Jefferys, a servant to two sisters, Eliza and Louisa Moreton, who have travelled to India seeking their fortune. Upon informing Caesar that he is free, Jeffreys maintains that "you're my friend and my equal" and promises to make Caesar "a lad of spirit, like an Englishman." Caesar's response to being freed is to pledge his devotion to Jeffreys:

(2) Friend oh vil you vite man be so kind to call poor black friend? de black mans he fight for his friend – bleed for his friend – die for him – starve for him – every ting for his friend. – But oh, Massa, I must call you Massa; for me feel, me love you like my old Massa.

On the one hand, Starke takes an abolitionist stance, imagining a system of subjugation as reformed. On the other, the underlying order is upheld. In insisting on addressing Jeffreys as "Massa," Caesar signals his position in that racialized order. He may be liberated from the legal apparatuses of institutional slavery, but he willingly and enthusiastically reaffirms his servitude.

Nussbaum (2004, p. 164) argues that Starke's drama figures England as offering the appearance of liberty to subjugated peoples like Caesar, "though the fact that his blackness may limit that freedom goes unmentioned." I would push Nussbaum's point even further. It is not merely possible that Caesar's blackness may constrain his selfhood and his agency. By having him replace one "Massa" with another, Starke defines Caesar's Englishness by his blackness. His status as a free man may change, but his status as a servant remains constant.

Other lexical features can function similarly in encoding African diasporic subjectivities as deficient, deviant, or otherwise rationalizing their own subjugation. In the statistical overview, I argued that self-address can function to displace a speaker's subjectivity. Signification of oneself does not take the form *I* or *me* (as it does in standard dialogue), but instead occurs non-pronominally. In some occurrences, characters address themselves by their own name (*Smutta*, *Negombo*, *Snowball*, *Cuffee*, *Quashie*). In others, the self-address underscores the character's racial identity (*black man*, *blacky boy*, *bad neger*, *negro maid*, *black dog*). One dominant pattern is the appearance of *poor* with a character's name or a racial identifier in self-address (*poor Cubba*, *poor Wowski*, *poor negro man*, *poor black*, *poor nigger*, *poor servant*, *poor black rascal*, *pore niggah*). Phrases that include *poor* make up 36% of self-address features in African diasporic dialogue.

Inserts are another of the more widely dispersed lexical features. The most common insert in African diasporic dialogue is *golly* (an amelioration of *God*), which accounts for 14% of the category. In the nineteenth century, the interjection appears to have an association with African diasporic speakers. That association is articulated in a story by Henry Hesketh Bell (1897), "His Highness Prince Kwakoo." Originally published in *The Idler Magazine,* the story chronicles the British adventures of an African conman. Bell describes main character's language as the "real English such as the missionaries […] spoke," as opposed to "the Christy-Minstrel sort that darkies are popularly supposed to interlard with 'Gollies' and guffaws." The association is evidenced further by distributions of *golly* in the source works, where the token occurs only in African diasporic dialogue. Other common inserts include variants of religiously related interjections (*garamercie*, *garamighty*, *gor amighty*, *goramity*, *lud a mercy*, *laws a massey*, *bress de lor'*). As the quotation from Bell suggests, these inserts often encode minstrelic exaggeration and buffoonery. The relationship

between comic stereotypes and inserts is apparent, for example, in the literary dialect of Sambo, who appears in George Cupples' (1850) *The Green Hand*:

(3) 'Golly!' chuckled the nigger, rolling the whites of his eyes and grinning like mad; 'oh sar, Misser Barton! dis 'ere shark riglar navligator! I 'clare to you, sar, um got chr'ometer aboard. Oh gum; berry much t'ink dis you own lost silber tickler, Misser Barton.'

Perhaps the most extreme expression of using inserts to signify the perceived deviance of African diasporic vocal culture occurs in Walter Besant's (1876a) *The Case of Mr. Lucraft*. Boule-de-neige, an African diasporic servant, literally clucks ("Cluck – cluck! Massa not angry with poor old Boule-de-neige"). Associating African diasporic voices with animal sounds has a long history that predates Besant. More than two hundred years before Besant, Edward Terry (1655, p. 16), in his travelogue, describes the language of the inhabitants of what is now Table Bay, South Africa as "inarticulate noise, rather than language," which he analogizes to "the clucking of hens, or gabling of turkeys."

Table 5.3: Frequencies of morphosyntactic subcategories in African diasporic dialogue.

Feature	N	% Global	Freq.
MORPHOSYNTACTIC-TYPE			
TOTAL	**3531**	**34.93%**	**133.04**
pronoun	778	7.70%	29.31
noun phrase	545	5.39%	20.53
verb phrase	1869	18.49%	70.42
adjective-adverb	32	0.32%	1.21
negation	195	1.93%	7.35
complementation	25	0.25%	0.94
discourse organization	87	0.86%	3.28

In addition to highlighting the salience of lexical features, Figure 5.1 makes clear the range of morphosyntactic features that have a relatively high dispersion in African diasporic dialogue. The highest frequencies are those related to the verb phrase (which accounts for 18% of morphosyntactic features) and those related to pronouns (which account for 8% of the category). The quantitative data aligns with historical accounts that stigmatize variants related to pronouns, as well as those related to verbs, their inflexions, agreement, and aspect. An interesting point of entry for an examination of this stigmatization is the debate surrounding the British and Foreign Bible Society's translation of the New Testament into "Negro-English" for circulation in Surinam (1829). The debate is noteworthy both because it occurred during the period surrounding the passage of Slavery Abolition Act in 1833 and

because it generated a good deal of discussion. The translation itself was undertaken to convert speakers of the Surinam creole – what is now known as Sranan and was then popularly called "talkee-talkee" – because it was mutually unintelligible with its related European languages English, Dutch, and Portuguese. Excerpt 4 comes from the Gospel of John:

(4) Bikasi Gado ben lobbi ala soema so, tee a gi da wan lobbi Pikien vo hem abra; vo ala soema, disi de bribi na hem, no moe go lasi, ma vo dem habi da liebi vo teego. Bikasi Gado no been seni hem Pikien kom na grontapo, vo a moe kroetoe kondre, ma vo a meki ala soema zieli fini helpi.

For God so loved the world that he gave his one and only Son that whoever believes in him shall not perish but have eternal life. For God did not send his son into the world to condemn the world, but to save the world through him.

The ensuing backlash in domestic Britain was intense. In a typical critique, the author derides the creole as more "lingo than a language" (Lockhart, 1830, p. 555). The author goes on to characterize the variety as follows:

(5) The language of the slaves in our sugar islands is as intelligible, when introduced in books, to English readers, as that of Mungo in the farce, and more so than the Scotch dialogues in Sir Walter Scott's novels. Any one might speak it, if he made himself acquainted with some half score words of foreign extraction which are most in use; all that he has else to do is to liquefy his English, speak straightforward, in contempt of case, number, mood, and tense, and throw grammar to the dogs.

The creole is thus presented as unstructured, as "liquefied," and that lack of structure is evident in its "contempt" of pronominal (case and number) and verbal (mood and tense) paradigms. In addition to the emphasis on features related to pronouns and verbs, it is worth noting that the passage references *The Padlock* (which the author calls "the farce"). The reference conflates the imagined dialect of Mungo with the language of the real-world speech community in Surinam. They are described as equally "unintelligible" – though, of course, the lack of mutual intelligibility between English and the creole was precisely the point for the Bible's translators. The erasure of differences between texts authored for entirely different audiences and realizing very different sets of features equates the language of all African diasporic speakers, whether fictional or real, and posits that language as defective and morally dangerous.[12]

Of the verb-phrase-type features, zero copula is the most dispersed (DP = 0.30), followed by invariant present (DP = 0.35) and invariant stem (DP = 0.40). One of the perhaps surprising results of the analysis is that despite the dispersion of the

[12] In addition to referencing *The Padlock*, the author also expresses derision for Walter Scott's use of literary dialect. The author, however, constructs a clear hierarchy, with the racialized literary dialect of Mungo construed as even more incomprehensible than the regional one.

zero copula in African diasporic dialogue, the feature is not statistically indexical of African diasporic literary dialect. Of the 222 coded features, zero copula has the fifth lowest F-score ($F = 0.89$) by ANOVA. By log-likelihood, the p-value shows only low significance ($p < 0.05$) in a comparison of African diasporic and Indian dialogue ($LL = 6.06$). In a comparison of African diasporic and Chinese dialogue, the zero copula is actually more frequent in Chinese dialogue, though again the difference shows only low significance ($LL = 4.43$).

The quantitative data appear to suggest that during the eighteenth century the zero copula was an indicator of general nonstandardness, occurring frequently in representations of all three groups of speakers. However, some contemporaneous accounts refine that picture somewhat. In his nineteenth century English grammar, for example, J. W. F. Rogers comments on the zero copula. He begins with a discussion of the clause "Socrates is just" (Rogers, 1883, p. 205). He notes that the construction requires the verb *to be* before the adjective *just* and contrasts it with the clause "Fish swim" in which the verb "can predicate immediately" (i.e., it does not require any intervening grammatical structure between the noun *fish* and the verb *swim*). From the comparison of these structures, he concludes that in a clause like "Sun bright" the adjective *bright* is being made "a verb of what grown people use as an adjective." Rogers interprets the zero copula as forcing adjectives to function as verbs. He comments further:

> (6) In children's prattle and in such broken English as Negroes and Chinamen often speak, many words are improperly employed as verbs which are not recognized as such in that polite usage which grammarians and logicians are supposed to cultivate.

For Rogers, the zero copula is a marker of both infantilized "prattle" and the "broken English" of subaltern speech. In using the zero copula to link the language of Chinese and African diasporic speakers with that of children, he figures the zero copula as an index of underdeveloped and ungrammatical English. That perception conforms to the quantitative patterns and further suggests a racialized valence to its indexicality. One possible explanation for the feature's status is its relative rarity in historical varieties of British English. In discussing the presence of the zero copula in the Knaresborough Daybook (a collection of daily reports written by the manager of a workhouse in the late eighteenth century), García-Bermejo Giner and Montgomery (2001, p. 356) suggest that its realization in the Yorkshire text is so unusual that it is likely idiosyncratic. That uncommonness in British varieties makes the zero copula a feature that can be used to distinguish not only standard voices from nonstandard ones, but

also rustic, regional, and working-class British voices from the voices of imperial subjects and subaltern peoples.

Table 5.4: Frequencies of pronoun-type features in African diasporic dialogue, where DP < 0.80.

Feature	N	% Global	Freq.	DP
PRONOUN-TYPE				
TOTAL	**778**	**7.70%**	**29.31**	
me subject	304	3.01%	11.45	0.57
him subject	160	1.58%	6.03	0.57
them subject	11	0.11%	0.41	0.77
you possessive	24	0.24%	0.90	0.73
him possessive	94	0.93%	3.54	0.61
demonstrative *them*	32	0.32%	1.21	0.72

After verb-phrase-type features, pronoun-type features are the next most frequent in the morphosyntactic category for African diasporic dialogue. Of the pronoun-type features, the most frequent and most dispersed are the object pronouns *me* (DP = 0.57) and *him* (DP = 0.57) being used as clausal subjects. As with the zero copula, object-pronouns-as-subjects are not statistically distinctive markers of African diasporic literary dialect, but they do exhibit some limited differences with Chinese literary dialect and even more substantial ones with Indian literary dialect. Subject *me* has a low F-score (F = 0.51) by ANOVA. A log-likelihood comparison shows no statistical difference with Chinese dialogue (LL = 0.00), yet a significant difference ($p < 0.0001$) with Indian dialogue (LL = 64.58). Subject *him* has a higher F-score (F = 4.33), though the significance of that value is still low ($p < 0.05$). A log-likelihood comparison once again shows a significant difference with Indian dialogue (LL = 54.83). It also shows a more robust difference with Chinese dialogue (LL = 9.74) than does subject *me*.

Nineteenth century descriptions of these pronominal features often suggest their racialization in ways similar to the descriptions of the zero copula. An anonymously authored short story, for example, suggests that it is "very easy to talk [American] *Indian* by the simple recipe of transposing the nominative and objective cases of the personal pronoun" ("A fast day," 1864, p. 689). Similarly, in his recollections of sailing to Liberia, Charles Rockwell describes the African crew-members recruited while docked in Monrovia. "[T]hey spoke a broken English," he claims, "in which the pronoun *me* was almost the only one used" (1842, p. 258). Even

more pointedly, the following quotation from the magazine *The Atlantic Monthly* links pronoun usage to the psychology of its users. It comes from an article titled "The Horrors of Santo Domingo" in which the author seeks to explain the causes of the Haitian Revolution that began in 1791. The excerpt is part of a passage describing the development of Creole French, which the author terms "a new colonial language." It is, according to the author, on the one hand "bright and sparkling" but on the other having "no grammatical reason" and resembling "the charming gabble of children":

> (7) These characteristics appear in the formation of the Creole French, in connection with another childlike habit of the negro, who loves to put himself in the objective case, and to say *me* instead of *I*, as if he knew that he had to be a chattel. (Weiss, 1863, p. 301)

Although the author is describing French, the grammar is exemplified entirely in English. The effect is a cross-linguistic signaling – evaluations about French Creole that also attach to English variants. Those evaluations infantilize African diasporic vocal culture (much like Rogers does in excerpt 6) and suggest speakers' psychological predisposition to servitude. Because they use first-person object pronouns as clausal subjects, they understand themselves as objects. Grammatical function is metaphorized, and linguistic structure is used to rationalize the material realities of slavery.[13]

In contrast to the relatively high frequency of morphosyntactic features, orthographic features constitute the least frequent superordinate category. In spite of their low frequency, eye-dialect-type features are moderately well dispersed in African diasporic dialogue (DP = 0.48). Underlying these two measures (low frequency and moderate dispersion) is the propensity for the feature to be lexically restricted or to appear only once or twice in works published before 1870. Works in which the use of eye dialect could be described as more sustained or systematic (i.e., works in which eye dialect appears multiple times and in multiple forms) are primarily published later in the nineteenth or early in the twentieth centuries. These include *Lutchmee and Dilloo* (1877), *Middy and the Moors* (1888), *Black Man's Ghost* (1889), and *Three Men on the Bummel* (1900). A rise in eye dialect is backed by log-likelihood comparisons of the three periods. A comparison of early (pre-1830) and late (post-1870) periods shows a statistically significant increase (LL = 12.10, p <

[13] Charles Loring Brace, who is best known for establishing the Children's Aid Society and for the massive relocation of orphaned children from the East Coast of United States to the Midwest, quotes the passage in (7). He uses it in support of his argument that "Languages are the best evidence of Race" (Stierstorfer, 1996, p. 157).

0.001). The rise in eye dialect corresponds to the hardening of racist, comic stereotypes, which proliferate after the American Civil War (Boskin, 1986; Jones, 1999). It also aligns with an increase in phonological marking, which I discuss later in the chapter. Both of these are evident in the following excerpt from Jerome K. Jerome's *Three Men on the Bummel* (eye dialect in bold):

(8) Yes, sar, dat's what I'se **cumming** to. It **wuz** ver' late 'fore I left Massa Jordan's, an' den I **sez** ter mysel', **sez** I, now yer jest step out with yer best leg foremost, Ulysses, case yer gets into trouble wid de ole woman. Ver' talkative woman she is, sar, very –

As I suggested in the statistical overview (§4.3.3), The anecdote is intended as a comic example of digression, and the speaker, Ulysses, is figured as a clownish fabulist. The exaggerations of his character are reinforced by the exaggerations of his voice, which is marked by a high number of both eye dialect and phonological features.

Table 5.5: Frequencies of phonological subcategories in African diasporic dialogue.

Feature	N	% Global	Freq.	DP
PHONOLOGICAL-TYPE				
TOTAL	**4599**	**45.49%**	**173.28**	
consonant substitution	2808	27.77%	105.80	
consonant deletion	693	6.85%	26.11	
insertion	284	2.81%	10.70	
vowel substitution	531	5.25%	20.01	
metathesis	12	0.12%	0.45	0.78
syllable deletion	232	2.29%	8.74	0.43
exaggerated	31	0.31%	1.17	0.79

As I noted at the beginning of this section, phonological features account for the largest percentage among the four superordinate categories in African diasporic dialogue (45%). The most frequent subcategory among phonological features is consonant substitution (see Table 5.5). Although there are a total of 31 different types consonant substitutions realized in African diasporic dialogue, only a small subset of those have dispersions where DP < 0.80 (see Table 5.6). The most dispersed and most frequent of these is *t*/*d*-for-*th* substitution. In fact, *t*/*d*-for-*th* substitution is the most dispersed feature overall in African diasporic dialogue (followed closely by address).

Table 5.6: Frequencies of consonant-substitution-type features in African diasporic dialogue, where DP < 0.80.

Feature	N	% Global	Freq.	DP
CONSONANT SUBSTITUTION-TYPE				
TOTAL	**2808**	**27.77%**	**105.80**	
t/*d*-for-*th*	1903	18.82%	71.70	0.22
b-for-*v*/*f*	595	5.89%	22.42	0.42
n-for-*ng*	96	0.95%	3.62	0.67
r-for-*l*	15	0.15%	0.57	0.76
f-for-*th*	32	0.32%	1.21	0.78

A list of the ten most dispersed phonological features (see Table 5.7) shows that in addition to three consonant substitutions (*t*/*d*-for-*th*, *b*-for-*v*/*f*, and *n*-for-*ng*), consonant deletions are common in the form of cluster reduction, word-final deletion, and word-initial deletion. Word-final deletions are largely restricted to rhotics (*-r*), which account for 65% of occurrences. Word-initial deletions most frequently occur with the pronouns *THEM* or *HIM* being realized as *em*, *um*, or *im* (43% of occurrences). Word-initial deletions are also frequent in the variants for *YES*, *iss* and *is* (35% of occurrences). The variants of *YES* were discussed in the previous chapter as productive environments for the realization of *i*-for-*e* vowel substitutions, as well. There, I noted that these respellings are enregistered indexes of African diasporic vocal culture, and, in fact, variants of *YES* make up 80% of all *i*-for-*e* substitutions in African diasporic dialogue. The other vowel substitution on the list of most dispersed features, *a*-for-*e*, was also discussed in the previous chapter. It appears most frequently pre-rhotically (63% of occurrences) in, for example *sarpant* for *SERPENT*, *whar* for *WHERE*, *sarve* for *SERVE*, and *sarvent* and *sarvint* for *SERVANT*.

Table 5.7: The ten most dispersed phonological features in African diasporic dialogue.

Feature	N	% Global	Freq.	DP
t/*d*-for-*th*	1903	18.82%	71.70	0.22
b-for-*v*/*f*	595	5.89%	22.42	0.42
syllable deletion	232	2.29%	8.74	0.43
cluster reduction	392	3.88%	14.77	0.46
word-final deletion	112	1.11%	4.22	0.57
a-for-*e*	83	0.82%	3.13	0.57
word-initial deletion	185	1.83%	6.97	0.61
-ee/*-y*/*-i* final	120	1.19%	4.52	0.63
n-for-*ng*	96	0.95%	3.62	0.67
i-for-*e*	89	0.88%	3.35	0.67

The only insertion included in the table is *-ee*/*-y*/*-i* final. The association of the feature with African diasporic voices – and some African diasporic communities in the Caribbean more specifically – is attested to by the designation of Surinam creole as "talkee-talkee" in the early nineteenth century debates regarding the Bible translation. In the corpus, the phonological feature is prominent, for example, in the dialogue Zebby, a servant in Barbara Hofland's (1816) *Matilda, or, The Barbadoes Girl* (*-ee*/*-y*/*-i* final insertions in bold):

> (9) Poor Zebby, delighted with the goodness of her young mistress, audibly expressed her pleasure, with all the characteristic warmth of her country, and not a little proud of those virtues which she fancied she had assisted to nurture. – "Oh," cried she, "dis be my own beautiful Missy own goodness; she **makee** joy in her mamma heart; she **makee** poor negro all happy – **singee** and **dancee** every body; no more whip, massa Buckraman – every body delight – every body glad – every body good Christian, when Missy go back!"

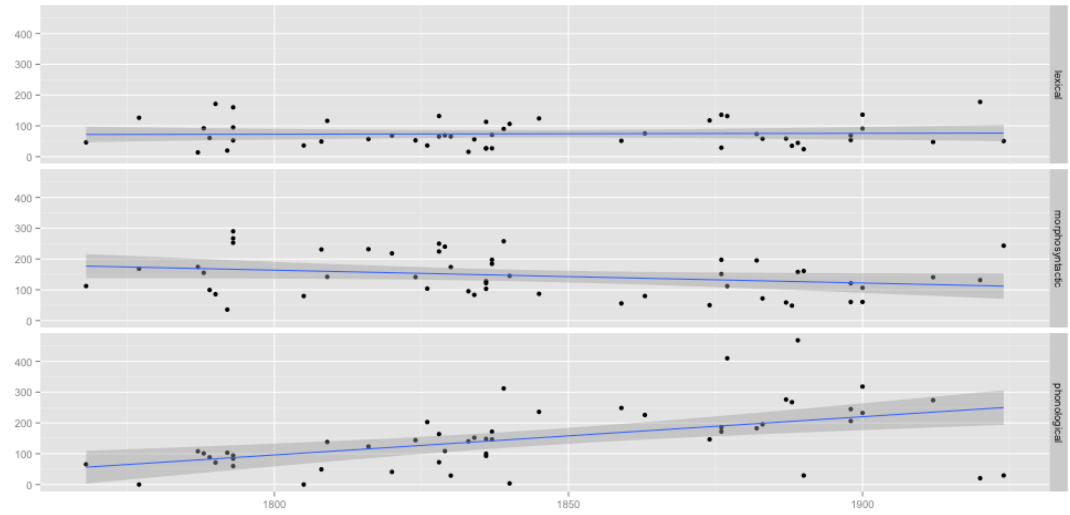
The example of Hofland's novel is prototypical not only because of its imagined Caribbean speaker but also because of its date of publication. Although *-ee*/*-y*/*-i* final insertions continue to appear in African diasporic dialogue throughout the corpus, they are most frequent in texts published before 1830. In a comparison of the early and middle periods, a log-likelihood comparison yield LL = 10.24 ($p < 0.01$) and in a comparison of the middle and late periods, LL = 21.88 ($p < 0.0001$). The frequencies of the feature in the early period are significantly greater than in the middle period, which, in turn, is greater than in the late period. Thus, the feature is one that exhibits a significant decline though the nineteenth and into the twentieth century. That trajectory actually runs counter to the trajectory of the phonological category as a whole, which increases over much of the nineteenth century, as we will see. That increase is one of the defining trends in the changing conventions of representing African diasporic vocal culture.

5.3 Diachronic trends in African diasporic dialogue

In the statistical overview, a regression analysis of overall frequencies in African diasporic dialogue revealed an increase in literary dialect features over time. Underlying that dominant trajectory, however, are differing trends, which we can see by separating African diasporic dialogue into its three main constituent categories. Figure 5.2 shows that the trend for lexical features has a nearly flat slope ($\beta = 0.03$) and that the trend for morphosyntactic features has a negative slope ($\beta = -0.42$). It is the phonological category that increases over time ($\beta = 1.24$) and, thus, appears to be

driving the overall rise in features. For the bulk of this section, I will be focusing on this trend in phonological features not only because the category plays a prominent role in shaping overall diachronic patterns in frequency, but also because it contains what are arguably the most indexical features of African diasporic literary dialect.

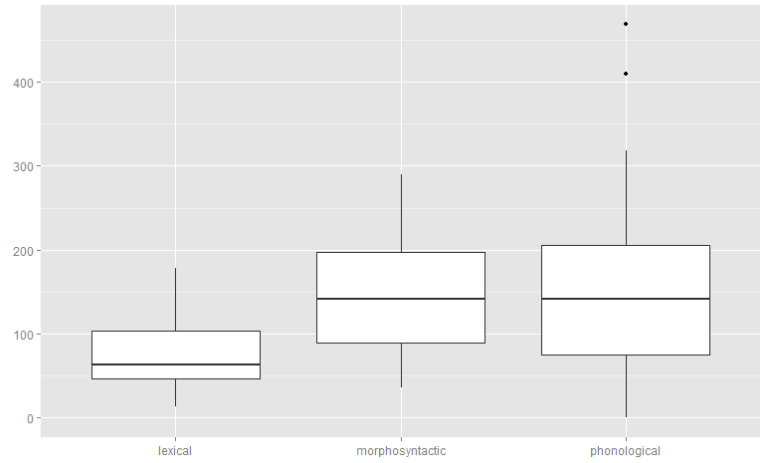
Figure 5.2: Scatter plots showing linear trends in frequency for the lexical, morphosyntactic, and phonological categories for African diasporic dialogue. The grey areas indicate the 95% confidence intervals.



That indexicality is partly suggested by the dispersion of a feature like *t*/*d*-for-*th* substitution, which is discussed above. More telling, however, is the ANOVA data that were presented in the statistical overview. The ANOVA data show that there are four features that distinguish African diasporic literary dialect. One of those, address (F-value = 8.50, $p < 0.0001$), is lexical. That feature significantly differentiates African diasporic and Indian from Chinese dialogue, as determined by a post-hoc Tukey test, but not African diasporic and Indian dialogue from each other. (It is important to remember, of course, that this is a categorical measure and that *MASSA* as an instantiation of the category is undoubtedly an iconized feature of African diasporic literary dialect. I examine address-type features in more detail in the next chapter.) The other three features that are significant belong to the phonological category: *t*/*d*-for-*th* substitution (F-value = 20.60, $p < 0.0001$), *b*-for-*v*/*f* substitution (F-value = 10.23, $p < 0.0001$), and cluster reduction (F-value = 5.11, $p < 0.001$). Additionally, *t*/*d*-for-*th* substitution has the third highest F-value overall (behind *l*-for-*r* substitution and *-ee*/*-y*/*-i* final insertion, both of which are representative of Chinese dialogue), and all three show high significance in differentiating African diasporic from both Indian and Chinese dialogue by a post-hoc Tukey test. In light of these results, it is apparent that: 1) iconized word forms like *MASSA* serve an important

signaling function in differentiating African diasporic dialogue from other racialized literary dialects, and 2) the most statistically significant markers of African diasporic literary dialect belong to a small subset of phonological features. I will examine the diachronic trends in phonological features shortly, but, first, I want to briefly discuss a few of the details related to the trends in the lexical and morphosyntactic categories.

Figure 5.3: Box plots for the frequencies of lexical, morphosyntactic, and phonological categories for African diasporic dialogue.

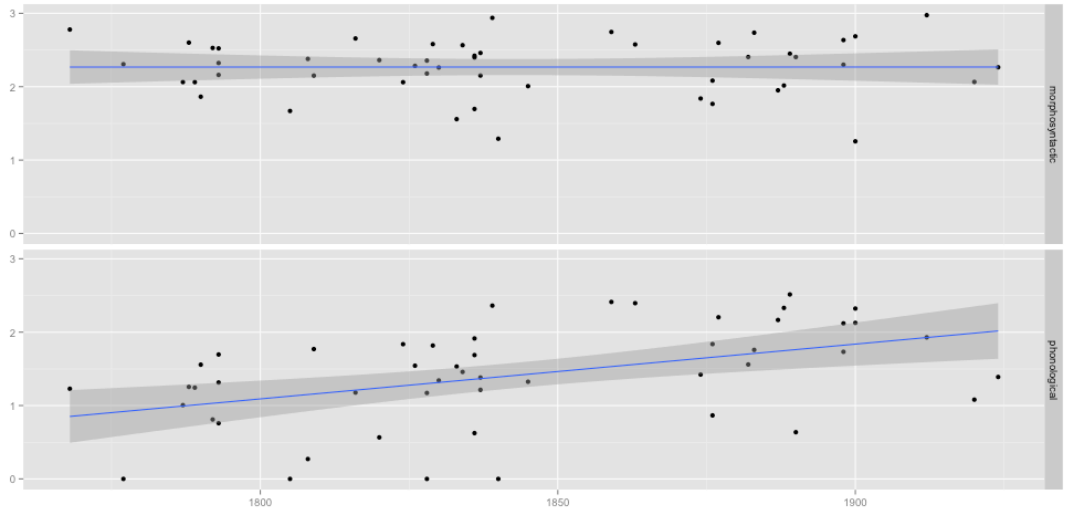


Lexical features have the lowest mean (74.21 per 1000) and standard deviation (41.74) of the three main categories in African diasporic dialogue. The relatively low interquartile range is illustrated in the box plots (see Figure 5.3) and is suggested by the more compact confidence interval in the regression plots (see Figure 5.2). If there were a positive or negative slope to the regression line, we might expect these properties to correspond to a relatively strong r-squared value. However, the slope for lexical features in Figure 5.2 is approaching zero; it is a nearly flat line. That suggests that there is no linear relationship between the two variables, time and frequency, for the lexical category. The results, in fact, predictably produce a very low r-squared ($r^2 = 0.0009$), and the lack of a linear relationship is confirmed by a low correlation measure (Kendall's $\tau = 0.04$). Put simply, the variation in the lexical category is unrelated to time.

The regression analysis of the morphosyntactic category produces a somewhat more robust result. The r-squared is higher, though still not particularly strong, which is reflective of the noisiness of the data ($r^2 = 0.07$, $F = 3.35$, $p < 0.1$). The correlation of time with frequency is negative (Kendall's $\tau = -0.18$, $p < 0.1$), which corresponds to the negative slope of the regression line. However, like the r-squared value, the

correlation measure is only moderately significant. Thus, while there is a downward trend in the frequencies of morphosyntactic features, that trend is only partially explanatory and is complicated by the variability of the category. By contrast, the upward trend in the phonological features explains the diachronic movement in the category far more conclusively. Both regression analysis ($r^2 = 0.25$, $F = 16.15$, $p < 0.0001$) and correlation (Kendall's $\tau = 0.42$, $p < 0.0001$) reveal a significant, positive relationship between time and frequency.

Figure 5.4: Scatter plots showing the linear trends in diversity for the morphosyntactic and phonological categories for African diasporic dialogue. The grey areas indicate the 95% confidence intervals.

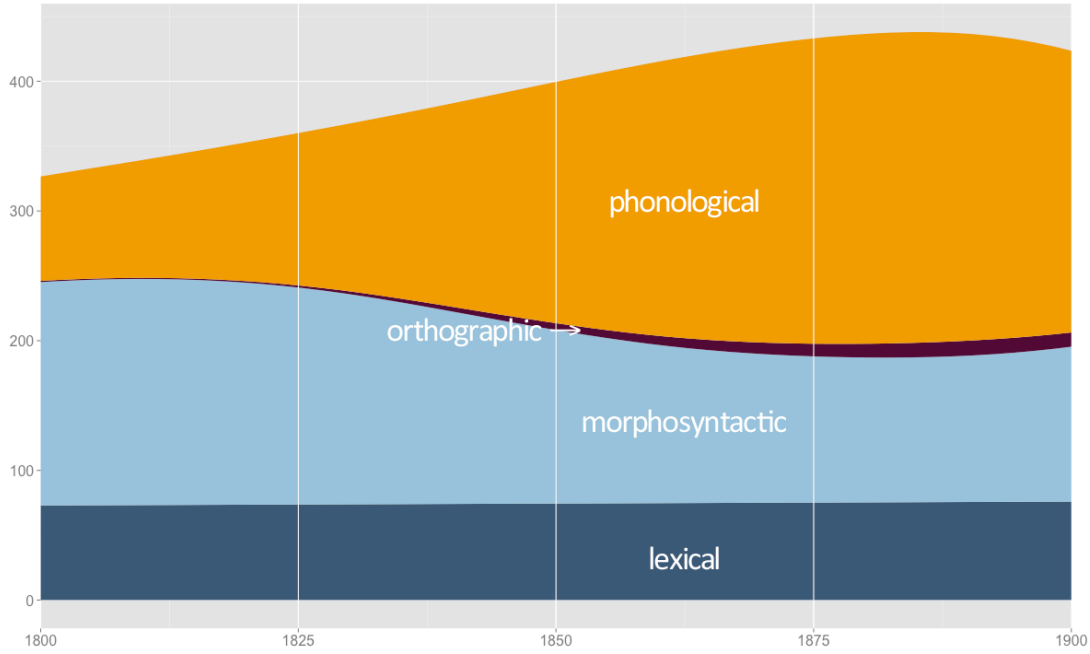


In addition to increasing in their frequency, phonological features also expand in their diversity. Figure 5.4 presents scatter plots for the two categories that exhibit changes over time: morphosyntactic and phonological. The morphosyntactic category shows no change ($\beta = 0.00$). The phonological category, however, shows rising diversity ($\beta = 0.007$, $r^2 = 0.21$, $F = 12.89$, $p < 0.001$). Thus, the category demonstrates parallel trends: one towards a greater frequency of features, and one towards a greater range of features.

These trends are the primary focus of the remainder of the section. The first step in this process is to carry out the regression analysis using an alternative model. In later chapters, I do this using approaches that include segmented regression and quantile regression. I am not applying those here primarily because they fare no better (and sometimes worse) in explaining patterns in the data under consideration. The model that I will be using – a generalized additive model, which combines properties of generalized linear and additive models (Hastie & Tibshirani, 1990) – improves the

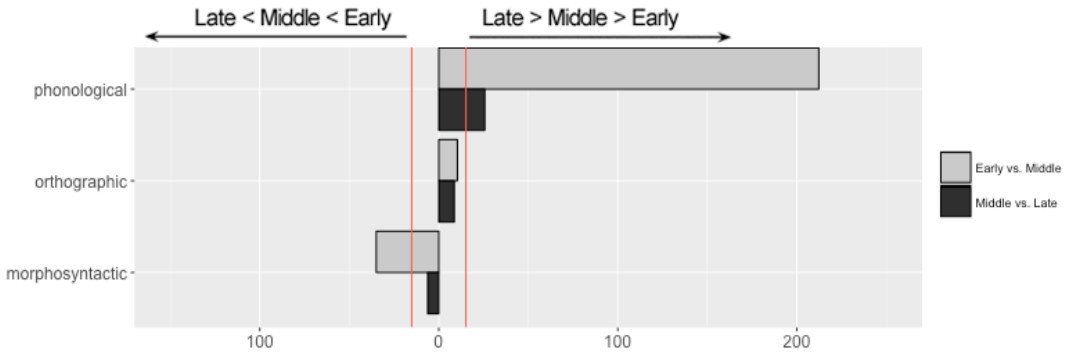
results not only for the morphosyntactic category (which the standard method of regression did not explain particularly well) but also for the phonological category (which it did). As an added benefit, the plots that are generated from this kind of analysis can be fit into stacked area charts, producing versions of streamgraphs that help to visualize changes in frequencies over time.

Figure 5.5: Stacked area chart showing the nineteenth century trends (using a generalized additive model) for frequencies of the four superordinate categories in African diasporic dialogue.



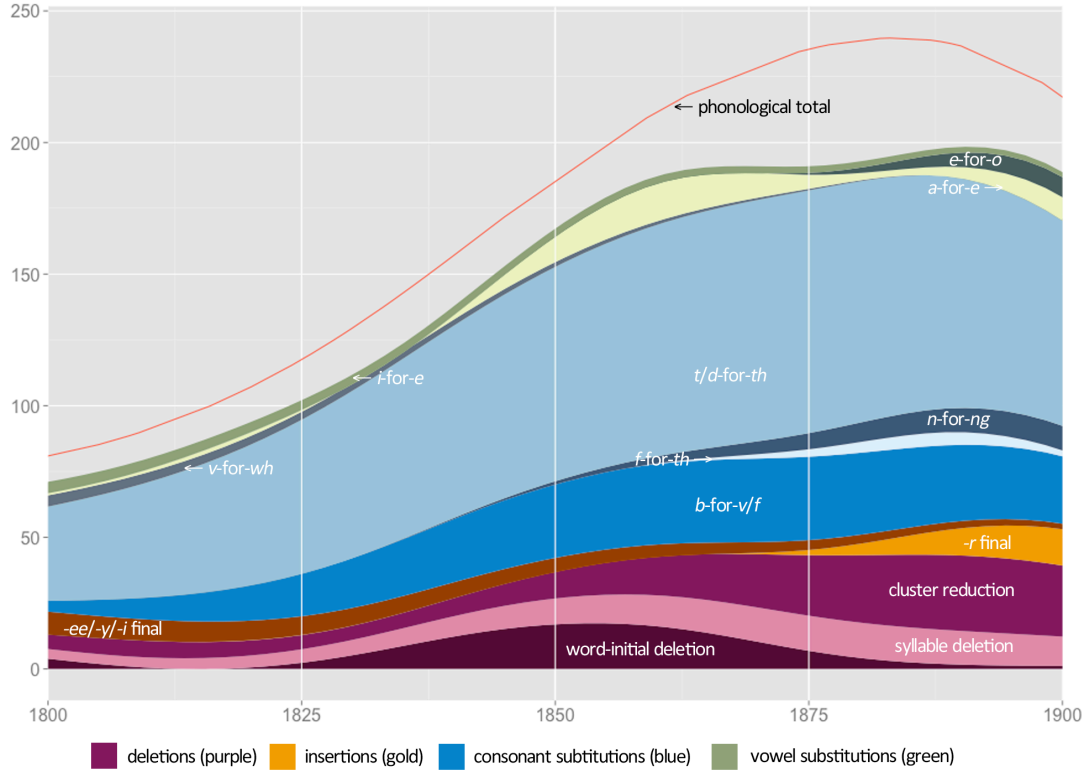
Application of a generalized additive model increases the r-squared for the morphosyntactic category from 0.07 to 0.17 and the r-squared for the phonological category from 0.25 to 0.44. A plot of the regression lines (see Figure 5.5) includes the orthographic category and focuses on the nineteenth century – the period during which phonological features undergo their rise. The chart suggests that the greatest increase in the phonological category occurs during the middle of the century. That rise appears to be mirrored by a decline in the morphosyntactic category over that same period. These changes are supported by log-likelihood comparisons. For all three categories for which there are significant changes across periods, the changes from the early (pre-1830) to the middle periods (1830-1878) are the most significant (see Figure 5.6). The transition from the early period to the middle one is particularly dramatic for phonological features (LL= 212.26).

Figure 5.6: Bar plot showing log-likelihood comparisons between the early (pre-1830), middle (1830-1880), and late (1880-1930) periods for the morphosyntactic, orthographic, and phonological categories in African diasporic dialogue. The red lines mark the points at which $p < 0.0001$.



Frequencies of individual phonological features, of course, do not change uniformly. Some increase in parallel with the growth in the overall category. Others, however, emerge in the middle of the century, and still others actually decline, exhibiting trends that run counter to the prevailing one. In order to illustrate some of that underlying complexity, the same techniques that were used in creating Figure 5.5 can be applied to a select number of features representative of the four phonological subcategories: deletions, insertions, consonant substitutions, and vowel substitutions. The resulting plot is presented in Figure 5.7.

Figure 5.7: Stacked area chart showing the nineteenth century trends (using a generalized additive model) for selected phonological features in African diasporic dialogue.



One of the patterns that the plot makes clear concerns the phonological features that are the most dispersed in African diasporic dialogue (*t*/*d*-for-*th* substitution, *b*-for-*v*/*f* substitution, syllable deletion, and cluster reduction). All four of those show increased frequencies through much of the nineteenth century, contributing to the dominant trend. Similarly, a number of features (e.g., *n*-for-*ng* substitution, *f*-for-*th* substitution, *-r* final insertion) emerge in the middle of the century, which is consistent with the rise in diversity indices. By contrast, two of the features that are included on the chart (*v*-for-*w*/*wh* substitution and *-ee*/*-y*/*-i* final insertion) have an opposing trajectory. Their frequencies decrease over the course of century. This is important for the reasons outlined below.

Up through the early nineteenth century, *v*-for-*w*/*wh* substitution appears to be an index for generic nonstandardness. This timeframe overlaps with the feature's relatively brief presence in the corpus – occurring first in Mariana Starke's (1788) *The Sword of Peace* and the last in Edward Howard's (1836) *Rattlin the Reefer*. Long before Starke's play, *v*-for-*w*/*wh* substitution had been used to represent a variety of nonstandard English speakers, such as the French Dr. Caius in Shakespeare's *The Merry Wives of Windsor* ("By gar, de herring is no dead so as I **vill** kill him") and the Irish Tegue O Divelly in Thomas Shadwell's *The Lancashire-Witches and Tegue O Divelly, the Irish-Priest* ("Now, I **varrant** you Joy, I **vill** do de Devil's business for him, now I have dis Holy-**Vater**").

As a constituent of African diasporic literary dialect, the feature most famously appears in a series of articles mocking the patrons and performers of a black-run New York theater, the African Grove. The articles were produced by the *National Advocate*, which was led by the journalist Mordecai Manuel Noah. One, originally published in September of 1821 and widely reprinted in both the U.S. and Britain, is an account of a production of *Richard III*. It describes Richard, "performed by a fellow as black as the ace of spades," striding upon the stage and delivering the line: "Now is de **vinter** of our discontent made glorus summer by de son of New York" ("African amusements," 1821). Noah's use of *v*-for-*w*/*wh* substitution is noteworthy partly because it coincides with the feature's lifespan in the corpus (from 1788-1836) and establishes it as a feature that circulates transatlantically. It is additionally noteworthy because of Noah's prominence and influence. Noah was such a prolific producer of literary dialect that Hay (1994, p. 13) refers to him as "the father of Negro minstrelsy." He was particularly influential with the British comic actor

Charles Mathews, who Hay (1994, p. 13) argues was the first person to bring "Noah's words to the stage."

Mathews is an important figure in propagating American minstrel traditions in Britain, primarily through the theatrical travelogues (like *Trip to America*) that he performed after touring the United States in 1822-1823 (Bratton, 1981; Robinson, 2001). In making the connection between Noah and Mathews, scholars have observed that in some iterations, Mathews' lyrics share significant overlaps in style with Noah's literary dialect (Davis, 2011; McAllister, 2003). In particular, there is a focus on their shared use of *v*-for-*w*/*wh* substitutions (Dennison, 1982, pp. 511-512). The work in the corpus with the most direct connection to Mathews is *Americans Abroad*, which was based on *Trip to America* and in which he starred. Interestingly, *v*-for-*w*/*wh* substitution does not occur in the dialogue of the African diasporic character Agamemnon (or anywhere else, for that matter) in *Americans Abroad*. Whether this absence is the result of the influence of Mathews' collaborator and the play's designated author, Richard Brinsley Peake, is difficult to determine.

Regardless, a little more than a decade after the premiere of *Americans Abroad*, *v*-for-*w*/*wh* substitution disappears from the corpus as a constituent of African diasporic dialogue. It does not, however, disappear from the source works. In Herbert Strang's (1912) *The Flying Boat*, for example, it occurs in the dialogue of a German military officer ("But surely you **vill** make complaint!"). It is also used to voice a Turkish Jew in Bracebridge Hemyng's (1900) *Jack Harkaway's Boy Tinker among the Turks* ("If it ish true dat de closhe makes de man, you **vill** do excellent **vell**, and de people **vill** not now run after you") and the Danish Jan Steenbock in John Hutcheson's (1889) *The Black Man's Ghost* ("Yous can go below; I **vill** keep ze **vatch**"). These later realizations are consistent with scholarship that has noted the feature's stereotypical associations with German- and Yiddish-speaking émigrés around the turn-of-the-century (Appel, 1957; Jones, 1999; Kersten, 2000).[14] Thus, *v*-for-*w*/*wh* substitution shifts over the course of the nineteenth century from indexing generic nonstandardness to indexing specific communities. This narrowing of associations is evidenced not only by the feature's changing associations with African diasporic literary dialect, but also by a corresponding change in Irish literary dialect.

[14] The Danish Jan Steenbock would seem to be the exception here. His dialogue, however, is heavily inflected by German as is evidenced by his use of *nein* for *no*.

There, the feature apparently disappears after 1750 (Sullivan, 1980, p. 199), almost a century before its decline in representations of African diasporic speakers.

Shifting associations and the fossilization of linguistic stereotypes similarly appear to affect the decline in and *-ee*/*-y*/*-i* final insertions. Their decline in African diasporic dialogue overlaps with the emergent practice of rendering Chinese speakers in literary dialect in the latter part of the nineteenth century. That emergence is the focus of a subsequent chapter, but here it is enough to note that *-ee*/*-y*/*-i* final insertion is a frequent and iconic feature in the representation of Chinese vocal culture (Bolton, 2003; Jones, 1999). As those associations solidify, the feature's earlier associations with African diasporic speakers wane.

There are, however, other forces at play, for the link between *-ee*/*-y*/*-i* final insertion and African diasporic literary dialect is already weakening before the conventions of Chinese literary dialect become widespread. That weakening is at least partly catalyzed by the widening influence of North American authors and their representational practices. Previously, I suggested that *-ee*/*-y*/*-i* final insertion is often associated with Caribbean creole speakers early in the nineteenth century and provided the example of Zebby, the Barbadian servant in *Matilda, or, The Barbadoes Girl* (see excerpt 9). Although the word-final vowel is not consistently applied to representations of speakers of Caribbean creoles, its emblematic status is signaled by the designations "talkee-talkee" and "taki-taki" for a number of Caribbean varieties (Lalla & D'Costa, 1990; Léglise & Migge, 2007). In the source-works, the anonymous author of *Marly; or, A Planter's Life in Jamaica* (1828) uses this designation when he refers to Jamaican Creole as "the negro corrupted dialect, or the talkee talkee language."

Apart from representations of Caribbean creole speakers, the feature circulates in the United States, though mostly in the eighteenth century. It occurs, for example, in a satirical letter published after the defeat of an emancipation bill in New York ("A letter from Cuffee," 1785): "De Legislatermen no make de poo nega free las Sataday, because dey no **makee** *two turd*: So de poo nega law no passe for dat de Legislatermen no habbe *two turds*." And in a similar letter published in the *Massachusetts Spy* ("For massatuse pie," 1782): "Wene court **makee** rate for hors, for beef, for shurt, and token, an ebery ting, dont fokes pa him? Wy **canee** no make rate for size too?" That said, *-ee*/*-y*/*-i* final insertions are not typical of representations of

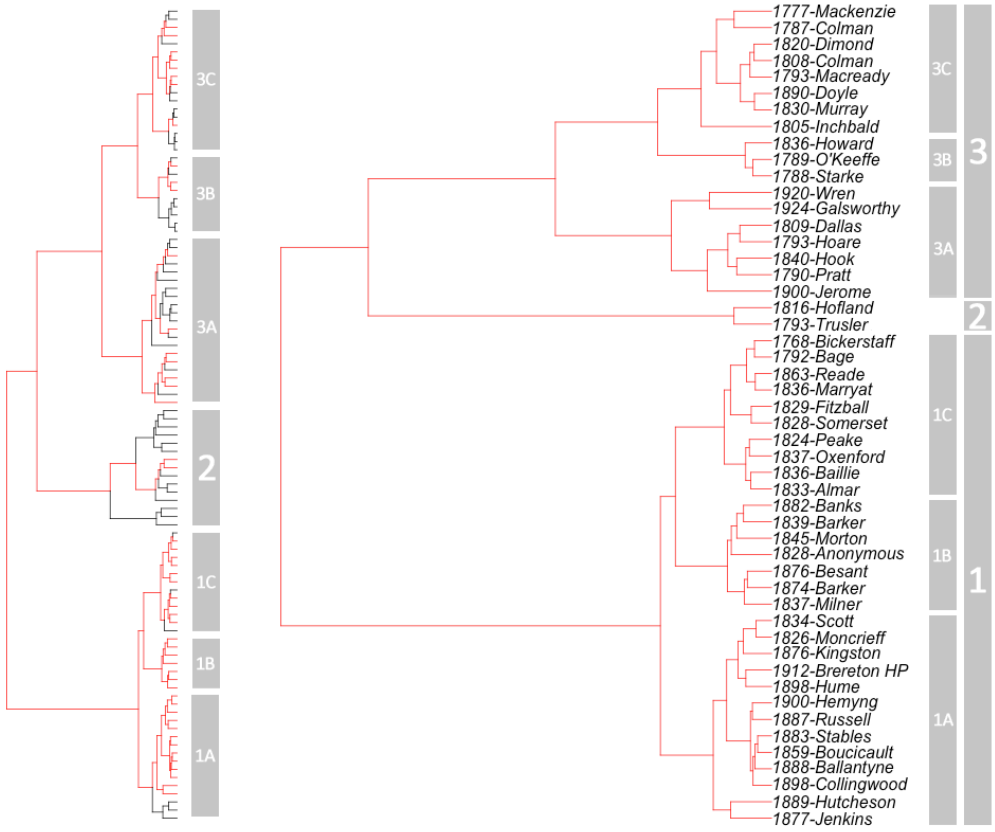
African diasporic speakers in America, particularly in works that circulate in the middle and later parts of the nineteenth century.

This difference is important, because one of the factors that likely contributes to the feature's decline is the increasing circulation of North American depictions of African diasporic vocal culture and their influence in Britain. During the same period that Figure 5.7 shows a rise in phonological features, there is an explosion of so-called "dialect novels" in the United States (Jones, 1999). Depictions of African American vocal culture play an important role in the movement – in works like Mark Twain's *The Adventures of Huckleberry Finn* and Joel Chandler Harris' *Uncle Remus* stories. A forerunner of those and similar works, Harriet Beecher Stowe's *Uncle Tom's Cabin*, is notable for its influence – inspiring not only sympathetic imitations, but also a reactionary genre of plantation literature – and its popularity in Britain (see, e.g., Holohan, 2013). Dion Boucicault refers to Britain's "Uncle Tom mania" in the letter that I quoted in the previous chapter (§4.4.3). Boucicault's evaluation is echoed by an American visitor to England in the 1850s who describes Stowe's novel as one of Britain's two "lions" – ideas that dominate the cultural discourse in their predatory and "unceasing repetition" (Tuckerman, 1854, p. 107). And in a testament to the global impact of Stowe's novel on nineteenth century perceptions of African American English, the linguist James A. Harrison protests its citation in the German philological journal *Anglia*, finding it "a shock to one's nerves to have 'Uncle Tom's Cabin' constantly cited in illustration of American Negro usage, phonetics, and philology" (Harrison, 1892).

The popularity of North American authors like Stowe coincides not only with the decline of word-final *-ee*/*-y*/*-i* in the corpus, but also with the emergence of features like word-final *-r* insertion, *n*-for-*ng* substitution, and *a*-for-*e* substitution. Just as the former is not present in her literary dialect, the latter of these figure prominently (Burkette, 2001). Moreover, the general rise of phonological features in African diasporic dialogue is consistent with the rise of dialect literature on the other side of the Atlantic. Commenting on the ubiquity of the "dialect novel" at the end of the nineteenth century, one critic laments that it has become "a sort of craze" that "may be regarded as a curse to the rising generation of fictionists" (de Leon, 1897, p. 680). Though dialect literature during this period is often framed as an American phenomenon, its practices, however regionally inspired, circulate and solidify new representational patterns across the Anglophone world.

5.4 Resemblances in African diasporic dialogue

Figure 5.8: A dendrogram zoomed for African diasporic dialogue. The numbered clusters on the right match their counterparts from the full dendrogram on the left. African diasporic texts are highlighted in red.



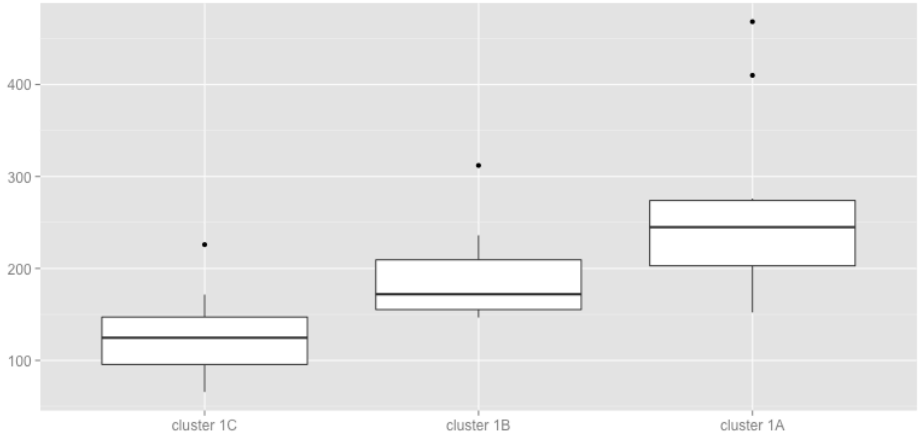
The illustration in Figure 5.8 is what is called a zoomed dendrogram.[15] On the left is the complete dendrogram that was presented in the previous chapter (§4.4.4). On the right is the "zoomed" portion – a dendrogram consisting of a subset of the data, which in this case is the African diasporic dialogue. The red highlighting shows where the clades on the right are situated in the larger structure on the left. For example, the top clade on the right (representing the African diasporic dialogue from Mackenzie's 1777 novel, *Julia de Roubigné*) corresponds to the third clade from the top on the left, which is the first red-colored clade. I have also numbered three clusters that I will be discussing in order to make their correspondences across the structures easier to identify.

One pattern that the dendrogram suggests is that, for a substantial subset of texts, the variation in African diasporic dialogue is built around a shared repertoire of

[15] Like the earlier dendrograms, Figure 5.8, too, was produced using the *APE* package for R (Paradis et al., 2004).

representational conventions. That there is a set of common features is illustrated by the large grouping of cluster 1 and its homogeneity. In the statistical overview, the heat maps similarly attested to the fact that few core features permeate cluster 1, though the frequencies of those features vary in cluster 1's three sub-clusters (see Figures 4.16 and 4.17). The most significant of those features are phonological: *t*/*d*-for-*th* substitution and *b*-for-*v*/*f* substitution.

Figure 5.9: Box plots for the frequencies of phonological features in the sub-clusters of 1.

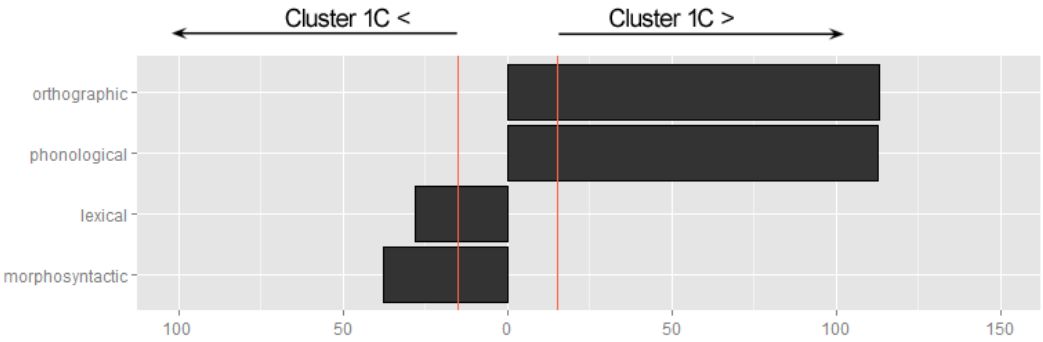


Our understanding of cluster 1 can be refined further by considering how the variation of phonological features across sub-clusters fits with the diachronic variation in phonological features discussed previously. The texts that populate sub-cluster 1C are primarily earlier texts, while those that populate 1A and 1B are primarily middle and later period texts. The frequencies of phonological features for the three sub-clusters are presented as a boxplot in Figure 5.9. As the boxplot illustrates, the mean phonological frequencies increase as you move down the dendrogram from sub-cluster 1C to sub-cluster 1A. Thus, cluster 1 captures both the conventions that are shared across time, as well as how increasing phonological frequencies shape evolving practices that group texts partly by time period.

Clearly, such groupings are neither uniform nor absolute. Within clusters, there are chronological outliers like Reade's *Hard Cash* in cluster 1C and Moncrieff's *Tom and Jerry* in cluster 1A. And even more obviously, texts from all periods fall outside those three sub-clusters. Many early texts congregate in clusters 3B and 3C. In fact, the split in early texts between clusters 1C and 3 appears to show at least two distinct lineages for African diasporic literary dialect: one (exemplified by cluster 1C) with underlying similarities to later conventions, and another (exemplified by the sub-

clusters of 3) realizing separate constellations of features. Structurally, this split results from differing distributions of features. These differences are clear in log-likelihood comparisons of the four superordinate categories (see Figure 5.10). The texts in cluster 1C are more orthographically and phonologically marked, whereas the texts in the sub-clusters of 3 are more lexically and morphosyntactically marked.

Figure 5.10: Bar plot showing log-likelihood comparisons between cluster 1C and the combined clusters 3A and 3B for the four superordinate categories in African diasporic dialogue. The red lines mark the points at which $p < 0.0001$.



These structural differences have ideological implications. Whereas cluster 1C is extracted from a relatively homogeneous cluster in the full dendrogram, clusters 3B and 3C are extracted from diverse clusters. For example, the dialogue of Amos from Elizabeth Inchbald's (1805) *To Marry or Not to Marry* is situated in a grouping of texts with Indian dialogue. Amos' dialogue realizes no phonological marking and has the lowest composite frequency (115.9) and second lowest diversity index (1.89) of all African diasporic dialogue:

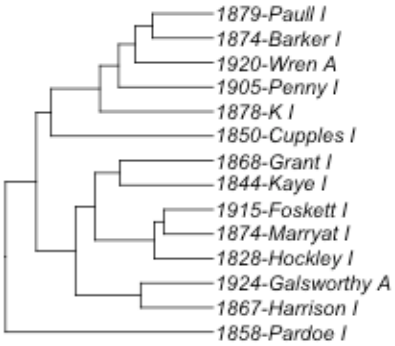
(10) Master, dear master, raise your head, and speak to poor servant, poor black, who has attend you from boy in his native country, followed you to your own, and is ready to follow you all the world over. Only tell him why you no eat why you no sleep and why big tear roll down from your eye?

The structure of Amos' dialogue, and other early instances like it, signify a kind of nonstandardness that is indicative of evolving racial categories and attitudes in the late eighteenth and early nineteenth centuries. These ideas are discussed at length in the following chapter. At this point, however, it is useful to note that some early examples of African diasporic literary dialect (like those in the sub-clusters of 3) encode generic linguistic and racial otherness – an otherness that is often expressed as the self-aware blackness and loyal servitude evident in (10). Other examples (like those in cluster 1C) imagine a racialized nonstandardness that is more distinctively figured as African diasporic, even as non-African diasporic identities (such as the Indian characters from

Neale's *The Port Admiral*, which were analyzed in the previous chapter) are mapped onto those signifiers.

In addition to the texts in clusters 3B and 3C, texts appearing in clusters 2 and 3A shed light on the variation that occurs in African diasporic dialogue. The pair of texts that appear in cluster 2 are situated on the full dendrogram in a cluster consisting primarily of Chinese dialogue. Hofland's novel was discussed above with attention to her use of *-ee*/*-y*/*-i* final insertion. The text that is paired with her novel, John Trusler's (1793) *Life; or, The Adventures of William Ramble, Esq.*, similarly realizes *-ee*/*-y*/*-i* final insertions in voicing Brutus, the slave of a former slave-trader, Mr. Raspe ("Me no **killee** Massa, – Bravo come and **killee** him"). Their shared and relatively frequent use of the feature is clearly a salient factor in their proximity to Chinese representations.

Figure 5.11: Cluster containing the African diasporic dialogue from *Cupid in Africa* (1920) and *The Forest* (1924).



More interesting still are the two texts paired at the top of cluster 3A: Christopher Wren's *Cupid in Africa* and John Galsworthy's *The Forest*. In the full dendrogram, they are embedded in a grouping that is otherwise comprised entirely of Indian literary dialect (see Figure 5.11). Their positioning reflects the imagined hybrid identities of the African diasporic characters in both works. In Wren's novel, Ali Suleiman is figured as an Ethiopian Muslim whose English is inflected by Swahili, Arabic, and British idioms:

> (11) "Bwana will wanting servant, ole chap," continued the negro, "don't it? I am best servant for Bwana. Speaking English like hell, sah, please. Waiting here for Bwana before long time to come. Good afternoon, thank you, please, Master, by damn, ole chap. Also bringing letter for Bwana… You read, thanks awfully, your mos' obedient servant by damn, oh, God, thank you, sah," and produced a filthy envelope from some inner pocket of the aforementioned night-dress, which, innocent of buttons or trimming, revealed his tremendous bare chest.

Note that Wren has Ali Suleiman using *bwana* (Swahili for *boss* or *master*) and *sah* as forms of address, not the conventional *massa*. This is important because, as was noted

earlier, address is a category that does not distinguish one form from another. Thus, there is an underlying pattern of coded features that suggests that Wren is aligning Ali Suleiman with aspects of Indian vocal culture – an unconventional alignment that is further confirmed by Wren's uncoded lexical choices. Ali Suleiman's connection to Indian vocal culture is, in fact, alluded to when he first appears and greets the novel's protagonist, Bertram Greene: "Jambo!" A moment of confusion follows because Greene does not understand "that the African 'Jambo' is equivalent to the Indian 'Salaam.'" Typically, glosses of foreign expressions are given in English, yet, here, the narration glosses Ali Suleiman's greeting in Urdu/Arabic, creating a suggestive juxtaposition.

The alignment between the African diasporic characters and Indian vocal culture is even more transparent in Galsworthy's play. Galsworthy figures his African diasporic characters as Sudanese Muslims and their dialogue makes frequent use of *sahib* as an address form:

(12) No can march if not eat. Lockyer Sahib tell men "Right about." Then obey – men march – all go back to river. Lockyer Sahib good – our officer – Strood Sahib –

Sahib has well-established associations with Indian literary dialect by the time Galsworthy authors his play. Galsworthy's specific use of the address form, as well as his and Wren's broader patterning of their dialogue, indicate both authors' efforts at fashioning regionally specific African diasporic voices that are inflected by Arabic and Islamic culture.

Figure 5.12: Cluster containing the African diasporic dialogue from *Americans Abroad* (1824) and *No Followers* (1837).

In some sense, the mapping of Indian vocal culture onto African diasporic subjectivities in the works of Wren and Galsworthy is an inversion of the mapping of African diasporic vocal culture onto Indian subjectivities in Neale's *The Port Admiral*, which I discussed in the previous chapter (§4.4.4). To conclude the analysis of African diasporic literary dialect, I would like to return to the cluster in which Neale's dialogue appears (see Figure 5.12). Neale's novel is part of a trifoliate grouping that includes the plays *Americans Abroad* by Richard Peake and *No*

Followers by John Oxenford's. I mentioned *Americans Abroad* earlier in this chapter in reference to Charles Mathews, upon whose travelogue the play is based and who played the lead (§5.3). Mathews has been posited as an important figure in spreading American minstrel traditions in Britain, in no small part because of his relationship to the American newspaper publisher M. M. Noah. But the networks of influence go far beyond Noah, which is why I want to examine Peake's play and its companion on the dendrogram, *No Followers*. Both serve to illustrate the complex currents of influence and ideology that circulate during a tumultuous period, as they straddle either side of the passage of the Slavery Abolition in 1833.

In *No Followers*, Lucius Lily quarrels with his drunk, white romantic rival, Toby Quondum. When Toby declaims Lucius as "an inferior order" and "nigger," Lucius replies:

> (13) No such ting! we men ob colour, de beauties ob de unibersal uniberse, we be de black spot on de domino – de white man be de white ob de domino, only put to show de black off to more advantage.

Lucius' declaration is an explicit inversion of the conventional figuring of race in the British imagination. Non-white characters are stereotypically positioned to contrast with and to rationalize the dominance of whiteness. As a resource in the signaling of that contrast, literary dialect "helps to underline the role of the standard," Blake (1981) argues, as much as it "emphasizes the deviant nature of the non-standard." In other words, literary dialect encodes not only the deviance of nonstandard speakers, but also the moral, political, and cultural authority of standard language vocal culture. In doing so, they reinforce the imperial systems that maintain that authority.

In *No Followers*, Lucius Lily makes this contrastive convention explicit and appears to turn it on its head. He rejects the dysphemic label and suggests that it is whiteness that exists only "to show de black off to more advantage." He is subversive in other ways, too. A number of the works in the corpus feature a nonstandard speaking African diasporic or Indian woman who is romantically partnered with an Anglo man. (I discuss these more in the following chapter). Lucius Lily, however, is the only non-European, nonstandard speaking man who is partnered with an Anglo woman. At the play's conclusion, he wins the affections of Mary Magnet, a servant who has been the subject of romantic competition between Lucius and Toby.

In the heat of that competition, the play's final scene has Lucius and Toby breaking a statue titled "Pity the poor African," which belongs to an emancipation society and is an apparent parody of the iconic image "Am I not a Man and a

Brother." As Mrs. Warnmore, Mary's employer, has decreed that Mary have "no followers" while she is out, Lucius Lily takes the place of the broken statue when she returns in order to disguise himself. The scene clearly satirizes middle class attitudes toward race. When Lucius camouflages himself as the statue, he quite literally embodies "the black image" – which is how Mrs. Warnmore refers to the figure, and the phrase serves as the play's subtitle. He becomes an emblem of British racial hypocrisy. Mrs. Warnmore's emancipation society may espouse pity for "the poor African," but she refers to her group as a "ladies nigger association," where "that sable gentleman will preside over the tea-pot." However, the primary purpose of the scene is not social commentary, but comic spectacle, and the central player in that spectacle is Lucius Lily. "Lucius is first and foremost a comic buffoon," Stierstorfer (1996, p. 157) writes, "and only as a side effect does his comic appeal to the audience carry a point." A review of the premiere reports approvingly that when Mary's "Negro inamorato assumes the dress and attitude of the fractured figure," he "creates a good deal of fun in the personification" ("Strand theatre," 1837).

The example of *No Followers* is instructive in its engagement with the conventions of contrasting racialized bodies and voices. It clearly does not cast aside those conventions. The positioning of Lucius Lily's literary dialect is clear evidence of this. It is situated in a relatively homogeneous sub-cluster of contemporaneous representations, which, in turn, is nested in the larger cluster of African diasporic dialogue. Lucius Lily's literary dialect is conventional for its time and in keeping with the conventions that propagate over time. Nonetheless, the play is unusually explicit in its acknowledgment of staging difference and its potential social significations. The play is also produced at a time of roiling debates about race and empire. The passage of the Slavery Abolition in 1833 is a pivotal moment.[16] The period leading up to it is marked by the circulation of increasingly hostile representations of African diasporic culture, representations that seek to position African diasporic people as unworthy of the freedoms that new legislation afforded them. "Bobalition" propaganda, for example, mocked Abolition Day celebrations, which observed the 1807 passage of the Act Prohibiting Importation of Slaves. Some of the broadsides include "Grand Bobalition or Great Annibersary Fussible" (1821), "Grand and Splendid Bobalition of

[16] For an analysis of the rhetoric leading up to the Act's passage and the debate's role in the formation of a British national identity, see Swaminathan (2009). For a discussion of the period following the Act's passage, see Huzzey (2012), and for the response in American South to the Act, see Rugemer (2004).

Slavery" (1822), "Grand Celebrashun ob de Bobalition of African Slabery!!!" (1825), and "Bobalition of Slavery" (1832). As their titles suggest, the texts ventriloquize caricatures like "Cesar Crappo," the "sheef marsal," whose use of literary dialect is intended to denigrate the social and cultural practices of real African diasporic communities.

Although "Bobalition" propaganda primarily circulates in cities in the northeastern United States, its influence can be found in Britain, as well. The artist Gabriel Shear Tregear, for example, reworks a series of engravings by Edward Williams Clay called *Life in Philadelphia* for publication in Britain and adds a piece that is a kind of visual adaptation of the American texts, complete with "politicized dialogue" that, as Jenna Gibbs (2014, p. 146) observes, "came directly out of Bobalition broadsides of the 1810s and 1820s." Titled "Grand Celebration Ob De Bobalition Ob African Slabery," Tregear's print takes up the discursive tradition of "Bobalition" propaganda and revisions the occasion as a response to the passage of British Slavery Abolition Act rather than the American Act Prohibiting Importation of Slaves.[17] In doing so, it figures British abolition into an American tradition of racist parody.

Such complex networks of transatlantic influence are part of the milieu that shapes the creation of works like *No Followers* during this period. That influence is evident not only in Oxenford's use of the abolition debates as source of comedy, but also in his portrayal of Lucius Lily as a paradoxical character: who, on the one hand, speaks in the "politicized dialogue" of the time (to borrow Gibbs' term), but, on the other, is romantically linked to Mary Magnet. Thus, the play at once subverts and reinforces stereotypes as it engages with the era's racial tropes, which circulate across the Anglophone world.

We can see similar influence in the text to which it is most closely linked on the dendrogram: *Americans Abroad*. In Peake's play, the character of Agamemnon is first heard off-stage singing "Opossum up a Gum Tree:"

(14) Possum up a gum-tree
Up he goes – up he go
Racoon in de hollow,
Down below, down below!

[17] One figure addresses "De Orator ob de day," who is presumably William Wilberforce, the leader of the British abolition movement.

The song is one that had been made famous by Charles Mathews (Jortner, 2009). After visiting the United States, Matthews staged his successful one-man show, *Trip to America*, in which he performed skits and songs while impersonating different American types. One of the most famous of these was an enactment of an African American Shakespearean performance. Mathews played the part of Caesar Alcibiades Hannibal Hewlett, a black actor delivering Hamlet's soliloquy. In a print version, Mathews (1824, p. 11) renders his lines: "To be, or not to be? that is the question; whether it is nobler in *de* mind to suffer, or tak' up arms against a sea of trouble, and by *oppossum* end 'em." The substitution of "opossum" for "opposing" then leads him to break into "Opossum up a Gum Tree."

Rather than being based on his observations, as he claimed, Mathews' performance seems to be a riff on Noah's widely distributed description of a staging of *Richard III* at the African Grove in New York. For Noah and Matthews, the language of Shakespearean performance serves a contrastive function much like the one Lucius Lily ironically calls attention to in *No Followers*. It has an additional ideological purpose, however, for it works not simply to encode white authority, but to stigmatize black identity. It is a metonym for larger social and cultural participation, conjuring the African diasporic actors as imperfect mimics. Noah disparagingly calls them "imitative inmates of the kitchens and pantries." And Noah and Matthews are not alone in using Shakespeare in this way. Tregear publishes a print in 1834 called "Othello, Desdemona Asleep" as part of his collection *Tregear's Black Jokes*. In the print, Othello, with a candle in one hand and a scimitar in the other, approaches the sleeping Desdemona and says:

(15) Yet I'll not shed her blood;
Nor scar dat Whiter skin ob hers dan snow,
And smoove as monumental alabaster.
Yet she must die, else she'll betray more niggers.

Also in 1834, Maurice Dowling premieres his *Othello Travestie* in Liverpool. The burletta recasts Othello as "A Moor of Venice, formerly an Independent Nigger, from the Republic of Hayti." A print celebrating the comedy's 1836 run at the Strand in London depicts William John Hammond in the lead delivering the lines:

(16) A Gypsy woman whose name wad Powel
To my poor moder she gab dat towl.

The sounds and images of minstrel Shakespearean performance stand as totems for white skepticism of political, economic, and social equality. Published a month before the account of *Richard III*, another article from the *National Advocate*

describing African Grove patrons concludes by portraying them as would-be bon vivants disengaged from abolition-related political events like the Missouri Compromise of 1820: “They fear no Missouri plot; care for no political rights; happy in being permitted to dress fashionable, walk the streets, visit African Grove, and talk scandal” ("Africans," 1821). When the account of the *Richard III* performance is circulated in the British press (Jerdan, 1821, p. 751), it is likewise framed as a comment on the social effects of African diasporic freedoms:

> (17) A New York Journal contains the following ludicrous account of the performances of a *negro amateur corps* in that city; to preface which it may be necessary to state, that the measures in Congress for the emancipation of the black slaves, are represented as having the effect of greatly exalting the notions of the coloured race.

Though common in the 1820s and 1830s, the paranoid mockery exemplified by minstrelic Shakespeare and “Bobilation” propaganda was neither universal nor unchallenged. James Hewlett, the African American actor Mathews caricatures, published a letter in the *National Advocate* ("Matthews," 1824), which addresses Mathews directly:

> (18) You have, I perceive by the programme of your performance, ridiculed our *African Theatre in Mercer-street*, and burlesqued me with the rest of the negroe actors, as you are pleased to call us –mimicked our styles – imitated our dialects – laughed at our anomalies – and lampooned, O shame, even our complexions. Was this well for a brother actor?

Though nothing like the direct rebuke of Hewlett, the character of Lucius Lily does not appear to wholly endorse the most aggressively paranoid forms of burlesque. Nonetheless, this is the context in which *No Followers* is authored and performed: a time of battles over the legal, social, and economic implications of abolition; and a time of increasing circulation of people, texts, and ideas around the Anglophone world.

5.5 Conclusion

Of African diasporic literary dialect, then, the following conclusions can be drawn regarding its structures in response to the first three research sub-questions (§1.3): 1) though iconized address forms are important indexes of African diasporic dialogue, the most statistically significant markers of African diasporic literary dialect are *t*/*d*-for-*th* substitution, *b*-for-*v*/*f* substitution, and cluster reduction; 2) over time, there is a significant increase in the frequency and complexity of phonological marking in African diasporic dialogue; and 3) early representations are primarily spread across three sub-clusters, but later ones show more homogeneity in their

resemblances, albeit with some remaining variation. Explaining those patterns falls to the fourth sub-question. To summarize some of the social and cultural forces that inform those patterns, I want to briefly frame them in the context of source work that does not meet the word-count threshold and, thus, is not included on the dendrogram: Samuel Foote's (1778) play *The Cozeners*.

The play stages a moment of cross-racial masquerade that throws into relief the issues undergirding the patterns described above. The plot involves Mrs. Fleece'em and Mr. Flaw's attempts to swindle Mr. and Mrs. Aircastle by convincing the Aircastles' son, Toby, to marry Mrs. Fleece'em's nonexistent niece.[18] The make-believe niece is supposedly an heiress, "an Indian woman, as rich as a Jew, from beyond the sea." In order to carry out the deception, Mrs. Fleece'em has her African diasporic servant, Marianne, play the part of the would-be bride. What ensues is a series of masquerades based on complexion. Because "her complexion will betray her at once," Mrs. Fleece'em stages the meeting between Marianne and Toby in the dark. Toby is further instructed to apply burnt cork to his face and "German blacking" to his eyebrows in order to approximate the "sallower hue" of "the natives of India."

With these machinations set up, the comedy of the scene turns on a moment of linguistic misrecognition. When Toby enters the darkened room, Marianne asks, "Who be dat dere?" Toby then remarks in an aside, "*Dat dere*? one may find out by her tongue she is a foreigner." Ragussis (2010) argues that this moment is a common trope in Georgian drama: the moment when dialect betrays a character's true identity. Later in the scene, when Toby draws up the shades and cries out, "Lord have mercy on me! she is turned all of a sudden as black as a crow!" Ragussis (2010, p. 53) contends that "skin color… confirms and supplements what her tongue has already revealed." I would elaborate this argument even further. Marianne's literary dialect is a signal to the audience. To reverse Ragussis' formation, it confirms for the audience what her skin color has already revealed. Her voice, however, does not disclose her identity to Toby. He already believes her to be a foreigner "from beyond the sea." The joke is that he fails to recognize *t*/*d*-for-*th* substitution ("Dat dere") as indexical of African diasporic vocal culture. As the analysis has shown, this phonological feature

[18] According to Sir Walter Scott's memoirs, Foote got the idea from an actual event involving the politician Charles J. Fox. He was convinced by a woman named Mrs. Phipps that she had connections to a Jamaican heiress – a type of woman that Scott says was known at the time as a "hyæna." Phipps attempted to extort money from Fox in order to provide him access to this heiress, who turned out to be nonexistent. Scott claims that the allusion to this incident was so apparent to the audience at the time that "the laugh was universal as soon as the black woman appeared" (Lockhart, 1838, p. 172).

is the most distributed among African diasporic texts and has the highest significance (by ANOVA and post hoc Tukey tests) for African diasporic dialogue. Moreover, it is one of the phonological features that is present in some eighteenth century examples of African diasporic dialogue like the *Cozeners* and *The Padlock*, but whose frequency increases dramatically throughout the nineteenth century. Thus, Foote is using a feature that is to become iconized but whose status is still nascent.

Prior to her being revealed, Marianne, in fact, uses a number of other marked features. She twice addresses Toby as "Massa," and of the 33 words in her dialogue, 11 are *iss* (for *yes*), which is a variant commonly associated with Caribbean speakers early in the corpus. The comedy, therefore, rests on the linguistic recognition of the audience and the misrecognition of Toby. Furthermore, part of the humor of Toby's reaction upon seeing Marianne, as Herzog (1998, p. 392) argues, is certainly predicated on British anxieties of cross-racial romance. However, it is important to note that Toby flees from Marianne because he believes her to be ghostly punishment visited upon him for forsaking his first love, Betsy Blossom. At the conclusion of the play, he refuses to join the other characters because "he says as how the house is haunted." Interestingly, his superstition plays as an inversion of the later trope, which has African diasporic characters being ridiculed for their mysticism and gullibility.

In many ways, *The Cozeners* is emblematic of the linguistic portrayals of African diasporic speakers in the earlier texts and presages later developments in the literary dialect. On the one hand, it is apparent that some variants have developed associations with African diasporic vocal culture and that those variants figure into routines of mockery. On the other, neither have those associations calcified into the kinds of linguistic stereotypes that are more common later in the nineteenth century, nor is the mockery as vicious as the "Bobilation" propaganda or the minstrelic Shakespearean performance that emerges in the 1820s and 1830s. Even though Toby's lack of linguistic knowledge is comic, that a character would confuse African diasporic and Anglo-Indian accents would be implausible by the middle of the nineteenth century. The angry reaction to the representations of Indian voices in *The Port Admiral* speaks to the increasing delineation among renderings of vocal cultures. The publication of *The Port Admiral* and its critique also coincides with the increased circulation of some of the most reactionary and pejorative burlesques of African diasporic identity, such as *Tregear's Black Jokes*. This concurrence is important. The delineation among vocal cultures depends on changes that occur to the conventions of

representing not only African diasporic dialogue, but also other communities of speakers. The specific differentiation between African diasporic and Indian voices, which so concerns critic of *The Port Admiral* in 1833, is informed by the decreasing frequency of literary dialect features in Indian representations, as much as it is by the increasing frequency of features in African diasporic representations. And just as the changes in African diasporic literary dialect are partly shaped by the Slavery Abolition Act of 1833 and the political debates surrounding its passage, the changes in Indian literary dialect are influenced by another political event that happens a mere two years later: Thomas Babington Macaulay's delivery of his "Minute on Indian Education" and the passage of the English Education Act, a discussion of which opens the next chapter.

Chapter 6
Imagining Indian Voices

6.1 Introduction

The previous chapter tracked the general movement of African diasporic dialogue toward increasing nonstandardization through the nineteenth and into the twentieth century. This chapter traces a similar, but opposing phenomenon: the general movement of Indian dialogue toward increasing standardization. Like the previous chapter, too, this one explores changing ideological currents and their intersections with changing representations of vocal cultures. An important historical inflection point that was discussed at length in reference to African diasporic representations was the passage of the 1833 Abolition Act, and the debate surrounding it. I want to open the analysis of Indian representations by positing an analogous inflection point that occurs almost concurrently: the delivery of Thomas Babington Macaulay's *Minute on Indian Education* on the 2nd of February 1835.

Macaulay went to India in 1834, where he served on The Supreme Council until 1838. This was a period during which the locus of political power in India was shifting from Company to the Crown. Two decades earlier, the passage of The East Indian Company Act established governmental control over territories held by the Company. In 1833, The Government Act of India revoked the Company's trade monopolies and turned it into an administrative entity. Macaulay's *Minute* was delivered in this context of political and economic reorganization.

Baman Das Basu (1922, p. 87) renders one of the *Minute's* oft quoted passages regarding the British goals for English language education in India as follows:

> (1) We must do our best to form a class who may be interpreters between us and the millions whom we govern; a class of persons Indian in blood and colour, but English in taste, in opinions, words, and intellect.

As it circulates in print in the nineteenth century, the quotation is largely the same but where Basu has "words" the older version has "morals" (Broughton, 1864). It is the nineteenth century version that is most frequently cited by scholars like Bhaba (1984) who read Macaulay's declaration as a quintessential expression of the importance of acculturation as a tool of colonial authority. In Bhaba's words (1984, p. 128), Macaulay is advocating for "mimic men" – Indians who are to be intermediaries

between colonial authorities and their subjects, who are to be extensions of colonial power, who are, as he puts it, "to be Anglicized" but "*emphatically* not to be English" (emphasis his).

The swapping of "words" and "morals" in not merely a curious inconsistency; it is telling. Beliefs about the English language and its functions as both the communicative medium for and repository of Western learning, culture, and morality are at the heart of Macaulay's proposal to make English the language of instruction in British East India rather than any of the local vernaculars. In his view, English is the apotheosis of Western linguistic and cultural development:

> (2) It stands as pre-eminent even among the languages of the West. It abounds with works of imagination not inferior to the noblest which Greece has bequeathed us; with models of every species of eloquence; with historical compositions, which, considered merely as narratives, have seldom been surpassed, and which, considered as vehicles of ethical and political instructions, have never been equalled; with just and lively representations of human life and human nature; with the most profound speculations on metaphysics, morals, government, jurisprudence, and trade; with full and correct information respecting every experimental science which tends to preserve the health, to increase the comfort, or to expand the intellect of man. Whoever knows that language, has ready access to all the vast intellectual wealth, which all the wisest nations of the earth have created and hoarded in the course of ninety generations. It may safely be said that the literature now extant in that language is of far greater value than all the literature which three hundred years ago was extant in all the languages of the world together. (Broughton, 1864, p. 3)

This is in stark contrast to his opinion of the "poor and rude" local vernaculars, which he claims "contain neither literary nor scientific information" (Broughton, 1864, p. 3). Thus, "words" and "morality" are inextricably linked. Cultural and social mimicry are predicated on linguistic mimicry. They rest on the imperial subject as speaker of English.

This ideological orientation to language – with its tensions between Anglicization and "Britishness" – informs the literary representations of Indian characters in British fiction during the nineteenth and early twentieth centuries. Although comparatively standard in some ways, the speech of Indian characters is frequently marked by common literary dialect features like *sahib*. In this way, their language can both reflect and rearticulate the figuring of the colonial subject as acculturated into the imperial project but still Other. These renderings are hardly uniform, however. In the statistical overview, the box plots showed that Indian dialogue exhibits the highest interquartile range in both its composite frequencies and diversity indices (§4.4.2 and §4.4.3). The range in the number and types of features that authors use to ventriloquize Indian characters suggests that there is less

conventionalization in representing Indian vocal culture, as well as less consistent imaginings of Indian subjectivities. Rather than contradicting the relationship between linguistic representation and ideologies of Anglicization, these inconsistencies underscore the ideological tensions at the intersections of language, race, and empire.

Finally, these tensions are articulated not only by the Anglicized voices of Indians, but also by the Indianized voices of Britons. Texts like Samuel Foote's play *The Nabob* reflect anxieties about the effects of empire upon domestic Britain, effects that are partly represented through hybridized language. Like the English of their colonized counterparts, the English of colonizers is sometimes portrayed as corrupted through contact. Such corrupted language is often the voice of an equally corrupted morality – linking once again "words" and "morals". Whereas the language of fictional imperial subjects can advertise the perceived moral underpinnings that justify imperialism, similar language in the mouths of colonial functionaries can advertise the fears of moral contagion contracted at the far reaches of empire and posing a threat to the heart of domestic Britain.

Thus, this chapter explores the literary dialect of Indian characters in nineteenth century British novels and plays and compares these patterns of linguistic representation to those of African diasporic and Chinese characters, as well as to those of "Indianized" British characters. The chapter is organized according to the structure established in chapter 5. The chapter begins with a discussion of patterns of coded features in Indian dialogue (§6.2). That is followed by an examination of diachronic trends (§6.3), and the chapter concludes with an analysis of resemblances (§6.4). Each of these sections addresses, in order, the first three research sub-questions (§1.3). Note, too, the computational techniques and conventions are the same as previous chapters. The analysis of feature patterns uses normalized frequency (which is calculated per 1000 words unless otherwise indicated) and deviation of proportions (DP), which is a dispersion measure (§4.2.1). The analysis of diachronic trends uses composite frequency, which is the normalized frequency of all coded features for a text, and a diversity index, which is a measure of complexity (§4.2.2). The examination of resemblances uses cluster analysis (§4.4.4). Additionally, throughout all the sections quantitative analyses are contextualized by artifacts from the imperial archive, like Macaulay's *Minute*, that shed light on attitudes toward English and its diverse speakers. Their combined analysis illustrates how ambivalence regarding the

imperial project in India, evolving notions of race, and contested identities animate the perceptions of linguistic variation.

6.2 Constituents of Indian dialogue

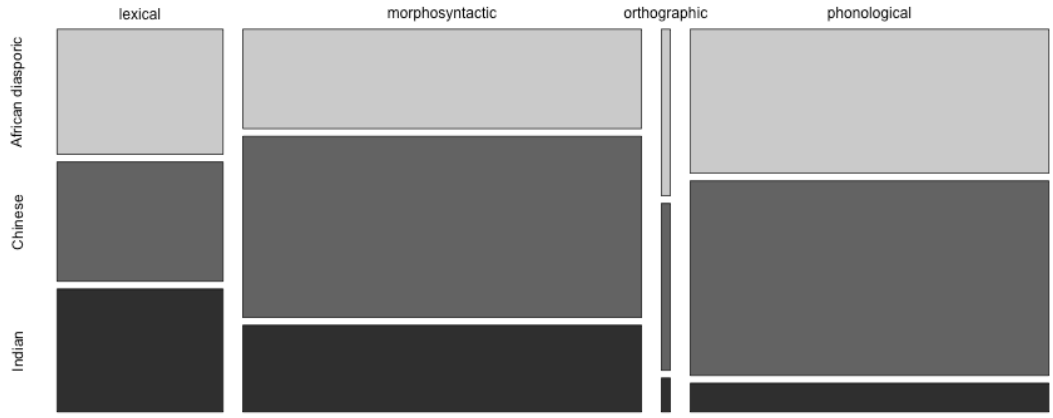
The Indian dialogue sub-corpus is smaller than the African diasporic dialogue sub-corpus – containing 16,639 words from 37 texts (§3.4). The first source work with Indian dialogue is High Kelly's (1774) *The Romance of an Hour*, which is published just six years after *The Padlock*. In spite of their nearly simultaneous emergence, there are many fewer texts in the corpus with Indian dialogue pre-1830 (7) than there are with African diasporic dialogue (25). This discrepancy may be the result of Indian characters, particularly Indian characters voiced in literary dialect, being less conventionally figured into early fictional works. Regardless, the upshot is that the sub-corpus contains fewer words during the early period, but is relatively balanced across the latter two periods (1830-1879 and 1880-1929).

Table 6.1: Frequencies of the four superordinate categories in African diasporic dialogue. *N* is the raw number of occurrences; *% Global* is the percentage a feature or category contributes to all coded features; and *Freq.* is the normalized frequency of a feature or category (per 1000 words).

Feature	N	% Global	Freq.
FEATURES-TYPE			
TOTAL	**3710**		**222.97**
lexical	1148	30.94%	68.99
morphosyntactic	1970	53.10%	118.40
orthographic	17	0.46%	1.02
phonological	575	15.50%	34.56

Table 6.1 shows that the greatest percentage of marking in Indian dialogue occurs in the morphosyntactic category (53%). That is followed by the lexical category (31%) and the phonological category (16%). In order to provide some perspective on these and other distributions, this chapter builds from chapter 5 by including additional comparative analysis. In the previous chapter, log-likelihood comparisons were used to a limited degree to contrast distributions across periods (§5.3). In this chapter, they are used more extensively to contrast distributions across the three speaker sub-corpora. These juxtapositions place the Indian dialogue data in conversation with the African diasporic dialogue data that has been previously analyzed, and they also establish additional foundational context for the chapter on Chinese literary dialect that follows.

Figure 6.1: A mosaic plot showing the relationship between feature categories and speakers based on normalized frequencies.



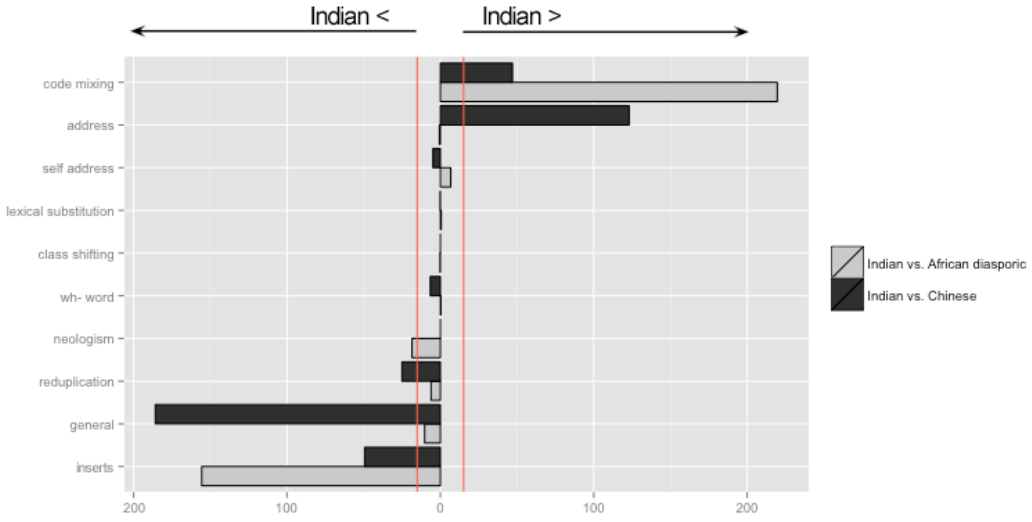
Before moving to statistical comparisons, the discussion begins with a visualization that illustrates how Indian dialogue is constructed very differently from either African diasporic or Chinese dialogue (see Figure 6.1). The plot uses cell volume to visualize the relationships among categorical data. The plot shows, for example, that the lexical cells for the three groups of speakers are roughly the same size. And, indeed, there are no statistically significant differences in the frequencies of lexical features. However, it is clear not only that the volume of the lexical cell for Indian dialogue is larger than the volume of the phonological cell for Indian dialogue, but also that the relationships between the lexical and phonological cells for the other speakers are proportionally very different. Again, these relationships are confirmed by log-likelihood comparisons. There are statistically significant differences in the frequencies of orthographic features in Indian dialogue (LL = 55.85 and 33.03 for comparisons with African diasporic and Chinese dialogue respectively) and morphosyntactic features (LL= 17.38 and 488.38), but these are dwarfed by the differences in phonological features (LL = 1963.04 and 1992.24).

Table 6.2: Frequencies of lexical features in Indian dialogue.

Feature	N	% Global	Freq.	DP
LEXICAL-TYPE				
TOTAL	**1138**	**31.11%**	**68.39**	
address	699	19.11%	42.01	0.39
self address	114	3.12%	6.85	0.53
lexical substitution	43	1.18%	2.58	0.56
code-mixing	168	4.59%	10.10	0.56
general vocabulary	84	2.30%	5.05	0.71
class shifting	11	0.30%	0.66	0.78
inserts	8	0.22%	0.48	0.86
wh- word	9	0.25%	0.54	0.89
reduplication	2	0.05%	0.12	0.97

Not only is the frequency of lexical features in Indian dialogue consistent with frequencies for the other groups of speakers, but also the constituents of the category are similar with a few important exceptions. The most frequent lexical feature is address, which mirrors patterns in African diasporic dialogue (see Table 6.2). Address comprises a relatively high percentage of literary dialect features in Indian dialogue (19%), and the frequency of the category is not significantly different from that in African diasporic dialogue (see Figure 6.2).

Figure 6.2: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for lexical-type features. The red lines mark the points at which $p < 0.0001$.



The broad consistency in lexical feature frequencies across groups of speakers is illustrated by log-likelihood comparisons (see Figure 6.2). Chinese dialogue is notable for its less frequent use of address and greater frequency of general vocabulary (issues that are discussed at greater length in §7.2). Indian dialogue uses significantly fewer inserts than either African diasporic or Chinese dialogue. The only lexical feature that figures more prominently in Indian dialogue than in the dialogue of the other groups is code-mixing. Code-mixing is the second most frequent lexical feature in Indian dialogue after address. Some potential functions of code-mixing are illustrated in the following passage from *Peregrine Pultuney* by John William Kaye (1844):

(3) "Master *Sirdar* – he bearer. Hindoo man – very good cast. He dress master – help make clean – take care master thing – keep key."

"Why, I can do all that myself," returned Peregrine Pultuney.

"No, master, not this country," said Peer Khan, putting his palms together, as natives always do, in respectful remonstrance; "not custom in this country, master. Inglis

> custom do plenty work, – this country custom master do nothing – *Kala-logue* do master's business."
>
> "And who the deuce are *Kala-logue?*" asked Julian Jenks, who had been listening attentively to this dialogue.
>
> "Native this country," returned Peer Khan, "master call *nigger*, I think."

The novel's protagonist, Peregrine Pultuney, and his friend, Julian Jenks, are newly arrived in Calcutta. The passage is part of a longer sequence that narrates their initiation into Anglo-Indian culture by Pultuney's servant, Peer Khan. As in the excerpt, that initiation follows a routine of Peer Khan using non-English terms that he then has to define for his uncomprehending interlocutors. In the excerpt, this happens with *sirdar* (variant of Hindi/Urdu *sardār*), which in this context means something like a valet, and *kala-logue* (variant of Hindi *kālē lōga*), which translates as "black people." The other terms in the longer passage (*bheesty* for water-bearer, *khitmudgar* for waiter, etc.) follow the same format. They are words related to the organization of the Anglo-Indian household, but more than that, they are words that outline the structure of the colonial order.[19] The protagonists – and the readers – are invited into an elaborate hierarchy that is at once foreign and familiar. It is one that involves caste (the *sirdar* being a "Hindoo man – very good cast"), but is racial at its foundation: the *kala-logue*, the "black people" take care of "master's business," while the British masters "do nothing." Too, the idleness of the British masters is positioned as morally excusable. The imperial subjects recognize that in other contexts, the English "do plenty work," but in India their indolence is "this country custom." As was often the case with fictional African diasporic characters, Peer Kahn endorses racial logics that dictate his own subservience. Indeed, the connection between the subjugation of African diasporic and Indian peoples is made explicit with Peer Kahn's final and dysphemic glossing of *kala-logue.*

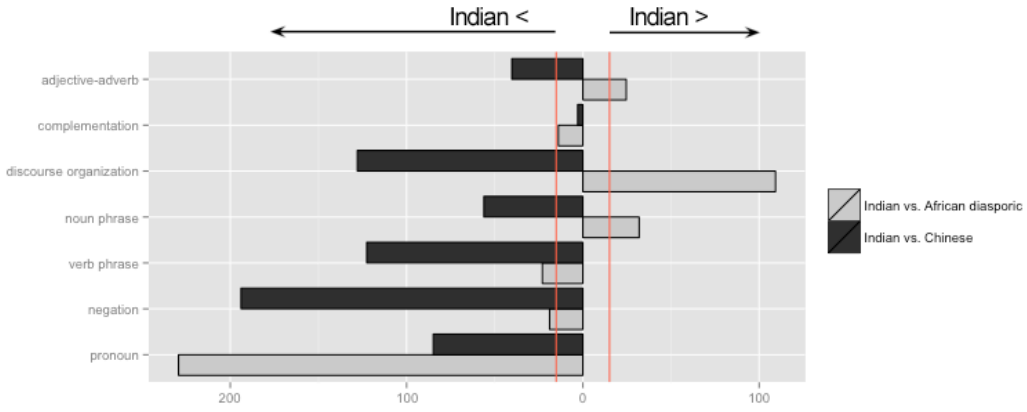
[19] During the nineteenth century, enumerations of the names and roles of servants in Anglo-Indian households were not confined to fiction. They were common features of travel narratives and guidebooks. Kaye's descriptions, for example, echo those provided in *The Hand-Book of India* by the journalist Joachim Hayward Stocqueler (1844), which was published the same year as the novel.

Table 6.3: Frequencies of morphosyntactic subcategories in Indian dialogue.

Feature	N	% Global	Freq.
MORPHOSYNTACTIC-TYPE			
TOTAL	**1928**	**52.71%**	**115.87**
pronoun	146	3.99%	8.77
noun phrase	487	13.31%	29.27
verb phrase	971	26.54%	58.36
adjective-adverb	58	1.59%	3.49
negation	68	1.86%	4.09
complementation	2	0.05%	0.12
discourse organization	196	5.36%	11.78

In its morphosyntactic marking, Indian dialogue diverges more from African diasporic and Chinese dialogue than it does in its lexical marking. In general, Indian dialogue realizes significantly fewer morphosyntactic features across subcategories (see Figure 6.3). There is less pronominal marking in Indian dialogue than in either African diasporic (LL = 229.36) or Chinese dialogue (LL = 84.90). However, discourse organization features are more common in Indian dialogue than they are in African diasporic dialogue (LL = 112.80), though they are more common still in Chinese dialogue (LL = 125.16).

Figure 6.3: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for morphosyntactic-type subcategories. The red lines mark the points at which $p < 0.0001$.



An article titled "Indian Scenes: Shopping" – published in *The Athenaeum* and authored by the travel writer Emma Roberts (1830) – illustrates both the role of discourse organization in the marking of Indian dialogue, as well as how distinctions between morphological and phonological marking can be weighted with social meaning:

(4) The merchant, or Soudargur. – whose name, if a Hindoo, may probably be Sankey Doss, or Dowbalut Sing; if a Mussulman, Maam Bucks, or some such appellation, – salutes the gentleman of the party, with the usual address: "Well, Sahib, what want?

> – all things got!" At this sweeping assurance, some luxurious, or perchance unheard of, article is named. The Baboo, shaking his head, yet nothing daunted, with an indescribable chuckle, replies: "All sold." This generally comprises all his English, except that when you complain of the prices of his goods, he may say, "Much money for freight – Captain very dear – make little profit – very poor man." His words, though few, are seldom, if ever mispronounced; – there is a slight Indian accent; but you never hear a native of Hindostan speak the gibberish which characterizes the African attempts at English. They take the liberty, however, of making considerable alterations in those English words which they have been compelled to adopt, to designate foreign productions – for instance, muffin is invariable called "mufkin"; and dumpling "dumpkin," by the native servants.

The interrogative clause *what want*, for example, contains a null subject, in addition to a null *wh*- auxiliary, which is a verb-phrase-type feature. Roberts construes these and other morphosyntactic features as a form of laconicism that is superior to any potential phonological marking by suggesting that an Indian speaker's "words, though few, are seldom, if ever mispronounced." She goes on to claim that any "Indian accent" is only "slight" and to figure this as evidence of a kind of linguistic and racial superiority compared to "African attempts at English."

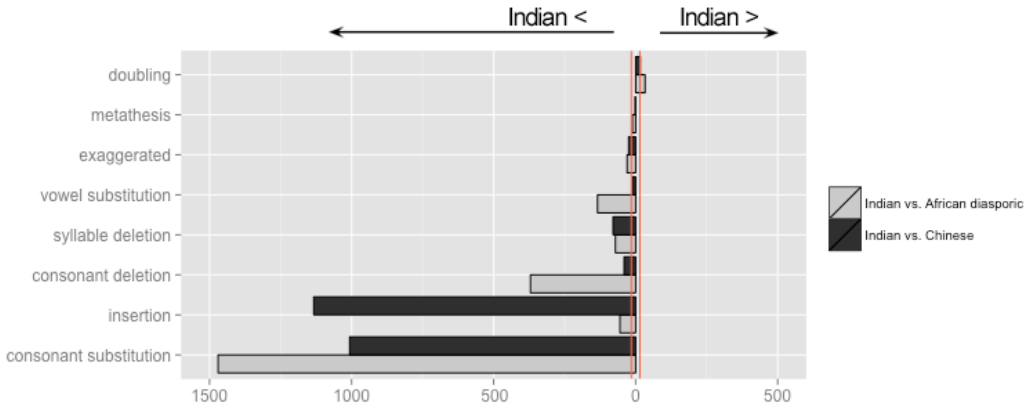
Table 6.4: Frequencies and dispersions of phonological features in Indian dialogue, where DP < 0.80.

Feature	N	% Global	Freq.	DP
PHONOLOGICAL-TYPE				
TOTAL	**575**	**15.72%**	**34.56**	
consonant substitution	253	6.92%	15.21	
consonant deletion	60	1.64%	3.61	
insertion	72	1.97%	4.33	
vowel substitution	113	3.09%	6.79	
syllable deletion	42	1.15%	2.52	0.71
doubling	35	0.96%	2.10	0.79
CONSONANT SUBSTITUTION-TYPE				
TOTAL	**253**	**6.92%**	**15.21**	
t/*d*-for-*th*	184	5.03%	11.06	0.64
CONSONANT DELETION-TYPE				
TOTAL	**60**	**1.64%**	**3.61**	
cluster reduction	25	0.68%	1.50	0.76
FINAL INSERTION-TYPE				
TOTAL	**52**	**1.42%**	**3.13**	
-*ee*/-*y*/-*i* final	24	0.66%	1.44	0.78
VOWEL SUBSTITUTION-TYPE				
TOTAL	**113**	**3.09%**	**6.79**	
ee-for-*y*	42	1.15%	2.52	0.76

The relative infrequency of phonological marking in Indian dialogue is immediately apparent in that data presented in Table 6.4. The table shows only five features with a deviation of proportions less than 0.80. Among those, only one – *t*/*d*-

for-*th* substitution – has a deviation of proportions less than 0.70. That feature is also one that declines over time. A comparison of early to middle period texts yields LL = 31.73 (with the early period being greater and $p < 0.0001$) and middle to late period texts, LL = 58.98 (with the middle period being greater and $p < 0.0001$). The decline is partly attributable to the practice of using literary dialect conventions associated with African diasporic vocal culture in ventriloquizing Indian characters, a practice that was particularly common in works published before 1830, and is discussed at length in a later section (§6.4.1).

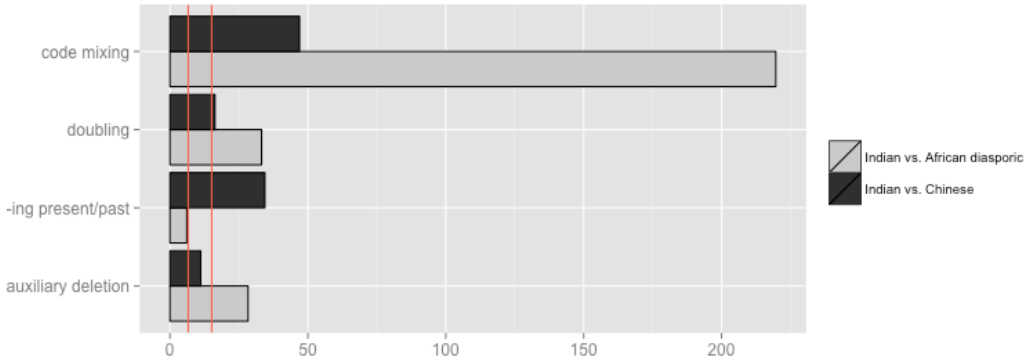
Figure 6.4: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for phonological-type subcategories. The red lines mark the points at which $p < 0.0001$.



Comparisons of the phonological subcategories emphasize just how little marked Indian dialogue is relative to African diasporic and Chinese dialogue (see Figure 6.5). The only phonological subcategory that is more frequently realized in Indian literary dialect is doubling. Doubling in Indian literary dialect appears predominantly in word-final *-t* (63% of all occurrences) or word-final *-s* (23%). It is also largely lexically restricted. All instances of doubled *t* occur in either *thatt* or *whatt*, and most instances of doubled *s* occur in *yess* (except for one instance of *iss* for *is*). Although it may have a significantly higher frequency in Indian dialogue, the feature's relatively high deviation of proportions (0.79) suggests that it is not a widely followed convention. It appears to be one that owes its circulation, however limited, to Rudyard Kipling. Among the source-works, both *thatt* and *yess* first occur in his novel *Kim* (1901). Bithia Mary Croker's (1902) novel *The Cat's Paw* also doubles the final consonants in *thatt* and *yess*, while also doubling the word-final *t* in *whatt*. *A Bottle in the Smoke* by Milne Rae, which is published a decade later, includes doubling exclusively in *thatt*. The phonological motivation of the doubling is not

entirely clear. In the novel *The Taming of the Jungle* (Charles William Doyle, 1899), one of the protagonists, Ram Deen, encounters a mysterious English woman in Northern India and perceives that "her soft 'd's' and 't's' showed that she had been born in India, and that she had spoken Nagari before she acquired English." The phonological quality that Doyle attempts to describe may be similar to what Kipling and his followers are trying to capture orthographically. Both may be efforts to represent adaptations of South Asian phonologies, which include voiceless dental or retroflex stops, to the English voiceless alveolar stop.

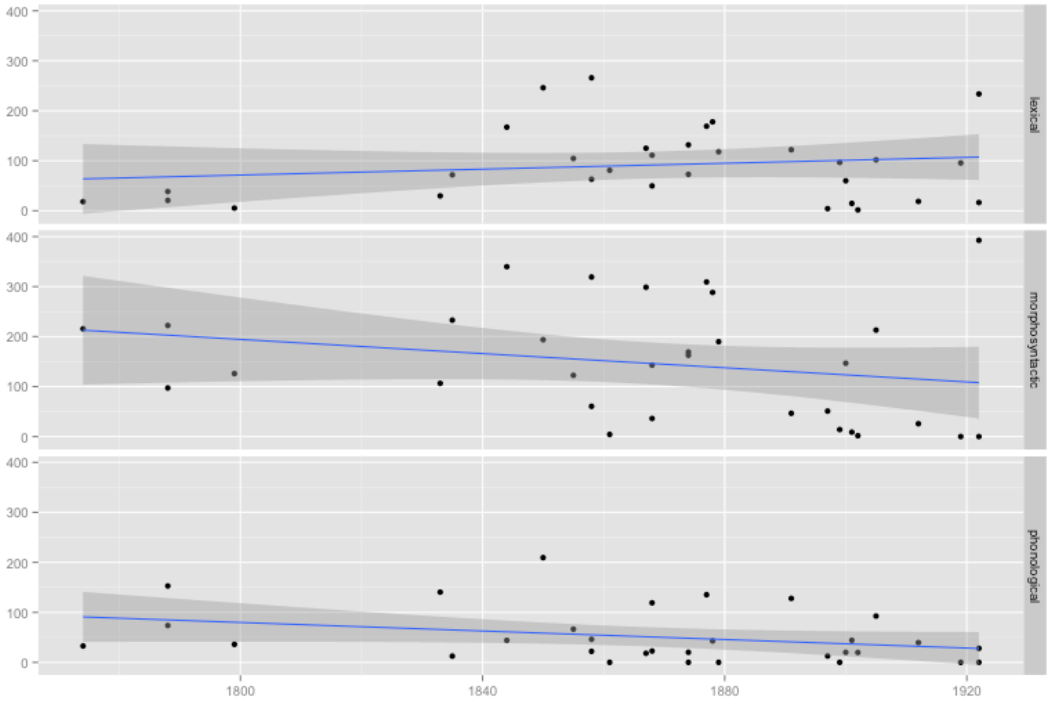
Figure 6.5: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue and between Indian and Chinese dialogue for the features that are more significantly frequent in Indian dialogue. The red lines mark the points at which p <0.001 and p < 0.0001.



Across all four superordinate categories, then, Indian dialogue only realizes four features with significantly higher frequency as compared to both African diasporic and Chinese dialogue: code-mixing, doubling, the present participle used for the present or past tense, and auxiliary deletion (see Figure 6.5). Of those features, code-mixing has a number of unique properties. It is the most dispersed (DP = 0.56 for code-mixing versus DP = 0.79 for doubling, DP = 0.84 for *-ing* present/past, and DP = 0.68 for auxiliary deletion). It is also the only feature that is significantly predictive of Indian dialogue according to an analysis of variance. The ANOVA data presented in the statistical overview shows twenty-two features that have an F-value where $p < 0.01$ (see Figure 4.6). Among those, only one shows $p < 0.01$ in distinguishing both Indian from African diasporic dialogue and Indian from Chinese dialogue according to a post-hoc Tukey's test. That feature is code-mixing. Put simply, it is the distinctive marker of Indian literary dialect, though it was not always that way. Code-mixing emerges as a convention as changing ideologies of race and empire influence the imaginings of fictive Indian identities.

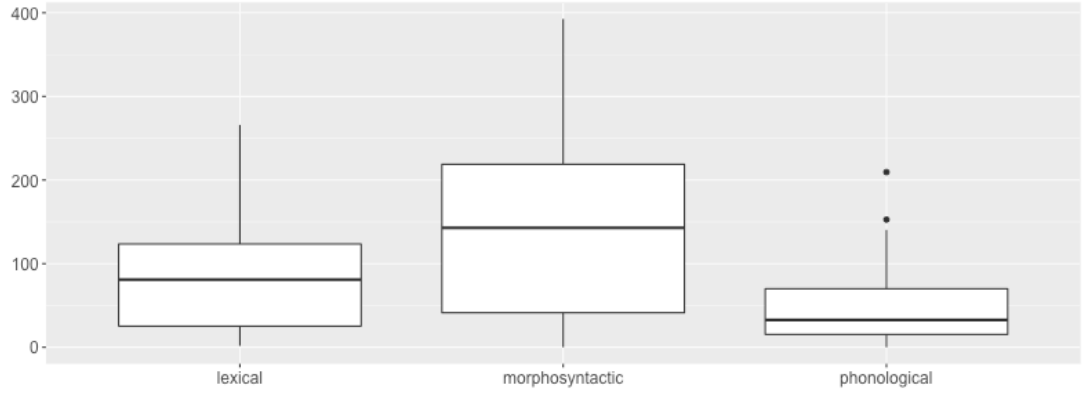
6.3 Diachronic trends in Indian dialogue

Figure 6.6: Scatter plots showing linear trends in frequency for the lexical, morphosyntactic, and phonological categories for Indian dialogue. The grey areas indicate the 95% confidence intervals.



In chapter 4, data were presented to show that there are parallel trends in Indian dialogue toward less frequency and less diversity of literary dialect features (§4.4.2 and §4.4.3). By these measures at least, the speech of Indian characters becomes less marked over time. Separated into the three main categories, the constituent trends tell a somewhat more complex story (see Figure 6.6). While morphosyntactic (β = -0.71) and phonological (β = -0.43) features follow the general trend of decline, lexical features appear to increase (β = 0.29). A complicating factor here is the same one we encountered with the overall trends in Indian dialogue: there are relatively high standard deviations in the data and relatedly relatively low r-squared values. In other words, the large spread of frequencies means that the trend lines are explaining less of the variation in the data. One place that we can see this is in the shaded confidence intervals, which are widest for the morphosyntactic and lexical categories. Boxplots for the categories also show that the frequencies of morphosyntactic features have a particularly wide distribution (see Figure 6.7)

Figure 6.7: Box plots for the lexical, morphosyntactic, and phonological categories for Indian dialogue.

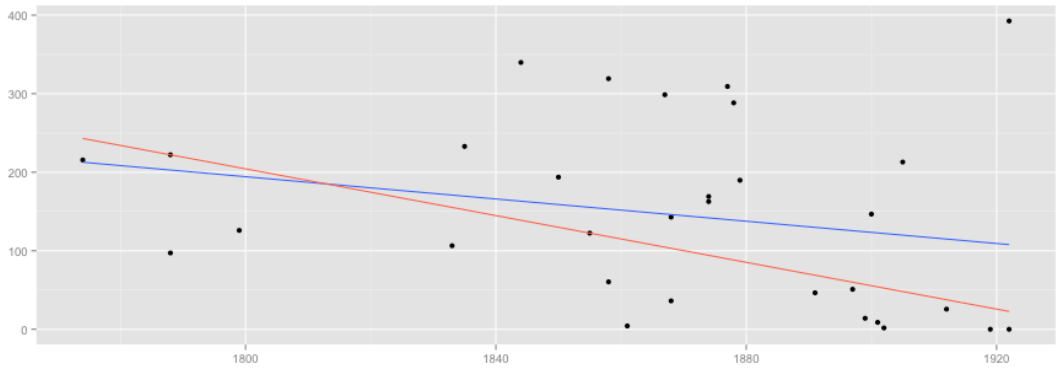


As I have discussed previously, the data are inherently noisy, and a low r-squared is not necessarily a problem (§4.2.3). The purpose of the regression analysis here is not to provide a precise prediction of the next iteration of fictional Indian dialogue, but rather to explain patterns in the extant data. The trend for the phonological category has the highest r-squared of the three categories ($r^2 = 0.097$). The F-statistic for the regression model ($F = 3.13$, $p < 0.1$) shows only marginal significance. However, a correlation measure (Kendall's $\tau = -0.26$, $p < 0.05$) suggests a significant negative correlation between time and the frequency of phonological features, corroborating the negative trend.

For the morphosyntactic category, the linear model appears less robust than it is for the phonological category ($r^2 = 0.059$, $F = 1.83$, $p > 0.1$). Again, however, correlation analysis (Kendall's $\tau = -0.27$, $p < 0.05$) suggests that the decline of features over time is significant. One of the issues with the regression model appears to be an outlier: a high morphosyntactic frequency in the 1922 novel *The Wireless Officer*. One way of mitigating that influence is to apply an alternative method of regression analysis. Quantile regression is one that uses conditional medians rather than means in its estimates, and in so doing is more resistant to outliers (Koenker, 2005). A quantile regression for the morphosyntactic category (see Figure 6.8) yields a steeper negative slope ($\beta = -1.49$) than what is produced using the standard least-squares approach ($\beta = -0.71$). Comparing them is difficult since quantile regression does not produce a measure equivalent to the coefficient of determination. However, we can compare their respective variable p-values ($p = 0.19$ for the standard model and $p = 0.015$ for the quantile model). The quantile regression model, therefore,

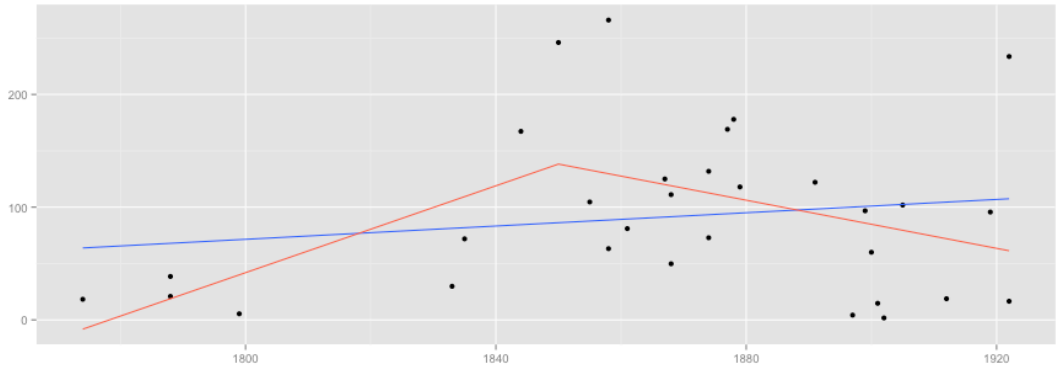
appears to explain the morphosyntactic trends more satisfactorily and confirms the trajectory of decline for the category.

Figure 6.8: Scatter plot showing linear trends over time for the morphosyntactic category for Indian dialogue using two different regression models. The blue line is based on a standard model and the red line on a quantile regression model.



The trajectory for the lexical category appears to be an interesting counterpoint to the other two in its apparent rise rather than fall. However, the standard regression model provides the most questionable results for lexical features ($r^2 = 0.026$, $F = 0.78$, $p > 0.1$). As we did with the morphosyntactic category, we can turn to an alternative. In this case, a segmented regression analysis proves to be a more effective approach (Wagner, Soumerai, Zhang, & Ross-Degnan, 2002). The lexical data partitions nicely into two segments with a breakpoint at 1844 (see Figure 6.9). The first segment has a positive slope ($\beta_1 = 1.95$), and the second has a negative slope ($\beta_2 = -1.10$). The results produce a much more robust coefficient of determination ($r^2 = 0.257$) and more persuasive variable p-values ($p < 0.05$ versus $p > 0.1$).

Figure 6.9: Scatter plot showing linear trends over time for the lexical category for Indian dialogue using two different regression models. The blue line is based on a standard model and the red line on a segmented regression model.



After further analysis, the general trends that are posited in Figure 6.5, for the most part, hold up, though they can be improved with a few adjustments. First,

changes in the morphosyntactic category are better represented by a somewhat steeper decline. Second, changes in the lexical category are more accurately described by a sharper rise into the middle of the nineteenth century, followed by a decline. Of these diachronic trends, I want to spend the remainder of the chapter focusing on two: the broad decline of overall feature frequencies in Indian dialogue and the countervailing trend of greater lexical frequencies into the middle of nineteenth century – lexical frequencies that are particularly marked by code-mixing and the development of an Anglo-Indian vocabulary. These trajectories form key elements in the emerging conventions for representing Indian vocal culture in fiction.

6.4 Resemblances in Indian dialogue

6.4.1 The clustering of early texts and imaginings of the "generic native"

In a preface to the fourth edition of *A Compendious Grammar of the Current Corrupt Dialect of the Jargon Hindostan* (Hadley, 1796, p. v), the author questions the practice of referring to "Hindooee" (i.e., Hindustani or historical Hindi-Urdu) as "Moors":

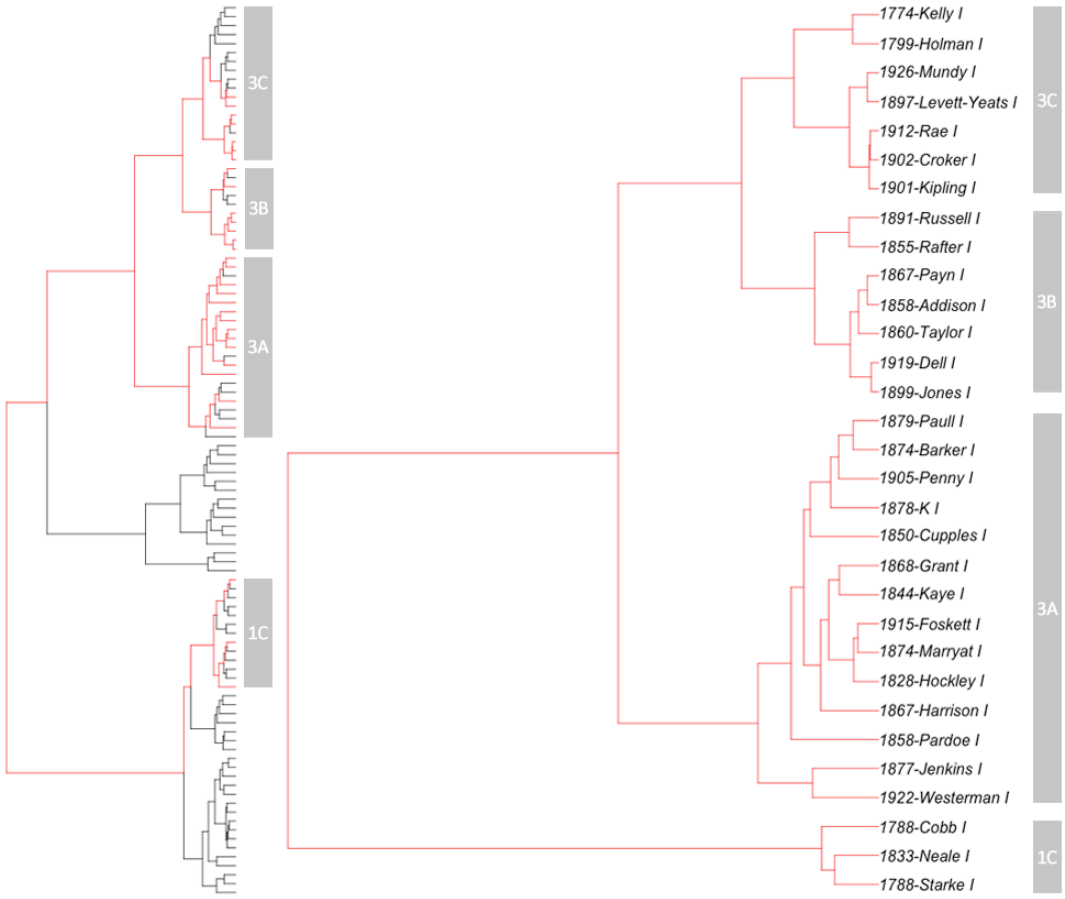
> (5) Why the Hindooee has been called Moors, and the people Moormen is not so easy to decide, unless from the association of that idea with every black person that we see; but they have been so miscalled from the earliest intercourse with India.

"Moor" as social category, Wheeler (2000) observes, was an elastic one in the eighteenth century, figuring Moorish exoticism as deriving from religious difference in some instances and complexion in others. As the above quotation attests, a similar equivocation is at work in the representations of Indians. On the one hand, there is the practice of grouping "every black person that we [the British] see," including the people of India, according to a racist aesthetic of generic blackness. On the other, there is resistance to Indians being "so miscalled," and other distinctions are drawn – some on the basis of class and British cultural accommodation.

These ideological contestations and their relationship to linguistic representation have been introduced in previous chapters. In the analysis of African diasporic dialogue, authors' efforts at differentiating African diasporic from other non-white and nonstandard speaking characters was discussed, with an emphasis on the ways in which such representations figure racist hierarchies. One such earlier example was Neale's (1833) *The Port Admiral*, which was examined in the statistical overview (§4.4.4). The dialogue of the Indian characters Jabbersagee and Jumsagee

was shown to be aligned with conventional voicings of African diasporic characters during the same period – an alignment that was criticized by a commentator who objected specifically to the use of *massa*. Thus, the novel and its detractors exemplify some of the tension between ideologies that position Indian voices and the subjectivities they advertise within a white/non-white binary and those that seek to place them in more elaborated taxonomies. In the dendrogram that is zoomed for Indian dialogue (see Figure 6.10), the dialogue *The Port Admiral* is part of a trifoliate grouping that includes two eighteenth century plays. That grouping is linked to two nineteenth century texts to form cluster 1C. (Note that the numbering of the clusters is the same for all dendrograms, following the template established in Figure 4.15.) Cluster 1C is embedded in the larger cluster (in the complete dendrogram on the left), which consists of primarily African diasporic dialogue.

Figure 6.10: A dendrogram zoomed for Indian dialogue. The numbered clusters on the right match their counterparts from the full dendrogram on the left. Indian texts are highlighted in red.



The four groupings that contain Indian dialogue (clusters 1C, 3A, 3B, and 3C) show some chronological coherence. Eighteenth century texts are located in cluster

1C (two texts published in 1788) and a pairing at the top of 3C (published in 1774 and 1799). Both of these are embedded in larger groupings that are heterogeneous by speaker, as the complete dendrogram on the left suggests. Middle and later nineteenth century texts are located in cluster 3A. In the statistical overview, it was demonstrated that this is the cluster where code-mixing is concentrated. (Refer to the heat map in Figure 4.16.) It is also part of a larger cluster that is more homogeneous than 1C. Cluster 3B is also relatively consistent for time period and speaker, as is the bottom pentafoliate grouping in 3C. The former consists largely of texts published in the mid- to late-nineteenth century, and latter of texts published in the early twentieth century. Furthermore, these two groupings contain ten of the eleven texts with the lowest overall frequencies (the only exception being the African diasporic dialogue from Elizabeth Inchbald's (1805) *To Marry or Not Marry*, which has the tenth lowest overall frequency). Thus, the movement from the bottom to the top of the dendrogram follows the rise of some features (like code-mixing), but a more general decline. It is also a movement that reflects the ideological changes accompanying the emergence of conventionalized representations of Indian speech, evidenced by the homogenizing of some clusters.

Figure 6.11: A subsection of cluster 1C, which includes three early examples of Indian dialogue.

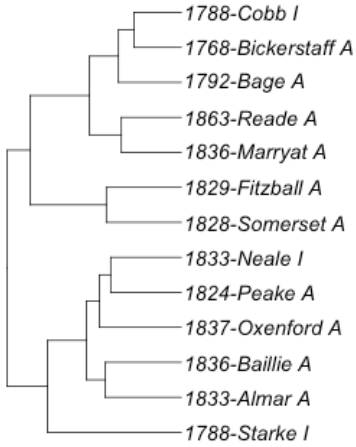


Figure 6.11 shows the larger context for the three early examples of Indian dialogue that are in cluster 1C. The bottom portion of the grouping was presented in the discussion of *The Port Admiral*. Neale's novel is compelling both because of its position on the dendrogram and the criticism of its alignment with contemporaneous African diasporic representations. Expanding that five text grouping by a couple of linkages reveals two eighteenth century works with Indian dialogue: Mariana Starke's

(1788) *The Sword of Peace* and James Cobb's (1788) *Love in the East; or, Adventures of Twelve Hours*. Like *The Port Admiral*, these works occupy locations on the dendrogram that raise questions.

The literary dialect of *Love in the East* has its closest analogue in the voice of Mungo from *The Padlock*. In the introduction, I noted the importance of Mungo as a dramatic and linguistic prototype (§1.2). The name is recycled in a variety of works well into the nineteenth century and is sometimes used, like Sambo, as a referent for any African diasporic male. Mungo also serves as a visual and linguistic touchstone in other media like the prints satirizing Jeremiah Dyson (see Figure 1.3) or Julius Soubise. In light of his enormous popularity, Carson (2007, p. 142) argues that Mungo's invention established a "new, now-comic, model of black masculinity." The pairing of *Love in the East* with *The Padlock* suggests that Cobb imitates Mungo's language in the creation of his Indian character, Rosario. Their similarities, however, do not stop at the linguistic. Rosario and Mungo also serve similar functions in the dramatic structure of the plays. In fact, the proximity of their voicings signals similarly figured identities. Rosario seems partly constructed from Mungo's "model of black masculinity," although he also appears less subversive. Bhattacharya (2006, p. 66) calls Rosario "the stereotypical ethnically unspecific 'native.'"

Love in the East is set in Calcutta, and Rosario, an Indian servant, plays a role similar to Mungo's in the comedy's romantic plots. In *The Padlock*, Mungo is left to guard Leonora, whom Mungo's master, Don Diego, intends to marry. Locking her in his house, Don Diego tells Mungo that no one is to enter. Leander, Don Diego's younger romantic rival, however, plies Mungo with music and wine. Once inside the house, Leander seduces Leonora. When Don Diego returns, a drunk Mungo mocks him: "Make no noise, I say; deres young Gentleman wid young Lady; he play on guitar, and she like him better dan she like you. Fal, lal, lal." Though initially angry, Don Diego eventually approves their marriage, admitting to Leonora, "I only am to blame, who should have consider'd that sixteen and sixty agree ill together." Mungo's role, therefore, is two-fold. On the one hand, his lack of guile in the face of Leander's manipulations confounds Don Diego's romantic designs. On the other, he is the unwitting catalyst to the play's happy resolution.

In *Love in the East*, Rosario is given a letter by Eliza, who at this point in the play is disguised as MacProteus, a Scotsman. Eliza, however, is only acting as an intermediary for Mrs. Mushroom, who wishes to lure Warnford, Rosario's master,

into a secret meeting. When tasked with delivering the letter to Warnford, Rosario mistrusts the mysterious circumstances surrounding the letter and fears that it portends ill:

> (5) Ah! my mind misgive me. – Dis letter be no honest, no say any ting on outside – all white and clean outside – nice and fair, like Missee – afraid though it be wicked and black within. – Poor Massa, why should Rosario give him bad letter? – He be good massa – give me money for my poor father – never say to me rogue – rascal – but always speakee kind, and call my own name.

Instead of bringing the letter to Warnford, Rosario gives it to Colonel Baton, a Frenchman recently arrived from Pondicherry. The misdelivery sets in motion a complex series of machinations that culminate with the elopement of Ormelina and Warnford and the marriage of Eliza and Stanmore.

Like Mungo, Rosario is pivotal to the play's comedic plot and romantic resolution. Also in both plays, that resolution is prompted by a failing – Mungo's failure to guard a house and Rosario's failure to deliver a letter – though there are also clear differences between their characters. In particular, Mungo is, as Miller (2009, p. 28) describes him, "a sassy, back-talking, physically comic slave." He, at least inwardly, resists the cruelty of Don Diego, who calls him a "perverse animal." In contrast, Warnford is portrayed as a kinder master and Rosario as a more tractable servant. Rosario notes in the above quotation, for example, that Warnford "always speakee kind, and call my own name." That said, their many similarities suggest their imbricated identities.

Their shared, imagined "blackness" not only is reflected in their roles in the plays' plots and social hierarchies, but also is advertised in their language. Lexically, like Mungo (as well as the Indian characters in *The Port Admiral*), Rosario uses *massa* as a form of address. Also like Mungo, Rosario uses *poor* in forms of self-address (***poor** Indian*). Phonologically, *t*/*d*-for-*th* substitution is present in the speech of both characters. Finally, the dialogue of both characters contain similar morphosyntactic features: the pronoun *him* as a possessive determiner (***him** damn insurance*), *me* as a clausal subject (***me** very good servant*), zero determiner *(have ∅ great mind to give it to ∅ Frenchman*), preverbal *no* (*he **no** come back*), invariant present tense (*he **walk** about*), and zero copula (*I ∅ so glad*).

The other Indian dialogue in cluster 1C comes from Mariana Starke's (1788) *The Sword of Peace*. As noted in the previous chapter (§5.2), the play participates in proto-abolitionist and anti-East India Company discourses that circulated in the late

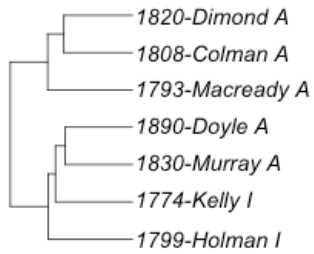
eighteenth century as concerns about the moral implications of imperial expansion grew after the Company-led victory at the Battle of Plassey established British rule in Bengal in 1757 (Moskal, 2000; O'Quinn, 2005). Setting her play in India, Starke critiques the corruption that accompanies imperial wealth largely though the character of the Resident, his underhanded economic dealings, and his sexual predation of Eliza Morton, who is newly arrived from England with her sister. As part of his scheme to seduce Eliza, the Resident coerces a Muslim merchant, Mazinghi Dowza, to imprison his romantic rival. In keeping with Starke's critiques of Company's rule, Mazinghi Dowza is portrayed sympathetically. He offers to provide Eliza with the money to free her suitor from captivity, and her sister, Louisa, describes him as "a charming black soul." As this quotation suggests, Mazinghi Dowza is also routinely marked by his complexion. He is never referred to as "a merchant," but always as "a black merchant." His language reinforces this marking. Like the other Indian dialogue in cluster 1C, it realizes iconized features of African diasporic vocal culture including *massa* as a form of address:

(7) Massa Edwards vas always good and civil – He alvay pay me honest ven he can, I sorry hurt him, good your honor's excellence.

Starke's play, thus, offers up Mazinghi Dowza as a kind of "noble native" – an idealization against which competing British actions, good and bad, are weighed. While he may be imbued with nobility, he is still figured into a racial and linguistic hierarchy. Starke, after all, does not advocate for an overthrow of the imperial order, just a more benign one. At the play's conclusion, the Resident is replaced by the more virtuous Mr. Northcote, resulting in rejoicing among "blacks and whites, masters and slaves, half casts and blue casts, Gentoos and Mussulmen, Hindoos and Bramins, officers and soldiers, sailors and captains."[20] Though embodied anew, British rule remains. Starke's equivocal stance regarding India is underscored by her equivocal stance regarding abolition. Her play asserts the fundamental humanity and liberty of the slave, Caesar (§5.2). Yet, it also has him vowing to Jefferys, his benefactor, to "sarve you faith-ly." One system of hierarchy is replaced by another, supposedly more benevolent one.

[20] *Mussulman* is a term for a Muslim, borrowed from Persian.

Figure 6.12: A subsection of cluster 3C, which includes two early examples of Indian dialogue.



The early texts with Indian dialogue that appear in cluster 3C are similar to those in cluster 1C in that they group primarily with contemporaneous African diasporic dialogue. Figure 6.12 shows the pairing of Hugh Kelly's (1774) *Romance of an Hour* and J. G. Holman's (1799) *The Votary of Wealth* linked to a trifoliate grouping of African diasporic dialogue that includes George Colman's (1808) *The Africans*, William Macready's (1793) *The Irishman in London* and William Murray's (1830) *Obi*. There are two Indian characters in *Romance of an Hour*. The first is Zelida, the daughter of "an Indian Omrah, or nobleman of great authority" and the focus of romantic competition between Brownlow and Colonel Ormsby. The other is Bussora, Zelida's servant, who is described as "so faithful a creature, and [having] a heart as sound as a biscuit." Bussora's voice is rendered in dialect, while Zelida's is not. The standardness of Zelida's speech is, in and of itself, telling. It clearly indexes her aristocratic status and distinguishes her from Bussora. However, it also indexes her mixed ethnicity – her mother is "an English woman," and her father was a man "who lov'd the English extremely."

Culturally, Zelida occupies a liminal space. Her nobility and mixed parentage validate her role as a marriageable prospect for the play's British, aristocratic men. Neither, however, is she fully British. She is similarly positioned between linguistic worlds, as her facility with languages allows her to navigate both European society and London's docks. Among her accomplishments, Sir Hector Stangeways lists her fluency in "English, French, and Italian." His wife, Lady Di, asserts that she speaks these European languages "[l]ike her vernacular tongue." Their son, Orson, adds, "Yes, she has a rare knack at her tongue; and I don't believe that there's ever a foreign merchantman in the whole Thames, but she's able to hail in her own lingo." Connecting her language to the docklands links her to both a lower-class area of London and one that "held either attractive or dangerous exotic connotations" for many British during this period (Fisher, 2011, p. 90). These associations are further

advanced in the play through the nauticalisms of Sir Hector (which are comprised of metaphors like "slacken his sails") and their perceived moral threat. Lady Di urges Sir Hector to "soften the coarseness of [his] phraseology, and use a little less of the quarter-deck dialect."

While Zelida's voice is unmarked from those of British characters, that of her Indian servant, Bussora, is represented in literary dialect. In introducing Bussora, the narrator warns the audience:

(10) His face, perhaps, too swarthy you may find;
"Yet see Othello's visage in his mind –"

In these lines, Kelly evokes Shakespeare and Desdemona's declaration of love for Othello before her father and the duke:

(11) That I did love the Moor to live with him,
My downright violence and storm of fortunes
May trumpet to the world. My heart's subdued
Even to the very quality of my lord:
I saw Othello's visage in his mind,
And to his honour and his valiant parts
Did I my soul and fortunes consecrate. (1.3.248-254)

Kelly urges his audience to see beyond Bussora's "swarthiness" and to recognize his nobility in the same way Desdemona penetrates Othello's physical skin to see the "visage of his mind" upon which is written "his honour and his valiant parts." After hearing Desdemona and Othello speak, the duke finds Othello "far more fair than black," In Kelly's prologue, he asks his audience to come to a similar assessment.

Kelly's Shakespearean appeal also serves to map Othello's Moorish-ness onto Bussora's Indian-ness, similar to the British convention of calling Indians "Moormen" that Hadley puzzles over his preface two decades later (see excerpt 5). Although this mapping may be an unintentional artifact of Kelly's reworking of the original quotation, it fits with the play's conflicted racial logic. When Bussora, for example, comments on Brownlow's good character, he says:

(12) Ah, no, you be too good; me saw you save black man's life, and no plunder in India. Besides, you have behaved like brother to my lady, place her with your own sister, and said oftener, than a thousand times, that there was no sin in have copper complexion.

Like the lines from Hadley, these have concomitant rhetorical effects. In part, these effects are introduced by Bussora's identification of Indians (and by implication himself) as "black." Earlier, Orson similarly describes Zelida as having a singing voice "like the mad negro that died in love for the ale-house girl at Portsmouth." Both of these quotations speak to the tensions in the play regarding Indian identity,

language, and race. In Bussora's quotation, although he identifies himself as "black," he alludes to Zelida's "copper complexion." It is unclear whether is lines are meant to distinguish "copper" from "black," and thus differentiate Bussora's Indian-ness from Zelida's hybridity. Although such a distinction would mirror the representations of their speech, it would contradict Kelly's admonishments in the prologue. On the one hand, Kelly wants his audience to see Bussora's inner nobility. On the other, he crafts Bussora's voice to advertise his place in a linguistic, racial, and social hierarchy. It is this kind of reading that leads Bataille (2000, p. 139) to place Bussora among "the noble savage figure[s] of Restoration and eighteenth-century primitivism" and to explicitly connect Bussora to Alfra Behn's and Thomas Southerne's (notably standard-speaking) Oroonoko.

It is equally possible, however, that Bussora's commentary on Brownlow's conduct toward Indians synonymizes "black" and "copper" and, thus, equates Zelida's racial status with his. This reading would make Kelly's treatment of race, while not radical, at least more ambivalent than Cobb's. When taken together with Orson's description of Zelida's singing, it would situate her as, at once, "black," "copper," "negro," Moor, Indian, and British. As an embodiment of ambiguous racial divisions, she would seem to presage British colonial anxieties voiced later by Macaulay and critiqued by Bhabha.

The equivocality of Kelly's views is reflected in his rendering of Bussora's voice. In the dendrogram, Bussora's dialogue groups with early African diasporic representations. It is also important to remember that texts in cluster 3A are embedded in the larger grouping that is comprised of primarily Indian dialogue. These stand in contrast to the texts in cluster 1C that are embedded in the larger grouping that is comprised primarily of African diasporic dialogue. In light of these distinct embeddings, it is perhaps not surprising that Bussora's dialogue realizes a number of features that are common in representations of African diasporic vocal culture, but exhibits differences, too. It realizes features like *t*/*d*-for-*th* substitution (*wrong* ***ting***, ***de*** *treasure*) and the use of *poor* in self-address (***poor*** *Gentoo*). There is, however, a complete lack of any nonstandard form of address. This absence is likely revealing for at least three reasons: 1) the salience of *massa* as a shibboleth; 2) the timing of play's premier during the height of Mungo's popularity; and 3) Kelly's explicit highlighting of Bussora as a dramatic invention. These factors suggest that Bussora is distanced

from Mungo much like the evidence from *Love in the East* suggests that Rosario is modeled on him.

Bussora's imagined dialect is not alone in its ambivalent situating of the Indian voice – at once linking it to and distancing it from a stereotyped generic "black" identity. J. G. Holman's *The Votary of Wealth*, first staged in 1799, introduces the character of Gangica, described as a "Gentoo." The play follows the scheming of Leonard Visorly (the "votary" of the play's title) and his attempts to defraud Julia Cleveland, a young woman recently arrived from India and heir to her thought-to-be dead father's estate. Gangica is Julia's Indian attendant, and Gangica's imagined dialect is paired with Bussora's on the dendrogram. It realizes many of the same features like *t*/*d*-for-*th* substitution (***dat*** *went out*), zero copula (*he* ∅ *bad man*), invariant *be* (*dat* ***be*** *my reward*), and null determiner in a noun phrase (*I am* ∅ *stranger*). Likewise, her dialogue lacks any nonstandard form of address.

Also like Bussora, Gangica is figured as an honest and loyal servant. When Gangica first appears on stage, Julia describes her to her mother:

> (13) Mother you must love Gangica for my sake; she has left her country and all her relations, because she would not part from me: therefore I must love her better than ever, and every body that loves me, must love Gangica.

Gangica's service to Julia is, thus, framed as more meaningful than her relationship to her own family, country, and culture. This framing orients her similarly to the other Indian characters whose voices are rendered in dialect. Her devotion to Julia speaks to her recognition of her place in a social, cultural, and racial hierarchy. That subordination of self-interest, then, is held up as a sign of her goodness. In describing her own capacity for self-sacrifice in support of her mistress, Gangica tells Sharpset, one of Leonard's former collaborators in his schemes:

> (14) Ay, dat I do – I would die for her. – Oh, I would do great deal more – I would live to bear pain in my limbs, and sorrow in my heart, to make her happy.

Impressed, Sharpset replies:

> (15) Well said, my little disciple of Brama! If the hallowed waves of the Ganges had any share in infusing this gratitude, I wish its stream lay near enough to be resorted to as a fashionable bathing place.

Gangica's altruism is very much in keeping with the models of servitude embodied by Bussora and Rosario. In her case, too, her willingness to make physical, psychological, and social sacrifices is made even more explicit. Her decision to choose a life with her British mistress over one with own "relations" further suggests parallels with Zelida's father. In *Romance of an Hour*, Zelida's father, Abdalla, "lov'd

the English extremely," married a British woman, and "had none of his country superstition on board his mind." He is figured as psychically aligned with British culture. Although he never physically leaves India, his and Gangica's crossings have much in common.

Their commonalities are further underlined by their shared status as romantic subjects of British desire. In Kelly's play, Zelida's father is married to a British woman. In Holman's, Gangica is pursued by Sharpset. The play, in fact, ends with her betrothal to Sharpset, but only at the behest of Julia. When Julia asks if she will consent, Gangica replies, "I do all as you please, ma'am." Her engagement, then, becomes another occasion for the display of her servitude. A review of the play published after its first staging notes approvingly that "we were not surprised to witness her capture the heart of Sharpset" (Wheble, 1799, p. 206).

That approval rests, in part, on Sharpset being a "character by the bye of a very subordinate cast" and, in part, on the reviewer's reading of Gangica as a "character of simplicity" and of "emotions arising out of pure Nature" (206). The description of Gangica in the review points to a tension at the heart of her portrayal: she is noble, "pure," and a suitable recipient for Sharpset's romantic attentions, yet her otherness – racial, cultural, and linguistic – is marked throughout the play. In the above quotation, for example, Sharpset refers to her as "my little disciple of Brama." At other times, she is variously referred to as "my little marigold," "this dusky piece of disinterestedness," "my little Gentoo," and "the fairest mind in a dark coloured case." On the one hand, her outsider status is clear, and her romance crosses a conventional boundary; the review notes the "improbability of [Sharpset's] falling in love with a woman of *her colour*" (emphasis the author's, 205). On the other, her outsider status positions her as an observer and judge of British society – its corruptions and immoralities, in particular – and her nobility ratifies those judgments. Bussora fulfills a similar role in *Romance of an Hour*. He, for example, questions the practice of British gentlemen dueling and facilitates a non-violent resolution to a romantic rivalry. For her part, Gangica casts doubt on European greed and helps to thwart Leonard's schemes. Moreover, she serves as a counterpoint to changing standards of British femininity. When she first responds to Sharpset's flirting, she says, "Oh, you mock – You not like my copper face." He replies, "Why not, my dear? In my mind a lady looks better with a face of copper, than of brass – And that is all

the fashion." In other words, the combination of Gangica's Indian-ness and selfless virtue is framed as preferable to British women's demonstrative vulgarity. Thus, Gangica functions as both a transgressive force and conservative one. She troubles racial boundaries while reinforcing systems of gender and class.

Before moving on from these early representations of Indian vocal culture, I want to call attention to a feature that occurs in the African diasporic dialogue of Caesar and the Indian dialogue of Mazinghi Dowza and Rosario, though not in the dialogue of either Bussora or Gangica: *v*-for-*w*/*wh* substitution. This substitution was discussed in the previous chapter as a phonological feature that has a long history, but appears early in African diasporic dialogue before fading from use. Thus, its presence in these dramas fits with the diachronic pattern. What is additionally interesting, however, is that it also appears in *Love in the East* in the Colonel Baton's speech ("I **vill** pay **vat** I owe"), as well as in Eliza's when she masquerades as the French-accented Baton ("**ven** dey give him little poke"). Although there are other overlaps between Rosario's and Baton's speech (e.g, invariant *be*), this one stands out for at least two reasons. First is the fact that Rosario's dialogue otherwise hews so closely to Mungo's. Second is the feature's established indexicality as a marker of theatrical French accents (e.g., in the in the speech of Dr. Caius in Shakespeare's *The Merry Wives of Windsor*).

The connection between French and Indian voicings in the eighteenth century points to the possibility that early representations of Indian vocal culture comprise a relatively fluid constellation of linguistic features: many drawn from racially paradigmatic models and others from generic notions of nonstandardness. The very first appearance of *v*-for-*w*/*wh* substitution in the corpus, in fact, occurs not in African diasporic dialogue, but in the speech of "a Gentoo woman" who is the wife of Frenchman. Adelaide, in *The Liverpool Prize* by Frederick Pilon (1779), is married to a French General, Monsieur Coromandel. Both are passengers aboard a French East-Indiaman captured by a British privateer. One of the play's plots involves Coromandel using the diamonds he received in marrying Adelaide to procure the consent to marry another woman. The cross-racial aspect of Coromandel's and Adelaide's marriage is consistently flagged (even as that marriage is being subverted) as Adelaide is regularly identified by complexion: a "dark-complexion'd lady," "the dingy lady," "the saffron-faced lady," "his copper coloured wife," and "his Nankin-coloured lady." At one point a character even questions if their marriage is genuine,

whether "this brown woman is really the General's wife?" In spite of these significations, Adelaide's language does not encode racialized difference like Rosario's, but is a version of theatrical French linked with Monsieur Coromandel's. Some of its most prominent features are the same ones that are in Colonel Baton's: invariant *be* and *v*-for-*w*/*wh* substitution ("Me **be** very glad to see you **vit** all my heart"). The most obvious difference between Adelaide's literary dialect and her husband's is his frequent code-mixing ("Now me vill discover von secret – **J'ai caché deux gros diamants**"). Adelaide only approximates that kind of code-mixing in her use of the French *diamants* for *diamonds*.

Because she has so few lines, Adelaide's literary dialect does not appear on the dendrogram. However, hers provides an interesting contrast to those more closely grouped with examples of African diasporic dialogue in cluster 1C, in part, because its influences are relatively transparent. While the literary dialects of characters like Rosario, Jabbersagee, Jumsagee, and Mazinghi Dowza link African diasporic and Indian vocal culture in encoding racialized identities, Adelaide's links French and Indian vocal cultures. The latter is not necessarily any more or less derogatory. Monsieur Coromandel is clearly an unsympathetic character. His name is a reference to the coastal region in India that includes Pondicherry and over which Britain and France competed for control. Only sixteen years before the play's debut, the Treaty of Paris was signed ending The Seven Years War and ceding an important victory to Britain. However, both countries continued to vie for regional dominance for another half-century. The conflict is a source of both humor and patriotism in the play. In the closing line, the captain of the privateer urges his crew (and the audience) to "once more have at the French."

In addition to instantiating the various ways that voices and the identities they advertise can be aligned in earlier texts, Adelaide's literary dialect is an inversion of the pattern that comes to define later representations of Indian vocal culture. Her voice is differentiated from her husband's (and French voices more generally) by an absence: a lack of French vocabulary. Over time, as cultural and linguistic contact between India and Britain expands and as reports of that contact increasingly circulate, voices – Indian and Indianized – become more conventionally marked by "Anglo-Indian" words and phrases. It is a process that arguably reaches its apex with the publication of Henry Yule's and Arthur Coke Burnell's (1886) *Hobson-Jobson* dictionary late in the nineteenth century (see, e.g., Anand, 2011). Code-mixing, then,

develops as an index of Indian vocal culture. What was once marked by absence becomes marked by presence.

6.4.2 The emergence of an Anglo-Indian lexicon and the "colonist style"

The first examples of code-mixing in the corpus occur in *The English in India* by William Browne Hockley (1828). Hockley served in The Bombay Civil Service and was reportedly fluent in Hindi-Urdu, Marathi, and Persian (Prakash, 1994, p. 85). He published a number of novels during the early part of the nineteenth century including *Pandurang Hàrì or, Memoirs of a Hindoo* (1826), *Tales of the Zenana; or, a Nuwab's Leisure Hours* (1827), and *The Vizier's Son or the Adventures of a Mogul* (1831). Gautam Chakravarty (2005, p. 97) observes that, among his works, *The English in India* stands out as an exception. While the others represent indigenous culture, *The English in India* is a "fictionalized social documentary" – an effort, in the author's words, "to pourtray the English in India as they really exist"

The literary dialect in the novel comes from the dialogue of Mahommed Sultaun who is a servant to John Tompkins, the Resident at an up-country station called Kirkpore. The narrator proposes that the interaction between Mahommed Sultaun and John Tompkins exemplify the "influence acquired occasionally by native servants over their European masters," specifically the efforts of Mahommed Sultaun to subvert the Resident's impending marriage in order to "control the household of his master, and to reap from that superintendence the advantages it naturally offered." Their relationship, therefore, appears to be a thinly veiled proxy for the imperial subjugation of India and the ratification of British power.

Hockley's Indian dialogue appears in cluster 3A and is the earliest text in that cluster. The cluster is comprised primarily of Indian dialogue. The heatmap that was presented in chapter 4 (see Figure 4.16) showed that one of the cluster's defining characteristics is its high frequency of code-mixing. Additionally, texts in the cluster tend to realize a moderate number of morphosyntactic features (particularly discourse organization-type features) and fewer phonological features. These constituents are evident in the following excerpt:

> (16) Cassim, that Seymour sahib dubashee, he eat little rice with me last night. He want Fatimah, mistress's ayah, for his wife; – I tell him his sahib give her new bangles, – want her live in his house. Then Cassim too much angry, – say I one lie-man, – say his master laugh at sahib's beard, and very often send little chit to mistress, till her horse ride morning time, – elephant not ride. Gora-wallahs got no sense; – go away

> far off; – Seymour sahib come on horse, – then he and mistress ride off together – same like this morning.

Earlier in this chapter, it was noted that many Anglo-Indian terms that circulate in fiction relate to household organization and taxonomies of servants. Two such terms are present in (16): *dubashee* (which Hockley glosses as *butler*, but elsewhere is used to designate an interpreter) and *gora-wallah* (which is a horse-keeper or groom). Also notable in the excerpt is the use of *sahib*. In fact, *The English in India* is the first source-work in which the form of address occurs. Thus, its use as a convention for voicing Indian characters in fiction emerges concomitantly with the use of code-mixing. Their attendant relationship is perhaps not surprising given that *sahib* itself is a loanword – borrowed into South Asian languages like Hindi, Urdu, Bengali, Gujarati, Marathi, and Punjabi from Arabic and, subsequently, gaining circulation in English during British rule. Together they form the lexical basis for a new style of representing Indian vocal culture and one that is clearly distinct from earlier examples.

In *The English in India*, code-mixing is not confined to the imagined dialects of Indian characters speaking in English. In the same conversation that includes (16), for example, John Tompkins asks Mahommed Sultaun, "[W]hy should they call you *jût wallah*, – if you have spoken the truth?" The term *jût wallah* (which Hockley glosses as *liar*) gets reanalyzed in (16) as *lie-man* (the *-man* nominal suffix standing in for the agentive suffix *-wālā*). Code-mixing is similarly woven into the third-person narration. Among the words and phrases that Hockley includes and glosses with footnotes are *bhoi* (palanquin-bearer), *burrah bebe sahib* (great lady), and *durbar* (an open hall). Sharma (2011, p. 7) refers to this use of code-mixing in nineteenth and twentieth century Anglo-Indian literature as the "colonist style." Some of the implications of this style are discussed in greater detail later in the chapter, specifically in the context of nabobs and Indianized English.

Another aspect of Hockley's representational practices that repeats in later works is his use of Shakespearean English to voice characters speaking in local vernaculars. The convention of representing Hindi, Urdu, and other South Asian languages in a faux-archaic English is not unusual. It occurs, for example, in *Kim* ("Hast thou eaten?") and *The K'haunie Kineh-Walla* ("And thou lovest her?"). By contrast, African diasporic characters are rarely voiced using either faux-archaic or standard English as a proxy for a foreign language. The only example in the source-

works is Quaco in *Hamilton King* (1839). The narrator notes that he "could mutilate" French and Spanish, but Quaco's Spanish is rendered in standard English ("I am a brother amongst you – your captain knows me well!").

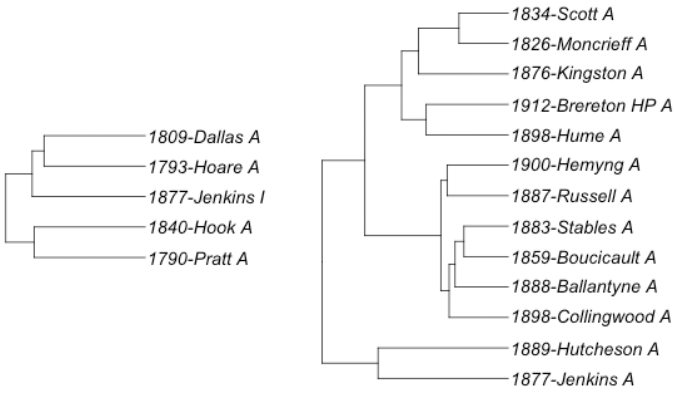
On the stage, Chinese and Indian characters had long been voiced in a Shakespearean style. (Examples of this earlier practice are discussed in the chapter on Chinese dialogue.) However, Hockley's stories are the first in the corpus to double voice characters – to ventriloquize them in a nonstandard variety when speaking English and in a faux-archaic variety when speaking another language. Instances of the latter occur in a conversation among the servants Cassim, Fatimah (17), and Mahommed Sultaun (18):

(17) "Hark ye, Cassim," said she; "if thou hast a grain of reason, thou must acknowledge thyself to be a most pitiful fool to allow Mahommed Sultaun to pick thy brains, as I see plainly enough he has been doing."

(18) "Not so fast, Cassim," said he, "I can tell you, you are not wanted. There is a great dinner to-night, and Fatimah has slipped away. I watched her to the empty bungalow yonder; – Seymour sahib never comes here, and you may guess where he is now."

Some of the potential ideological implications of double voicing are easiest to illustrate using an example that juxtaposes the imagined subjectivities of Indian and African diasporic characters. This happens in a number of works in the corpus, but one that I think is particularly instructive is *Lutchmee and Dilloo* by Edward Jenkins (1877). Before penning the novel, Jenkins (1871) wrote a polemic titled *The Coolie, His Rights and Wrongs*, which condemned the indentured servitude of Indian and Chinese laborers in British Guiana. After not getting the response he wanted, Jenkins decided that in fictional form his argument would gain more traction. The result is *Lutchmee and Dilloo*, a novel whose titular protagonists are an Indian wife and husband who are lured away from their homeland and forced to work on a Caribbean plantation.

Figure 6.13: Clusters containing Indian dialogue (left) and African diasporic dialogue (right) from *Lutchmee and Dilloo* (1878).



In the zoomed dendrogram (see Figure 6.10), Jenkins' Indian dialogue pairs with the Indian dialogue from *The Wireless Officer* in cluster 3A. The complete dendrogram reveals that the dialogue's immediate neighborhood is primarily made up of a number of early examples of African diasporic dialogue (see Figure 4.14). The African diasporic dialogue in Jenkins' novel, however, is paired with Hutcheson's novel *The Black Man's Ghost* and is part of a larger grouping of African diasporic representations largely published later in nineteenth century (see Figure 6.13). Part of what makes this clustering interesting is that the Indian characters are imagined as speaking in Creole. Lutchmee, for example, is described as "adopting the Creole patois of her new acquaintances" when she is queried about the whereabouts of her husband and replies, "No sabby, massa." Their use of Creole as a lingua franca is surely an effort at verisimilitude, but it is also used to figure their dislocation. Jenkins makes this clear by rendering their native language, which the main characters use to communicate with each other, in a formal standard English. When Lutchmee and Dilloo are reunited, Dilloo says to her:

(19) "I rejoice to see you here, my lily, and to clasp you once more in my arms. But this is not the kind of place I had hoped to find when I listened to that cursed recruiter, and came away here in search of riches I shall never win. My poor Lutchmee," he said, stroking her hair with his supple hand, "you know not what you have come to in looking for your lost Dilloo. How unhappy you will be!"

The double voicing aligns the Indian characters with the British ruling class as much as it highlights their being placed in an alien world. Part of Jenkins' political purpose is to depict the injustice of the Indians' indentured servitude, and the standardness of their native language signals their proximity to normative British culture.

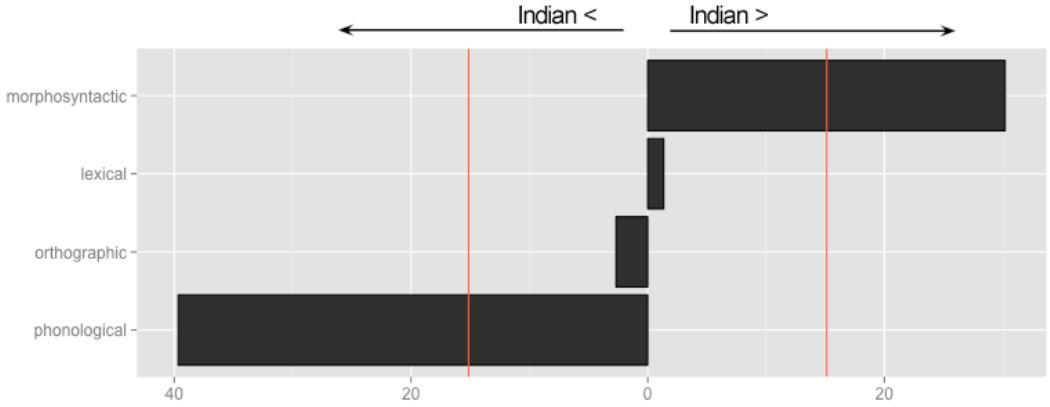
In addition to drawing from earlier stage conventions, the use of standard or archaic English to render Indian languages is also likely influenced by nineteenth century Sanskritists and the Indo-European hypothesis. The hypothesis (forwarded by William Jones and widely debated in the nineteenth century by popular figures like the philologist Max Müller and his friend, the historian and novelist Charles Kingsley) suggested that some Indian and European languages are cognates. One implication of that shared linguistic heritage was a shared ancestral one (see, e.g., Abberley, 2015; Olender, 1992). The sense of a shared linguistic past can engender a sense of linguistic and cultural overlap in the present, as Robert Spence Hardy (1863, p. 20), the General Superintendent of the Wesleyan Mission in South Ceylon, makes clear:

(20) The dark and dreamy Brahman and the pale and practical European, once chased each other under the shade of the same tree, and lived in the same home, and had the same father, and spoke to that father in the same language; and though the difference is now great, both in outward appearance and mental constitution, not more certainly do the answering crevices in the cleft rock tell that they were once united, than the accordant sounds in the speech of the two races tell that they were formerly one people; and this unity is proclaimed every time that they address father or mother, or call for the axe, or name the tree, or point out the star, or utter numbers.

Jenkins' sympathetic stance toward his Indian characters is further highlighted by his differentiation of African diasporic voices from Indian ones, even though the Indian characters are voiced speaking an imagined Creole that is the lingua franca of the novel. In a passage describing the main African diasporic character, Sarcophagus, the narrator characterizes his speech as being an imitation of evangelist preaching, "strongly interlarded with words of many syllables, of which the meaning and fitness were mere matters of chance to him." Further, his linguistic facility is metaphorized as animalistic, and by extension his vocal instrument, his mouth, is a destroyer of words, a pulverizer of meaning:

(21) You could explain more to him by signs than by words. If you tossed him a bundle of words, he used them as a gorilla would use a bundle of sticks. He unaccountably mixed and twisted them up together, he tore them to shreds between his teeth. Some fibres might remain, but they gave dubious testimony of the original form or shape of the communication.

Figure 6.14: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue from *Lutchmee and Diloo* for the four main categories. The red lines mark the points at which $p < 0.0001$.



Based on the descriptions of Sarcophagus' language, we might expect a higher frequency of literary dialect features in his dialogue that in the dialogue of the Indian characters. Indeed, this is the case (674.00 versus 618.36). A log-likelihood comparison of the feature categories shows that the greatest difference occurs in phonological marking, while the Indian dialogue actually realizes significantly more morphosyntactic features (see Figure 6.14). In this way, Jenkins' novel illustrates a number of trends. It serves as an example of the linguistic differentiation between

African diasporic and Indian vocal culture. In late eighteenth and early nineteenth century texts, we saw a variety of examples in which Indian and African diasporic characters are ventriloquized according to shared conventions, encoding shared subjectivities. In Jenkins' novel, we would expect to find similarity, yet he constitutes the literary dialects of his Indian and African diasporic characters distinctly. Thus, his dialogue follows a broader trend toward differentiating Indian vocal culture – though the specific constituents of his Indian dialogue are idiosyncratic in an effort to represent a particular sociolinguistic environment. His African diasporic dialogue (while arguably more viciously exaggerated than others) is not at all idiosyncratic. It is in line with the trend toward increased phonological marking, which was discussed in the previous chapter.

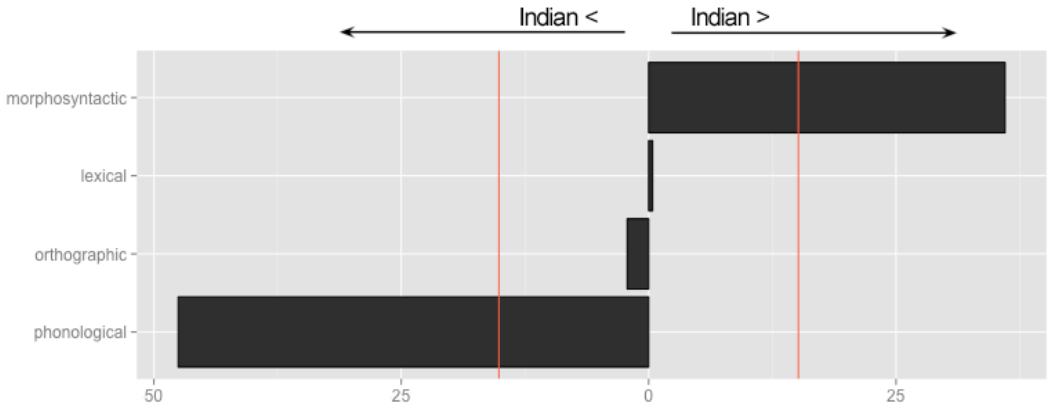
One of the implications of Jenkins' differentiation is an endorsement of a racist hierarchy, advertised through language, that positions Indian communities liminally (between Anglicized and English according to Bhabba's figuration) and that pathologizes African diasporic identities. Jenkins, of course, is not alone in figuring such hierarchies. A similar view is articulated in the excerpt from Emma Roberts' article on shopping in India in which she contrasts the "slight Indian accent" with "the gibberish which characterizes the African attempts at English" (see excerpt 4). Roberts' assessment echoes Edward Terry's seventeenth century travelogue, which I cited in the previous chapter. In it, he describes African language as "inarticulate noise," as akin to the "the clucking of hens, or gabling of turkeys" (1655, p. 16). This description stands in stark contrast to his characterization of Urdu, which he calls "a language which is very significant, and speaks much in few words" (1655, p. 217).

Other source-works construct similar hierarchies – expressed through logics of complexion, language, or both. Consider, for example, the texts that form a pair at the top of cluster 3A (*Levelsie Manor* by Susannah Paull and *With a Stout Heart* by Lucy Barker). In *Levelsie Manor*, Susannah Paull's (1879) British characters explicitly articulate distinctions based on complexion. While a young girl and her mother are waiting for an Indian ayah to arrive by train, the girl asks if it is true that the ayah is "a black woman." "Not black, Gerty, but dark brown," her mother corrects. In *With a Stout Heart* (1874), the author holds up an African diasporic servant, Julius Caesar, to ridicule because of his officiousness with Indian servants. His actions are figured as a comical overreach of the authority conferred upon him not just by his station, but also by his race:

(22) Julius Caesar had not improved in modesty since we knew him before. He bullied the native servants frightfully, treating them like dogs; and though blacker than any of them, was perpetually bringing up their colour against them. It was ludicrous to hear him, with his jetty skin, flat nose, thick lips, and woolly head, addressing the people of the country, with their fine features, and often merely copper-coloured complexions, as, "ugly black niggers."

His body serves to index a position that is presumably subordinate to the Indians he presides over. He is "blacker than any of them" and appears "ludicrous" in comparison to "their fine features." The dysphemism that Julius Caesar uses to address the Indian servants, Barker implies, would be more appropriately directed at himself.

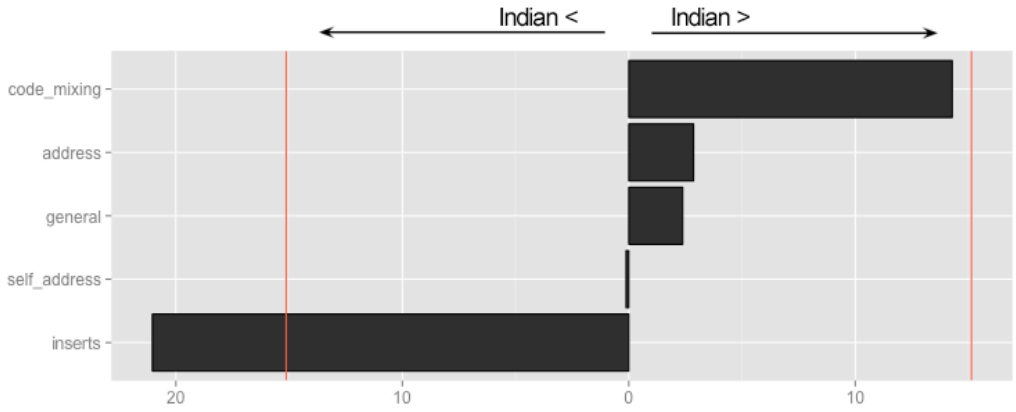
Figure 6.15: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue from *With a Stout Heart* for the four main categories. The red lines mark the points at which $p < 0.0001$.



A log-likelihood comparison of Barker's Indian dialogue to her African diasporic dialogue reveals distributions that are remarkably similar to those in *Lutchmee and Dilloo*, even though Barker imagines a cultural and linguistic environment very different from Jenkins' novel (see Figure 6.15). *With a Stout Heart* is an example of a literary subgenre that emerges in response to the Sepoy Rebellion in 1857. These works often reflect British anxieties about the fragility of imperial rule in India in the wake of the rebellion and articulate beliefs in the moral authority of that rule (see Chakravarty, 2005). A common trope in the figuration of the latter is a loyal servant who protects her or his British household from mutineers. In both fiction and journalistic narratives, this archetypal character is routinely a "faithful ayah" who safeguards her young British charges. In Barker's novel, there are two such characters: an ayah and another servant named Firmall. They are embodied endorsements of the imperial order. Their loyalty is proof of the effectiveness of British administrative policies and a rebuke of mutineers like the tailor who

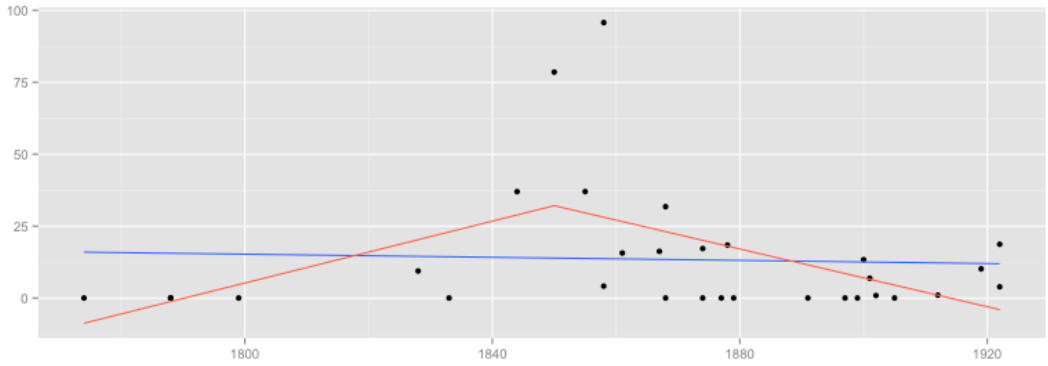
"savagely" warns the wife of the protagonist, Mrs. Dunbar, that "plenty too long have black peoples borne white peoples."

Figure 6.16: Bar plot showing log-likelihood comparisons between Indian and African diasporic dialogue from *With a Stout Heart* for the lexical subcategories. The red lines mark the points at which p < 0.0001.



It is worth noting that as with other dialogue in cluster 3A, code-mixing plays a role in Barker's imaginings of her Indian voices. When the lexical category (which Figure 6.15 shows as being statistically undifferentiated) is broken out into its component features, a comparison demonstrates that code-mixing and inserts are salient in marking Indian and African diasporic dialogue (see Figure 6.16). The fact that code-mixing-type features fall just below the level of highest significance ($p < 0.001$ but $p > 0.0001$) is due to their relatively low number ($n = 6$). This is not unusual for a work published later in nineteenth century. A scatter plot of code-mixing-type features in Indian dialogue reveals a situation similar to what was shown for the lexical category as a whole (see Figure 6.17). Again, the regression is better modeled as segmented rather than simply linear ($r^2 = 0.281$ versus $r^2 = 0.002$), with a rise into the middle of the nineteenth century followed by a decline.

Figure 6.17: Scatter plot showing linear trends over time for code-mixing for Indian dialogue using two different regression models. The blue line is based on a standard model and the red line on a segmented regression model.



Interestingly, even as the frequency of code-mixing generally declines in fictional Indian dialogue, it continues as an ever more prominent index of Anglo-Indian literature, particularly associated with canonical authors like Rudyard Kipling. The power of that association is attested to by parodies like "Burra Murra Boko," a short story comically attributed to "Kippierd Herring" and appearing in the satirical magazine *Punch* (1890). Its defining features are italicized, onomatopoeic respellings that stand in for code-mixing ("the heart within me became as a *Patoph Buttah* under the noon-day sun"). What may appear to be a paradox – declining frequency and rising indexicality – is not a paradox at all. Anglo-Indian vocabulary had been linked just as much, if not more, to "Indianized" British voices – to the "colonist style" – as it had been to Indian voices speaking in English.

We saw this in the first source-work to use code-mixing, Hockley's *The English in India*, which includes Anglo-Indian terms in the narration and the dialogue of British characters. This type of code-mixing actually predates *The English in India*, appearing in eighteenth century works like Samuel Foote's *The Nabob*, which is first performed in 1772. The vocabulary of the titular nabob (his "strange jargon," as another character calls it, exemplified by his use of *jagghire* for an annual income) signals a transformation that is further confirmed by a complexion "tinged by the East." The "tinging" of both his words and his flesh suggests miscegenated cultural values brought on by living abroad and by newfound wealth that affords him power beyond his station. The fear of such moral corruption being returned to domestic Britain is a common theme running through nabobian imagery in the eighteenth and nineteenth centuries (Nechtman, 2010). The nabob in the play *A Cure for the Heart-ache* (Morton, 1797) is unsubtly compared to a viper who "came to the abode of peace and innocence, and disseminated his poison."

Not only are the imperial themes of Foote's play replicated in subsequent works, but so too is the encoding of those anxieties onto bodies and into language. In *The English in India*, Miss Albany jokes that if John Tomkins is appointed Resident, he will develop "a countenance dyed yellow by the united influence of curry, bile, and mulligatawny." Henry Barkley Henderson's (1829, p. 146) collection *The Bengalee: Or, Sketches of Society and Manners in the East* describes the language of an Indigo planter Mr. Neilman as inflected by contact with Indian culture as Miss Albany fears John Tompkins' physiognomy will be:

(23) It is necessary to inform my readers, that my new companion, Mr. Neilman, had adopted, in his phraseology, a most happy, or, at all events, a most unceasing admixture of Hindoostanee aids and expletives. Half his native English had now given way to bad Hindoostanee. Thus he never dines, only *khana-khats*; he never touches wine, it is all *shraub* with him, or rather *beer-shraub*, his only beverage. When he inspects his Indigo fields, he takes a *dékh* at the plant, or *chuls* over the *kates*: he calls Alport his old *doost*; and conversing with his good lady, a little *bat-cheet* with the *beebee-sahib*! Without premising this, it would be difficult to follow Mr. Neilman through his present Eurasian, or Anglo-asiatic illustrations in conversation. But such of my Readers as may find it difficult to keep pace with him, I can safely recommend them to the able expositions of that eminent eastern philologist and linguist, John Borthwick Gilchrist, LL.D. and Author of a very opportune work, – "The Orienti-occidental Tuitionary Pioneer!"

Anxieties that link linguistic mixing to other kinds of sociocultural hybridity are not unique to the context of British Empire in India. An article in *Chamber's Edinburgh Journal* ("Short notes on the West Indies," 1845, p. 4), for example, claims that new arrivals from England are admired for "the purity of their language and pronunciation" because for long-term residents, "even the heads of respectable families are often themselves not free from a 'touch of the negro brogue.'" In one of Sax Rohmer's (1922, p. 169) short stories, "The House of Golden Joss", Ma Lorenzo, a London woman who is "half Portuguese," "catch[es] the infection of that pidgin-English which is a sort of esperanto in all Asiatic quarters" because of her "long association with the Chinese." Fears of linguistic contagion, thus, are attached to a variety of contexts and vocal cultures. However, they appear particularly acute in relation to India, which may partly result from the prominence of nabobs in the domestic imagination. Nechtman (2010, p. 232) argues that attacks against nabobs arose from their perceived role "in bringing South Asians; South Asian animals; South Asian foods, clothing, architectural styles, and *languages* home with them to domestic Britain" (emphasis mine).

Even so, not all depictions of hybridized language are negative. The preface to *The English in India*, Kent (2014, p. 94) contends, sets out to defuse anti-nabob attitudes. Congruent with that position, Hockley's take on John Tomkins' code-mixing appears positive. Tomkins' familiarity with the languages and customs of India helps him to be "[r]espected by the higher class of natives" and "beloved by the lower." An even earlier example of imperial "frontier language," as Lewis (1991, p. 11) terms it, occurs in Phebe Gibbes' (1789) *Hartly House, Calcutta*. Upon her arrival in India, Sophia Goldborne, the novel's narrator, tells her audience that "the European world faded before my eyes, and became orientalised at all points" (Gibbes, 1789, pp. 10-11). Her language, too, undergoes a process of "orientalization," and, like

Hockley, Gibbes affirms this process. Among the "customs of the East" that Sophia Goldborne "adores," is the eschewing of having servants "speaking in broken words." Instead, the colonists "learn to ask for what they want in Gentoo phrases; and making English the vehicle only of polite conversation" (Gibbes, 1789, p. 60). Even Henderson, who mocks the "Anglo-asiatic illustrations" of Mr. Neilman (23), includes a glossary of Anglo-Indian terms in *The Bengalee*, suggesting that his mockery is at least somewhat tongue-in-cheek and self-effacing.

6.5 Conclusion

The conventions of Indian literary dialect follow remarkably different paths in their evolution from those of African diasporic literary dialect. Primarily as the result of changing phonological trends, African diasporic literary dialect becomes increasingly marked over time. In contrast, Indian literary dialect becomes less so. The one caveat to that general decline is a rise in lexical marking into the middle of the nineteenth century – a rise largely driven by the emergence of code-mixing as an enregistered constituent of Indian and Anglo-Indian vocal cultures. That code-mixing is as much a part of the "colonist style" as it is of the fictional speech of Indian characters is important. For even though there is a decline in the frequency of lexical features in Indian dialogue later in nineteenth century, code-mixing continues as widely recognized (and occasionally parodied) hallmark of Anglo-Indian literature into the twentieth century.

In spite of their clear differences, African diasporic and Indian literary dialects have similarities, too. For one, there is no statistical difference in the frequency of address. This is because both groups are typically imagined as servants, and address forms often ratify their position in a social hierarchy. This parallel is perhaps unsurprising in early texts where Indian and African diasporic literary dialects are clustered together in a couple of locations on the dendrogram, a reflection not only of their shared structure but also of what that shared structure implies. In many early works, Indian and African diasporic subjectivities are imagined within a racialized order that groups Indian and African diasporic bodies and voices into one undifferentiated category. That category fairly soon fragments in response to changing racist ideologies. Fictive Indian voices become distinguished by lexicon and by fewer markings – particularly phonological markings – overall. Concomitantly, Indian bodies are positioned within elaborated taxonomies. These interconnected

systems of classification reflect sometimes conflicted notions of kinship and difference, of the perceived efficacy of imperial rule and the illusions of supremacy needed to sustain it.

As a way of navigating such tensions, lexical marking remains a salient index of Indian vocal culture. It can not only delineate West from East, but also encode a matrix of anxieties when those imagined categories are thought to be collapsing. Address, in particular, can signal subordinate alterity even as the demography of fictional India becomes more diverse in later works (populated by characters like the threatening tailor in *With a Stout Heart* or the bureaucratic babus in *A Galahad of the Creeks*, *Kim*, and *The Wireless Officer*). Address is further interesting in that its frequency does not distinguish Indian from African diasporic literary dialect (though their paradigmatic forms, *sahib* and *massa*, do by the middle of the nineteenth century). However, as was presented in chapter 4, address does distinguish Chinese from both Indian and African diasporic dialogue according to an analysis of variance. The lower frequency in Chinese dialogue hints at how differently Chinese identities are imagined, as well as the sociocultural contexts that affect its emergence.

Chapter 7
Imagining Chinese Voices

7.1 Introduction

In the previous two chapters, we have been examining diachronic trends in the use of literary dialect as a representational resource. The first of these explored the increasing frequency of literary dialect features, particularly phonologically motivated respellings, in voicing African diasporic characters. The second looked at an opposing trend: the decreasing occurrence of literary dialect features in voicing Indian characters. In both chapters, there was an identification of inflection points and a discussion of the confluence of political, social, artistic, and commercial factors that helped to shape those transitional periods. Together, the two chapters capture a kind of symmetry, showing how such forces can influence divergent trajectories for two groups of speakers.

As this chapter will show, the history of representing Chinese voices presents an entirely different picture. The conventions of Chinese literary dialect emerge almost a century later than the others. They also appear to be particularly influenced by American authors in California who reacted to the conditions of cultural contact in the region precipitated by the gold rush, which began in the middle of the nineteenth century. Much like some of the African diasporic representations discussed in chapter 5, these images of Chinese voices and identity circulated transatlantically, thus affecting practices in Britain. The political and social conditions in Britain were ripe for taking up the increasingly sinophobic imagery as Britain was engaged in military and economic conflict with China during this same period.

As with the previous chapter, this one follows the outline established in chapter 5, which divides the chapter into three main subsections that address, in order, the first three research sub-questions (§1.3). It begins with an account of the constituent structure of Chinese dialogue (§7.2). That is followed by a discussion of changes over time (§7.3), and the chapter concludes with an examination of resemblances (§7.4). Also as in previous chapters, the statistical analyses use frequencies of features and feature categories that have been normalized per 1000 words (unless otherwise indicated), deviation of proportions (DP) as a dispersion measure (§4.2.1), and hierarchical cluster analysis (§4.4.4). There is, however, one

departure from the chapters analyzing African diasporic and Indian dialogue. In the middle section that explores diachronic changes, the previous chapters have used regression analysis to model trends in the frequency and complexity of literary dialect marking. But unlike the other two sub-corpora, data for Chinese literary dialect are not available for the entire span of the corpus, as was noted in the statistical overview (§4.4.2). Its relatively late development means that modeling change must be done from an alternative perspective.

Rather than seeking to explain the shifting constituents of literary dialect over time, the diachronic analysis tells the story of emergence. Telling that story necessitates the marshaling of alternative kinds of evidence. For the first time since the introduction (§1.4), data from Google Books are included in order to illustrate some shifts in the representations of Chinese people and culture over the course of the nineteenth century. Notwithstanding the integration of such ancillary quantitative data, the diachronic analysis relies more heavily on qualitative data than the previous two chapters. Archival evidence has been used throughout the study, of course, as a means of attending to the fourth research sub-question and contextualizing various quantitative patterns. In this chapter, that evidence takes an even more central role in demonstrating how the emergence of Chinese literary dialect is influenced not only by a shifting imperial landscape, but also by the sociolinguistic history of Chinese Pidgin English and nineteenth century discourses surrounding the variety. When these patterns are taken up by writers of fiction, the result is more consistent representational practices than what we have seen in either African diasporic or Indian dialogue, which is reflected in more coherent clustering on the dendrogram.

There are, however, outliers. These include dialogue from adolescent adventure novels by Frederick Sadleir Brereton and Robert Michael Ballantyne, as well as dialogue from romances and domestic melodramas by Elizabeth Meade, W. Somerset Maugham, and Thomas Burke. Burke's representations of Chinese voices and culture are particularly interesting in their relationship to those produced by his contemporary, Sax Rohmer. The two authors structure their Chinese dialogue differently, and those differences reflect varying sympathies and hostilities toward their Chinese characters and what those characters represent. But there are overlaps, too. Both authors participate in the imagining of the neighborhood of Limehouse as "London's Chinatown." These imaginings recontextualize linguistic variants associated with Chinese vocal culture creating a social and linguistic space that is

connected to older patterns of enregistering Chinese voices, but one that is also creating associations with new fears and fascinations.

7.2 Constituents of Chinese dialogue

Because it has roughly a third of the chronological coverage of the other two sub-corpora, the Chinese dialogue sub-corpus is the smallest, containing 7,971 words. That dialogue was drawn from 39 different source works, very similar to the number of source works that supplied the Indian dialogue sub-corpus (37). The earliest source work with Chinese dialogue is Henry Addison's (1858) *Traits and Stories of Anglo-Indian Life*. That dialogue is assigned to an unnamed character as part of an anecdote. The first dialogue from a named and recurring Chinese character appears a year later in Caroline Leakey's (1859) *The Broad Arrow*.

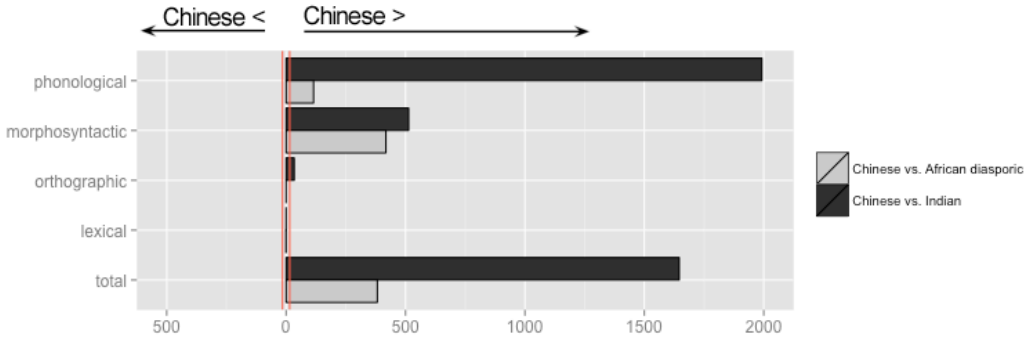
Table 7.1: Frequencies of the four superordinate categories in Chinese dialogue. *N* is the raw number of occurrences; *% Global* is the percentage a feature or category contributes to all coded features; and *Freq.* is the normalized frequency of a feature or category (per 1000 words).

Feature	N	% Global	Freq.
TOTAL	**4366**		**547.74**
lexical	529	12.12%	66.37
morphosyntactic	1931	44.23%	242.25
orthographic	40	0.92%	5.02
phonological	1866	42.74%	234.10

In its distributions of the four superordinate categories, the Chinese dialogue sub-corpus hews closely the patterns we saw in the literary dialect corpus generally (see Table 7.1). Morphosyntactic and phonological features predominate and are roughly balanced (44% and 43%, respectively). Lexical features trail at 12%, and orthographic features account for only 1%. In their frequency, however, these patterns exhibit significant differences when compared to the other two sub-corpora. Given what we have seen in previous chapters, it is not surprising that Chinese dialogue realizes significantly more literary dialect features in the morphosyntactic and phonological categories in comparisons with Indian dialogue (see Figure 7.1). What is perhaps more surprising is that there are significant differences in comparisons with African diasporic dialogue, too, though these differences are smaller particularly among phonological features. In fact, of the three sets of data, Chinese dialogue realizes the most features overall. There are historical reasons for this higher rate of

marking, as well as implications for how texts are grouped on the dendrogram. I will examine both of these issues later in the chapter.

Figure 7.1: Bar plot showing log-likelihood comparisons between Chinese and African diasporic dialogue and between Chinese and Indian dialogue for the four superordinate categories and total composite frequency. The red lines mark the points at which $p < 0.0001$.



As was discussed in the previous chapter, the lexical category is the only one that exhibits no significant statistical differences in overall frequency among the three data sets. However, like Indian dialogue, the constituents of that category are uniquely distributed in Chinese literary dialect (see Table 7.2). The clearest difference is the lower frequency of address in Chinese dialogue ($LL = 123.06$, $p < 0.0001$ in comparisons with Indian dialogue; and $LL = 149.01$, $p < 0.0001$ in comparisons with African diasporic dialogue). As with African diasporic and Indian characters, Chinese characters are often imagined in subservient roles, as cooks and servants, like the "faithful celestial" John Jong in Frederick Brereton's (1912) *Under the Chinese Dragon*, who refers to the protagonist, David Harbor, as *misser*, *masser*, and *excellency*. The very first example of Chinese literary dialect in the corpus, in fact, is notable for its use of address. From Henry Addison's (1858) *Traits and Stories of Anglo-Indian Life*, the story in which the example appears concerns a Chinese miniaturist. In carrying out his task, he uses a compass to meticulously map the pockmarks on the face of the European who commissioned his portrait. The miniaturist utters only one line: "I tell you, massa, I tell you; me measure little holes in massa's face, to put 'em in picture, massa." The anecdote is an indictment of European vanity, as well as a comedy of cross-cultural miscommunication. The dialogue, itself, is remarkable for its use of the iconized address form *massa*. In fact, there are only a few other examples of this specific variant occurring in Chinese dialogue: in Brereton's *The Hero of Panama* ("Allee lighty, **Massa** Jim") and Herbert Strang's *The Flying Boat* ("Yes, sah: **Massa** Leinhadt velly fond smoke").

Table 7.2: Frequencies of lexical features in Chinese dialogue.

Feature	N	% Global	Freq.	DP
MORPHOSYNTACTIC-TYPE				
TOTAL	**529**	**12.12%**	**66.37**	
general vocabulary	209	4.79%	26.22	0.44
lexical substitution	23	0.53%	2.89	0.46
address	128	2.93%	16.06	0.48
self address	76	1.74%	9.53	0.48
reduplication	16	0.37%	2.01	0.62
wh- word	13	0.30%	1.63	0.65
class shifting	5	0.11%	0.63	0.65
inserts	38	0.87%	4.77	0.71
code-mixing	21	0.48%	2.63	0.72

As we will see later in the analysis, Brereton's Chinese dialogue groups with African diasporic representations on the dendrogram, though Strang's does not. The excerpt from *Traits and Stories of Anglo-Indian Life* is too short to analyze using hierarchical clustering; however, its constituents recall the Indian dialogue from *The Port Admiral*, which was critiqued for its use of *massa*. That earlier work was consistent with voicings of African diasporic characters during the period of its publication, and in doing so it positioned the subjectivities of Indian characters as generically non-white. Similarly, Addison's representation – as well as a few others like Brereton's – appear to align, at least partly, with Caldwell's (1971, p. 124) assessment of early Chinese stereotypes in California. He suggests that mid-nineteenth century stereotypes "had an anti-color bias which generalized that all people of color [...] were in one degraded and inferior category." For some Chinese and Indian characters, fictional African diasporic speech functions as the default model for that "one degraded and inferior category," with *massa* a paradigmatic signifier.

Though such linguistic and ideological figurings are apparent in the source works, the significantly lower frequency of address in Chinese dialogue suggests that there are substantive differences in the imaginings of many Chinese characters. I would propose that one factor is the range of subjectivities and roles that are made available to Chinese characters. African diasporic characters are figured almost exclusively as servants or slaves. Indian characters, too, are often imagined as servants. However, they are also depicted as merchants, soldiers, thieves, and nobles. Of course, many such non-servant characters are not voiced in literary dialect.

Chinese characters are accorded a similar, if not broader, range of subjectivities. There is, for example, Mrs. Sweetapple, the "Chinese-Anglified" wife of a British missionary in *China Coast Tales*, or Sin Sin Wa, the opium smuggler in *Dope*.

While the frequency of address lags in Chinese dialogue, Chinese dialogue leads in the frequency of general vocabulary. The significance of these differences is borne out in log-likelihood comparisons (LL = 185.88, $p < 0.0001$ versus Indian dialogue; and LL = 149.13, $p < 0.0001$ versus African diasporic dialogue) and analysis of variance. The ANOVA data from chapter 4 showed that general vocabulary is the only member of the lexical category for which Chinese dialogue leads in frequency and which significantly distinguishes Chinese literary dialect. The most common general vocabulary token in Chinese dialogue is one that occurs in Indian and African diasporic dialogue, as well: *SAVEY* and its variants (e.g., *sabbey*, *sabee*, *savee*, *savvee*), which account for 18% of the subcategory. Among words that are unique to Chinese dialogue, 16% are variants of three reduplications: *CHOP CHOP*, *CHOW CHOW*, and *CHIN CHIN*. Most of these are noted, for example, in a late nineteenth century description of "Pidgin English" that is published in *Pro and Con* (Hamilton, 1872):

> (1) The vocabulary consists of a few words of French origin, such as *savey*, one or two from the Portuguese, many common Chinese expressions, such as *chin-chin*, a salutation, *chop-chop*, for quick, *man-man*, which means stop, *lalilong* man, a thief, with plentiful use of the word *pidgin*, which appears to be applied with the utmost impartiality, to a variety of most incongruous phrases.

Such descriptions, though common in the later part of the century, are not restricted to that period. Earlier examples are more likely to identify the variety as "Canton English" rather than "Pidgin English," as does this one from the historian and lawyer George Wingrove Cooke (1858, p. 59), which circulates in both Britain and North America and presages the description from *Pro and Con* in a number of its specifics:

> (2) The basis of this "Canton English" – which is a tongue and a literature, for there are dictionaries and grammars to elucidate it – consists of turning the "r" into the "l," adding final vowels to every word, and a constant use of "savey" for "know," "talkee" for "speak," "piecey" for "piece," "number one" for "first class," but especially and above all the continental employment of the word "pigeon."

Even earlier, the doctor and author Charles Downing (1838, p. 99) glosses several terms, using *chow chow* as an occasion to mock Chinese culinary culture. He suggests that when the word is "applied to little dogs and tender rats, […] it is spoken with great gusto." In fact, the documentation of many tokens extends at least into the

eighteenth century – the three reduplicative terms appearing, for example, in *A Narrative of the British Embassy to China in the Years 1792, 1793, and 1794* (Anderson, 1795).

Table 7.3: Frequencies of morphosyntactic subcategories in Chinese dialogue.

Feature	N	% Global	Freq.
MORPHOSYNTACTIC-TYPE			
TOTAL	**1931**	**44.23%**	**242.25**
verb phrase	796	18.23%	99.86
noun phrase	391	8.96%	49.05
discourse organization	270	6.18%	33.87
negation	196	4.49%	24.59
pronoun	192	4.40%	24.09
adjective-adverb	82	1.88%	10.29
complementation	4	0.09%	0.50

In contrast to lexical features, Chinese dialogue realizes significantly more morphosyntactic features across all subcategories, with only a couple of exceptions (see Table 7.3). It does not realize more complementation-type features; however, those are so infrequent throughout the corpus that assigning meaning to any distribution is difficult. The more interesting sub-type is the pronoun category, where Chinese dialogue trails African diasporic dialogue by a moderately significant margin ($LL = 6.17$, $p < 0.05$). The most frequent (11.54) and most dispersed (DP = 0.69) pronoun-type feature in Chinese dialogue is *me* as clausal subject ("**me** pilot-man many years on Canton river"). That also holds true for African diasporic (11.45, DP = 0.57) and Indian dialogue (4.39, DP = 0.72). One noteworthy pronoun-type feature that is unique to Chinese dialogue is *my* as a clausal subject ("**My** go longside opium houso"). Though the feature is less frequent than *me* as a clausal subject (5.39), its dispersion is similar (DP = 0.58). It is also one that is noted in the article from *Pro and Con*, which is quoted in (1). In addition to listing vocabulary, the article claims that, in Chinese Pidgin English, "I, me, my and mine, are all expressed by one word, *my*." Interestingly, *my* as an object does appear in the corpus ("What fo' you pinch **my**"); however, there are only five occurrences in three different texts.

The most dispersed morphosyntactic features in Chinese dialogue include the zero determiner and the zero copula (see Table 7.4). These I have discussed elsewhere as common across the corpus (also being the most dispersed morphosyntactic features in Indian dialogue and the third and first most dispersed morphosyntactic

features in African diasporic dialogue, respectively). The last five features listed on Table 7.4 are all related to reduced structures of various kinds. Null particle, null modal, and null *wh*- auxiliary are verb-phrase-type features. Null subject and null preposition are discourse-organization-type features. Of these, all but null *wh*-auxiliary have significant F-values by ANOVA and distinguish Chinese dialogue from African diasporic and Indian dialogue.

Table 7.4: The ten most dispersed morphosyntactic features in Chinese dialogue.

Feature	N	% Global	Freq.	DP
MORPHOSYNTACTIC-TYPE				
zero determiner	257	5.89%	32.24	0.24
zero copula	187	4.28%	23.46	0.27
no preverbal	168	3.85%	21.08	0.29
invariant present	106	2.43%	13.30	0.30
invariant stem	137	3.14%	17.19	0.33
null particle	72	1.65%	9.03	0.34
null subject	133	3.05%	16.69	0.35
null modal	105	2.40%	13.17	0.36
null *wh*- auxiliary	28	0.64%	3.51	0.40
null preposition	65	1.49%	8.15	0.44

The other morphosyntactic feature with a significant F-value and a high dispersion (DP =0.29) is preverbal *no* ("he **no** savee anything"). In nineteenth century descriptions of Chinese Pidgin English, the realization of the feature before the modal *can* is particularly marked ("You **no** can help him"). The German linguist Karl Lentzner, for example, calls *no can do* "a favourite negative" of the variety (Lentzner, 1891, p. 180). Occurrences of preverbal *no* before the modal *can* account for 18% of the total. More tellingly, there are only two occurrences of *cannot* in Chinese dialogue and no occurrences of the contraction *can't*. Furthermore, *no can* appears in only one African diasporic text (Galsworthy's *The Forest*) and two Indian texts (Westerman's *The Wireless Officer* and Rafter's *Percy Blake*).

Nineteenth century reports of Chinese Pidgin English that describe its morphosyntax often explain features affecting discourse organization and elision as a product of language contact. The entry for "Pidgin English" in *Chamber's Encyclopedia* (Chambers & Chambers, 1880, p. 360) notes that its "syntax is usually formed by arranging the words according to the Chinese order," for example. Although this quotation is rather neutral, such descriptions are routinely accompanied by evaluations of the kind we saw in characterizations of African diasporic vocal

culture. This same entry from the encyclopedia is a perfect illustration. It dispassionately observes that "earnest students recognize in [Pidgin English] a new language in embryo, and predict its ultimate status as an accepted tongue, believing that it will be a powerful aid in 'westernizing' China, Japan, and India." Elsewhere, however, it calls Pidgin English a "grotesque form of speech" and a "mongrel dialect" that "def[ies] all known grammar."

Some of the descriptions in *Chamber's Encyclopedia* appear to be informed by a widely circulated article printed in the *Chinese Repository* more than four decades earlier. The *Chinese Repository* was published in Canton by the American Elijah Coleman Bridgman in order to support Protestant missionary activities; thus, the article, "Jargon Spoken at Canton," is presumably intended to inform an audience made up of missionaries and supporters abroad, rather than titillate or shock a domestic audience. Yet, it is more aggressively pejorative than the encyclopedia entry. Like the entry, it suggests Canton English "disregard[s] of all rules of orthography and syntax" (Bridgman, 1836, p. 430). It goes on, however, to claim that it is "an evil," that through its use, "the king's English is murdered" (Bridgman, 1836, p. 433). Further, it makes an explicit connection between Chinese and African diasporic vocal cultures, marking both as "corrupted" and "gibberish":

> (3) The gibberish in use among the negroes in the West Indies, and the corrupted French spoken at the isle of France, resemble this jargon more than any other dialect with which we are acquainted. (Bridgman, 1836, p. 432)

An iteration of the article that appears two years later in a London periodical, *The Penny Illustrated Paper*, emphasizes the language's status as a form of pathological violence. Its words are "grievously mispronounced" and "oddly perverted from their proper meaning," the article proclaims, which results in English "suffer[ing] a mutilation by the tongues of the people of China" (Knight, 1838, p. 190).

Table 7.5: Frequencies of phonological subcategories in Chinese dialogue.

Feature	N	% Global	Freq.	DP
PHONOLOGICAL-TYPE				
TOTAL	**1866**	**42.74%**	**234.10**	
consonant substitution	896	20.52%	112.41	
consonant deletion	84	1.92%	10.54	
insertion	689	15.78%	86.44	
vowel substitution	87	1.99%	10.91	
syllable deletion	96	2.20%	12.04	0.50
exaggerated	11	0.25%	1.38	0.77

The allusion in the quotation to "meaning," of course, pejorates the lexicon of Chinese Pidgin English, as much as it does its morphosyntax. Likewise, the references to pronunciation and "tongues" mark the variety's phonology. The references to phonology are unsurprising, given that the enregisterment of Chinese pronunciation predates the *Chinese Repository* article by at least one hundred years. They also portend the salience of phonological features in fictional representations of Chinese vocal culture that emerge later. In the Chinese dialogue sub-corpus, phonological features occur in frequencies similar to morphosyntactic features (234.10 versus 242.25). Most phonological features fall into either the consonant substitution or insertion subcategories (see Table 7.5). Following the pattern we saw in African diasporic dialogue, in spite of their frequency, the range of consonant substitution-type features with DP ≤ 0.80 is relatively limited. Although 18 different types of consonant substitutions are realized in Chinese dialogue, only 8 have DP ≤ 0.80. The most frequent and dispersed of these, *l*-for-*r* substitution, is among the most indexical features in the corpus, and I discuss that feature shortly. First, however, I want to look at two other substitutions that may not be as immediately associated with Chinese literary dialect: *b*-for-*v*/*f* and *ch*-for-*t*.

Table 7.6: The ten most dispersed phonological features in Chinese dialogue.

Feature	N	% Global	Freq.	DP
PHONOLOGICAL-TYPE				
TOTAL	**1866**	**42.74%**	**234.10**	
l-for-*r*	604	13.83%	75.77	0.29
-ee/*-y*/*-i* final	616	14.11%	77.28	0.32
syllable deletion	96	2.20%	12.04	0.50
b-for-*v*/*f*	36	0.82%	4.52	0.55
cluster reduction	40	0.92%	5.02	0.60
word-final deletion	23	0.53%	2.89	0.61
ch-for-*t*	38	0.87%	4.77	0.67
word-initial deletion	19	0.44%	2.38	0.68
i-for-*e*	7	0.16%	0.88	0.71
s-for-*sh*/*ch*	7	0.16%	0.88	0.74

These are the second and third most distributed consonant substitutions in Chinese dialogue and the fourth and sixth most distributed phonological features (see Figure 7.6). Both exhibit degrees of lexical restrictedness, which affects their indexicalities. For example, *b*-for-*v*/*f* substitution occurs in 13 distinct words, but *hab* is the only variant with a count higher than three. As such, it makes up 58% of all *b*-

for-*v*/*f* substitutions. Of course, the preponderance of *hab* is influenced by the frequency of all variants of *HAVE*. In fact, *hab* accounts for 30% of all *HAVE* variants in Chinese dialogue. This is roughly the same opportunity percentage as *ribber* (33%). Though that variant occurs only once, there are only three occurrences of *RIVER* in any form. The situation for *ch*-for-*t* substitution is even more well-defined. The feature is realized only as a variant of two words: *WANT* and *GOT*. The variant *wanchee* accounts for 68% of the category and 44% of all realizations of *WANT*. The variant *gotchee* accounts for 32% of the category and 27% of all realizations of *GOT*. For these consonant substitutions, it would appear, therefore, that lexically specific forms (*hab*, *wanchee*, and *gotchee*) are particularly indexical, even if their restrictedness is not absolute.

The most frequent and most dispersed consonant substitution in the Chinese dialogue sub-corpus is one that is arguably among the most indexical features in the entire corpus (along with word-final *-ee*/*-y*/*-i* insertions and *t*/*d*-for-*th* substitution): *l*-for-*r* substitution. It has the highest F-value by ANOVA (66.37, $p < 0.0001$). In addition, it is also among the categories (at any level of resolution) that exhibit the most significant differences in comparisons with African diasporic and Indian dialogue: LL = 1734.05 and LL = 1361.83, $p < 0.0001$, versus African diasporic and Indian dialogue, respectively. As noted previously, the Chinese pronunciation of English has a long history of enregisterment, and this is particularly true of the pronunciation of *r*. Like some of the lexical features we saw earlier (e.g., *chow chow*, *chop chop*), *l*-for-*r* substitution has been associated with Chinese speakers of English since at least the eighteenth century. In the *Historia Litteraria*, for example, the Scottish historian Archibald Bower (1732, p. 161) asserts:

> (4) The Chinese pronounce the Words of other Languages according to their own Elements, and change our Letters B D R X Z, which they have not, into P T L S S. Thus instead of Maria, they say *Ma li ya*; instead of Crux *cu lu su*; instead of Spiritus, *su pi li tu su.*

Similarly, the British lexicographer Thomas Dyche (1740, p. 457) suggests in his dictionary's entry for the letter *L*:

> (5) [I]t is remarked of several people, as the *Chinese, &c*. that those words which have *r* in them they cannot pronounce, but change it into *l*, as for *Petrus* they say *Petlus*, *Francis Flancis*, *&c.*

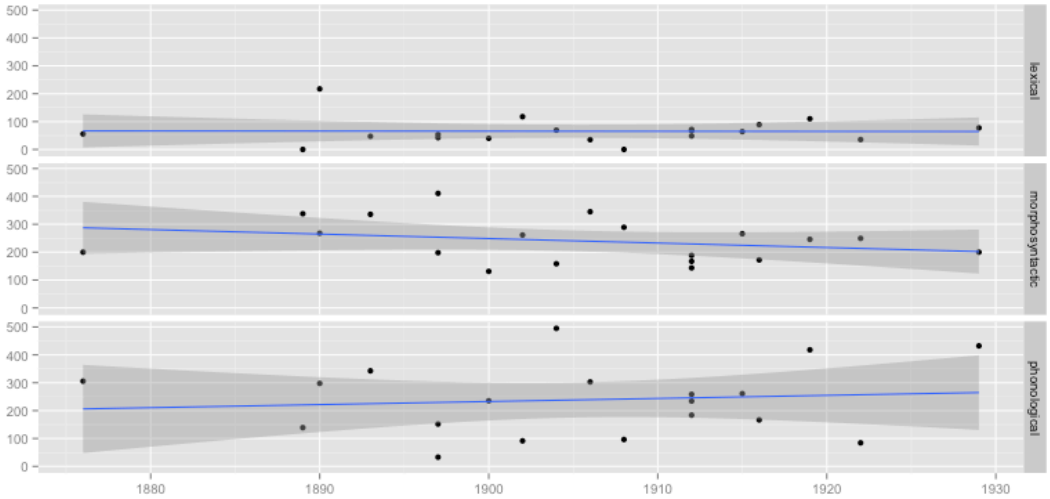
The other highly distributed phonological feature in Chinese dialogue is word-final *-ee*/*-y*/*-i* insertion. Like *l*-for-*r* substitution, it, too, has a highly significant F-value by ANOVA (31.97, $p < 0.0001$) in distinguishing Chinese from African

diasporic and Indian dialogue. I noted in the analysis of African diasporic dialogue that the feature is present in early nineteenth century representations of African diasporic speakers, but largely disappears from those representations later in the century (§5.3). Here, we can see that one force driving that change is just how closely they are associated with Chinese vocal culture during that later period. The feature is equally as frequent (77.28) and almost as dispersed (DP = 0.32) as *l*-for-*r* substitution. Only two texts with African diasporic dialogue realize similar frequencies of the feature: Hofland's (1816) *Matilda, or, The Barbadoes Girl* (82.98) and Trusler's (1793) *Life, or, The Adventures of William Ramble, Esq.* (61.07). And both of those are early works.

Word-final *-ee*/*-y*/*-i* insertions also parallel other frequent features in Chinese literary dialect in that they appear in travel narratives and descriptions of Canton long before being adopted as conventions in fiction. Occurrences in the article from *The Chinese Repository* include *catchee*, *makee*, *muchee*, and *wantchee*. These and their variants are among the most common realizations of word-final *-ee*/*-y*/*-i* insertions in the Chinese dialogue sub-corpus (accounting for 8%, 6%, 6%, and 8%, respectively). Also in the early part of the nineteenth century, Charles Toogood Dowling (1838, p. 280) complains that in hawking their wares, Chinese merchants "drawl out the syllables to unreasonable length." He then exemplifies his complaint with repeated word-final insertions: "What thing-ee you – wantee-shee? Can catchee all same – shele – insectee – fanee?" Similarly, in his book on nautical navigation, Charles Lynn, a commander in the East India Company's naval service, recalls an interaction with Chinese sailors. According to Lynn (1821, p. 148), "using their own broken English," they warn him and his crew about a coming typhoon: "All man talkee Joss too muchee angeree; you too muchee take care." An anonymously published travelogue titled *The Englishman in China* is even more plain in singling out the feature. "The great secret in speaking this dialect," the author writes of Chinese Pidgin English, "is to add *ee* to the end of your words, as, *makee*, *walkee*, *talkee*, *showee*, *singee*" (*The Englishman in China*, 1860, p. 42). That this more explicit description appears later is probably not a coincidence, for *The Englishman in China* is published just as the traditions for representing Chinese vocal culture in Anglophone fiction are undergoing a radical change.

7.3 Diachronic trends in Chinese dialogue

Figure 7.2: Scatter plots showing linear trends in frequency for the lexical, morphosyntactic, and phonological categories for Chinese dialogue. The grey areas indicate the 95% confidence intervals.



The diachronic data for Chinese dialogue is not particularly revealing (see Figure 7.2). The trends for the three most frequent superordinate categories are largely flat. There is a slight decline in morphosyntactic frequencies (β = -1.61) and a slight rise in phonological frequencies (β = 1.11. However, the r-squared calculations suggest that these linear relationships are not particularly explanatory (r^2 = 0.00 for the lexical category, r^2 = 0.07 for the morphosyntactic category, and r^2 = 0.01 for the phonological category). This is, of course, predictable. Because the practice of representing Chinese speakers in nonstandard literary dialect does not begin until relatively late – at least in literature – the span of data is compressed by a hundred years for Chinese dialogue, as compared to African diasporic and Indian dialogue.

For those previous groups of fictional speakers, diachronic changes have been tracked by variations in feature (or feature category) frequency. Why do phonological features increase in African diasporic dialogue? Why do code-mixing features increase in Indian dialogue? And so on. Because the situation for Chinese dialogue is so different, different kinds of questions need to be asked. In lieu of thinking about changes in frequency, we can think about changes in state. Why do representations of Chinese speakers change from a state without literary dialect, a null state, to a state with literary dialect, a positive state? Why does the practice emerge when it does? What are conditions that facilitate its emergence? By its nature, this kind of analysis is more reliant on qualitative data than the diachronic analysis that has been undertaken

up to this point. There is no quantitative data to compare. I will, however, attempt to flesh out the quantitative picture, at least a little, using the Google Books data that I introduced in the first chapter (§1.4).

In the late eighteenth and up through the mid-nineteenth centuries, Chinese characters are typically voiced using a standard variety. The convention arises partly because many early Chinese characters are imagined as speaking Mandarin (or sometimes another dialect), which is then rendered in a standard English – like the character Zamti, a "Mandarine" in Arthur Murphy's (1759) version of *The Orphan of China*:

(6) China is no more; –
The eastern world is lost – this mighty empire
Falls with the universe beneath the stroke
Of savage force – falls from its tow'ring hopes;
For ever, ever fall'n!

Lines such as these recall the depictions of Hindi, Urdu, and other Indian languages discussed in previous chapters. Like those depictions, these representations of Chinese voices encode an ambivalent association with a non-Western imperial culture. "[T]he ambivalence toward the idea of the Chinese empire," Yang (2011, pp. 17-18) argues, "stemmed […] from its affiliations with classical antiquity, and hence its role as cultural mediator between civilized and uncivilized regions of the world." On the one hand, China, "this mighty empire," is connected to idealizations of Greek and Roman culture, and thus to British imperial culture, which is their imagined heir. By implication, Chinese voices and British ones share a common lineage. On the other hand, China is also a part of, or in the above excerpt synonymous with, the "eastern world." As such, it is differentiated from the "western world," its culture, and its legacy. It is constructed, as Yang says, as a borderland.

This figuring of China is made explicit in the poet William Whitehead's prologue to Murphy's play. Whitehead informs the audience that the drama "boldly bears Confucius' morals to Britannia's ears." Although these "fresh virtues" come from "eastern realms," the audience will recognize in them themes "echoing Greece." According to Whitehead, however, China is a deficient exemplar of empire, whose flaws are presented for the edification of Britons. China's fall at the hands of the Tartars, which is referenced in (2), is interpreted as the result of the deification of its royalty. The British, however, can be more secure in their imperial ambitions because "[f]rom nobler motives our allegiance springs."

The rendering of Chinese voices in standard English affirms the fundamental nobility of its people and culture, while its exoticism is often advertised through other means (manners, customs, dress, etc.). Such voicings remain the convention into nineteenth century. Even as more aggressively derogatory stereotypes emerge related to Chinese language and culture, the most indexical features of Chinese literary dialect (*l*-for *r* substitution and *-ee* final insertion) appear infrequently in Anglophone literature prior to 1860.

The transition toward more sinophobic representations is evident, for example, in James Planché's (1848) comedy *The King of the Peacocks*, which premiered in London the day after Christmas in 1848. The play uses a number of tropes that commonly co-occur with literary dialect in later renderings of Chinese characters. For one, the play uses "John Chinaman" as a generic identifier. "Chinaman" as a racialized term originates in and gains frequency through much of the nineteenth century. The play also makes reference to "Chinese people eat[ing] 'bow wow.'" The depiction of imagined Chinese culinary customs, particularly the eating of dogs and rats, is a common way of figuring a combination of otherness and deviance. Finally, the play alludes to "Chinee lingo" and has a French character (Soyez Tranquille, a chef) who speaks in literary dialect. The voice of the play's Chinese character (Poo-lee-ha-lee, the captain of a junk), however, is rendered not in any fictional "Chinee lingo," but in a nautically inflected English ("Avast, there ma'am").

Both *Traits and Stories of Anglo-Indian Life* and *The King of the Peacocks* are transitional texts, works that anticipate the emergence of new conventions for representing Chinese vocal culture. An early North American example of those conventions appears in *A Live Woman in the Mines* by Alonzo Delano (1857), which is published one year before Addison's collection of stories. The play is set in the California frontier during the gold rush – Delano, himself, having spent time as a prospector. In the play, the Chinese character has four lines in which he warns of an Indian attack:

(7) Chinaman. Me help! Me help! Shooty me! Bang me shooty! One, tree, five hundred Indian! O! O! O!

Pike. Shoot you, bang you, two or three hundred Indians? What the devil do you want with so many Indians?

Chinaman. No, no, no! Pop! Bang! Bullet shooty me!

Old Swamp. Indians shoot you?

> Chinaman. Gold prospect, me hill over. Par one dol,ar [*sic*] – one dollar, two bit – one dollar half. Indian come! Me bang! Bang! Bullet! Pop me! Two, tree, five hundred!
>
> […]
>
> Old Swamp. The Diggers are upon us, boys – let's meet them on the hill and surprise them
>
> Pike. And lick them before they have a chance to scalp Short-Tail. [All rush out, except CHINAMAN, with a "Huzzah!"]
>
> Chinaman. Chinaman no fight; Chinaman skin good skin; keep him so. Mellican man big devil – no hurty bullet him.

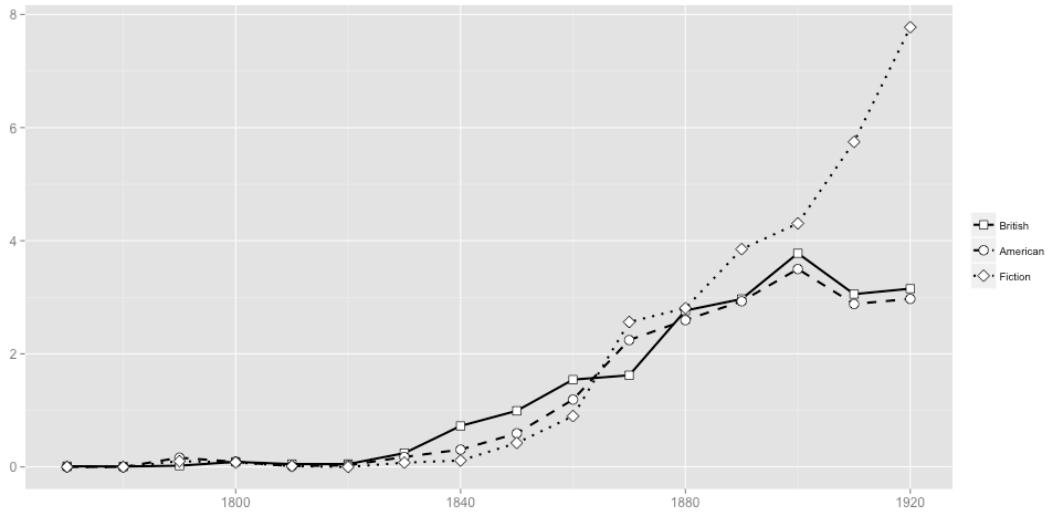
As in the anecdote from *Traits and Stories of Anglo-Indian Life*, the scene's comedy is predicated on miscommunication. In Addison's work, the miscommunication is cultural; the Chinese portraitist does not understand the European's expectations. In the scene from Delano's play, the miscommunication is linguistic; the Chinese prospector has difficulty in communicating his warning. Both works also identify their Chinese characters only as "Chinaman." Addison actually glosses the term as regionally specific, "as [what] we call him in Bengal."

The use of this identifier is not incidental. The invention of "Chinaman" as a racialized subjectivity and the ventriloquizing of that subjectivity using literary dialect appear to go hand in hand. One way to demonstrate this relationship is simply to note that every source work that contains a Chinese character voiced in literary dialect also contains the lemmatized token *CHINAMAN* except for one. *The Happy Adventurers* (Middleton, 1922) has a character named Ah Kew, who is also a Chinese servant; he speaks only twenty nine words.

The diachronic trajectories for the token are additionally suggestive. In order to illustrate these trends, I want to return to the data from Google Books that I introduced in the first chapter and connect it to the data from the source works. The first occurrences of *CHINAMAN* in the source works are in the 1850s, which coincide with the beginning of the rise in the Google Books data for English fiction (see Figure 7.3). Additionally, 91% of the 1238 occurrences of *CHINAMAN* in the source works appear after 1890, the same period that shows increasing frequency for fiction in Figure 7.2. The trends in the Google Books data and the source works, in fact, are highly correlated (Kendall's $\tau = 0.81$, $p < 0.0001$), even though there is a selection bias in favor of the token in the course works. Of course, all of the same caveats apply to the Google Books data that were discussed earlier. Nonetheless, the triangulation with the data from the source works points to a provocative relationship between the

increasing frequency of *CHINAMAN* and the evolving practice of ventriloquizing Chinese characters using literary dialect.

Figure 7.3: Frequencies (normalized per million words) of lemmatized *CHINAMAN* in the Google Books data tables from 1770-1930.



In addition to their shared use of the term *CHINAMAN*, British works like *Traits and Stories of Anglo-Indian Life* and American ones like *A Live Woman in the Mines* are instructive for their settings. The earliest examples of Chinese literary dialect in the corpus tend to be from works that are set outside of domestic Britain, out in the empire. The anecdote of the miniaturist in Addison's collection takes place in Agra, India. Another early example, *The Broad Arrow* by Caroline Leakey (1859), is set in Port Arthur, Australia. The novel's protagonist, Maida, is wrongfully convicted of murdering her child and is sent to a convict colony in Tasmania. The novel includes a Chinese character named Opal, who is the servant to a "convict mistress," Mrs. Evelyn. Opal's dialogue is the first in the corpus to include *l*-for-*r* substitutions ("All **light** den – Opal **welly** glad"), though it does not have any word-final *–ee*/*-y*/*-i* insertions. Opal is described as a "Chinese worshipper" of Mrs. Evelyn, who, in his words, "luff dat plitty light laddie vely much."

These early examples conform to the models of Anglo-Indian literature that are being published during this same period. Novels like *Peregrine Pultuney* (1844), which was discussed in the previous chapter, were influenced by the changing conditions in India – the contact and conflict that shaped British desires and anxieties regarding its empire. Similar changes affected Sino-British relations. Britain and China were engaged in a series of conflicts over access – economic and religious – to China's markets. The First Opium War ended in 1842 with the signing of the Treaty

of Nanjing (Brook & Wakabayashi, 2000). The treaty ceded control of Hong Kong to Great Britain, and established ports for foreign trade in Amoy, Canton, Foochow, Ningpo, and Shanghai, effectively ending the earlier Canton factory system (Van Dyke, 2005). In an effort to further broaden its trade interests, Britain sought to renegotiate the Treaty of Nanking in the mid-1850s. This eventually led to the Arrow War, which ended with the ratification of the Treaty of Tianjin in 1860 (Wong, 1998). That treaty expanded British control of Hong Kong to the Kowloon peninsula and established the rights of Christian missionaries to proselytize in China (Munn, 2013). Missionary activity is one of the factors that motivated a resistance led by the Yihetuan (or the "Boxers") at the end of the nineteenth century – an event that figures in a number of source works, most prominently in Henry Charles Moore's (1906) *Afloat on the Dogger Bank*.

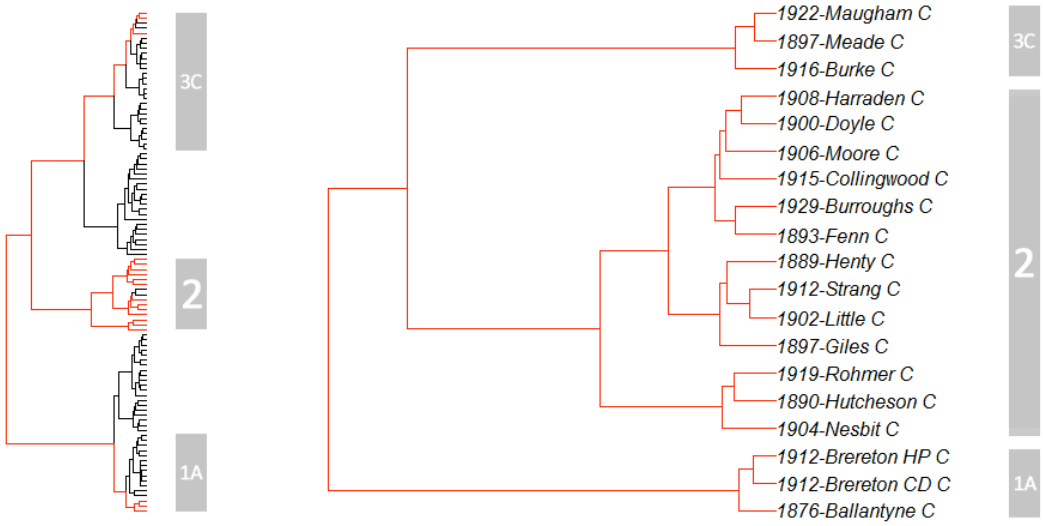
The political, military, and economic interactions between China and Britain inform not just British, but also American imaginings in the middle of the century. For example, a review of *Yankees in China, or a Union of the Flags* describes the play as "founded on the present quarrel between the British and Chinese" (Snowden, Sigourney, & Embury, 1840, p. 208). Fairly quickly, however, the contact that animates American representations of Chinese people and culture is domestic. Accordingly, the locations for such imaginaries become national rather than international, with San Francisco and its environs serving as an important site for both invention and circulation.

The setting for *A Live Woman in the Mines* is, therefore, an instructive counterpoint to the settings of *Traits and Stories of Anglo-Indian Life* and *The Broad Arrow*. While the latter British works take place out in the empire, Delano's American play is set in California. That location is informed as much by demographic changes that occurred in American West as its British counterparts were by imperial economics and politics in Asia. In the middle of the nineteenth century, there was rising Chinese immigration into the western United States, and particularly California. Coolidge, in a demographic study from early in the twentieth century, claims that the West Coast Chinese population stood at 7,370 in 1851 and reached 132,300 in 1882. Chen (2000) argues that Coolidge's figures are largely accurate, though they may, in fact, understate the case. Citing federal census data, he calculates that the Chinese population in San Francisco alone went from 2,719 in 1860 to 12,022 by 1870. These demographic changes brought Chinese and European-American cultures into

increasing contact. That contact fostered xenophobia, which in some of its more extreme expressions feared an overthrow of the European-American social order brought on by waves of Chinese immigration. The "yellow peril" discourse of this period augurs the passage not only of the 1882 Chinese Exclusion Act in the United States, but also of the 1901 Immigration Restriction Act in Australia and the 1923 Chinese Immigration Act in Canada.[21]

7.4 Resemblances in Chinese dialogue

Figure 7.4: A dendrogram zoomed for Chinese dialogue. The numbered clusters on the right match their counterparts from the full dendrogram on the left. Chinese texts are highlighted in red.



All of these histories (sociolinguistic, economic, military, cultural, etc.) influence the patterns of clustering we see on the dendrogram (see Figure 7.4). Most obviously, the clustering of Chinese dialogue is more consistent than either African diasporic or Indian dialogue. Most texts are situated in cluster 2. The others aggregate into two trifoliate groups at the ends of the dendrogram. The reasons for this relative consistency are several. First, the other two types of literary dialect circulate for nearly a hundred years longer, and thus undergo the changing conventions that have been documented in previous chapters. Second, the ANOVA results presented in the statistical overview showed that the range of highly significant features distinguishing Chinese from African diasporic and Indian dialogue is much more robust than it is for the others (§4.4.1). The implications of Figure 4.6 were clear. Chinese literary dialect

[21] For an examination of the Chinese Exclusion Act in the United States, see, for example, Andrew Gyory (1998). For analyses of Australian and Canadian policies toward Chinese immigration see John Fitzgerald (2007) and Lisa Mar (2010), respectively.

has a stronger "signal" than either African diasporic or Indian literary dialect. The more robust set of identifiers yields a more coherent grouping on the dendrogram.

Finally, representations of African diasporic speakers, in particular, are influenced by an array of regional varieties and traditions for representing those varieties. We saw this in the various influences Caribbean and North American conventions have on British representational practices. Chinese literary dialect, by contrast, emerges more specifically from the social and linguistic conditions of Canton. The perception of a specifically "Chinese English" consolidates with the increased recognition of and discussion about "Canton jargon" in the early nineteenth century (Bolton, 2000, 2002, 2003). Because of its role as China's sole, official port for European and American trade between 1747 and 1842, Canton was an active site of linguistic contact. From this contact emerged a "jargon called *Canton-English*," which the same article from *The Chinese Repository* that is quoted in (3) describes as the lingua franca not only between Chinese and English speakers, but also among all foreigners who did business in the "factories" (or hongs) along the Pearl River, partly because the learning of Chinese by foreigners was outlawed (Bridgman, 1836, p. 432).

In my earlier discussion of the article, I noted that it – and other descriptions of "Canton Jargon" or "Pidgin English" like it – were widely circulated in newspapers and periodicals in the nineteenth century. These, I argue, affected the uptake of the variety into fiction and its construction as a literary dialect. And in this case, there is a piece of explicit evidence substantiating those links. One of the first examples in the corpus of Chinese literary dialect appears in a story by Eustace Wilberforce Jacob (1863) from *Something New, or, Tales for the Times*. The character, A-ping, is a Chinese servant who speaks only twenty-five words. Following his brief dialogue, another character Mr. Courtney, reads a parody of *Norval's address* from John Home's (1757) play *Douglas*. It is given to him by his companion, Dr. Compton, who asserts that the parody is "a receipt compounded by an American gentleman at Shanghae" and that it "will make you laugh if it does nothing else." The address begins (with the original text from *Douglas* in italics):

(12) My name belong Norval, top-side that Grampanie-hill,
My fader – you savey my fader? Makee pay chow-chow he sheep.
He smallo heartie man, too muchee likee that dollar; gala!
So fashion wanchee keep my counta one piecee chilo, stop he own side.

My name is Norval; on the Grampian Hills

My father feeds his flocks; a frugal swain,
Whose constant cares were to increase his store.
And keep his only son, myself, at home.

Before being appropriated by Jacob, this parody actually appeared in a number of other publications. In Britain, it was printed in *The National Magazine* (Saunders & Marston, 1862, p. 109), and an iteration published in *The United Service Magazine* (Pollock, 1863, pp. 364-365) the same year as Jacob's story claims that it was "penned some eight years ago by an American gentleman." The article continues, "It will appear I fancy a very incoherent piece of literature to those who have not an intimate acquaintance with this wondrous lingo, which bears a slight resemblance to the Anglo-Nursery dialect." The parody was reprinted in publications as diverse as *The Overland Monthly* and *The Mission Field* into the twentieth century, often alongside a similar "translation" of Longfellow's poem "Excelsior" (which is rendered as "Topside Galah").

The wholesale incorporation of the artifact into a literary work stands as a clear illustration of the processes of circulation, influence, and imitation. Jacob even imports phrases from the address directly into the dialogue of A-ping with minor variations (e.g., "one piecee chilo" in the address becomes "one smallo piecee cow chilo" in A-ping's dialogue). Moreover, the reference to "the Canton tongue" that prefaces Norval's address in *The United Service Magazine* solidifies the link from descriptions of "Canton Jargon" like the one in *The Chinese Repository*, to the parody of Norval's address, and ultimately to fictive voicings of Chinese characters like A-ping.

In addition to the relatively coherent clustering, there are two other elements of the dendrogram I would like to point out before moving on to an analysis of the outlier groupings. The first of these is the positioning of Ling-Wong's dialogue from Harry Collingwood's (1915) *A Chinese Command*. In the discussion of the corpus composition (§3.2.1), I stated that the inclusion of Ling-Wong's dialogue was a borderline case. While Ling-Wong is identified as Korean, he is explicitly and repeatedly described as speaking "pidgin English." On that basis, I opted to retain the dialogue. The clustering shows that, in fact, Ling-Wong is ventriloquized using a constellation of features that aligns with conventional voicings of Chinese characters. This is at least one historical data point in support of the argument made by scholars like Chun (2004), which posits that stereotypes of Chinese language and customs come to be applied to a generalized Asian identity in contemporary culture.

The other noteworthy facet of the dendrogram is the bifoliate grouping of the dialogue from two Frederick Brereton novels: *The Hero of Panama* and *Under the Chinese Dragon*. Given that they share the same author and are both published in the same year (1912), we would expect their features to be similar. That the dendrogram underlines their similarities supports the validity of the overall approach. Their pairing also highlights an issue that this particular study is not designed to address, but one that is intriguing nonetheless. The bulk of the analysis has been invested in exploring intersections of historical ideological currents and changing linguistic representation. Occasionally, however, the analysis has bumped up against questions of individual authorial style, as it did in the discussion of M. M. Noah and *v*-for-*w*/*wh* substitution in African diasporic dialogue (§5.3). The question of how much variation is attributable to the constraints of literary dialect conventions versus the stylistic idiosyncrasies of a given author is not inconsequential. In his analysis of style, Jockers (2013, pp. 92-93) demonstrates that is easier to classify a chunk of text by its author than by the work from which it is extracted. In other words, one can more accurately identify that a chunk of text was written by Charles Dickens than that it comes from either *Great Expectations* or *David Copperfield*. In short, individual stylistic tendencies are strong.

The grouping of Brereton's texts, I think, attests to the fact that literary dialect is not immune to those tendencies. In Brereton's case, he creates an amalgam of Chinese and African diasporic literary dialect conventions. In the following examples (the top from *The Hero of Panama* and the bottom from *Under the Chinese Dragon*), the blending of indexical features is clear:

(13) A cork, sah; I's got the velly thing. You wantee someting to push in dere. Ching hab plenty fine cork. (*The Hero of Panama)*

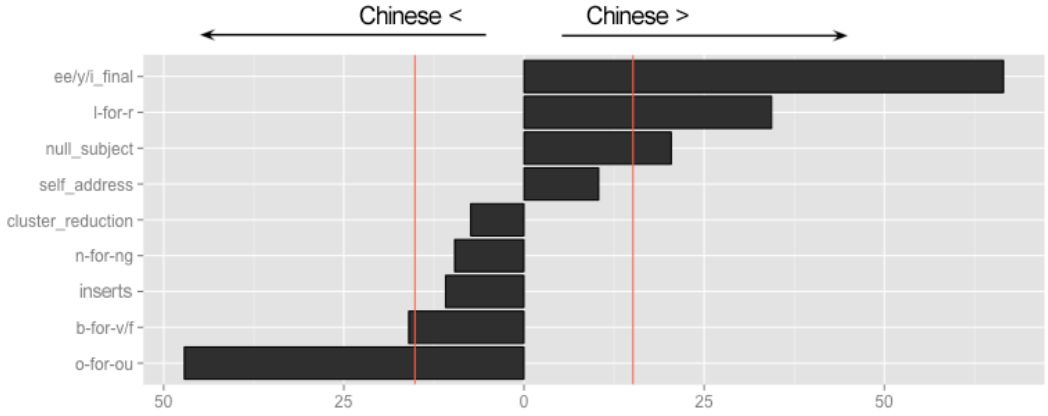
Jong say dat allee lightee. Watch, den no easy to be cut to piecee. Neber know who or what comin' along. P'laps dere robbers. Dey make mincemeat of de lot of us before you have time to breathe. (*Under the Chinese Dragon)*

Brereton's dialogue combines two of the most significant identifiers of Chinese dialogue (*l*-for-*r* substitution and word-final *-ee*/*-y*/*-i* insertion) and two of the most significant identifiers of African diasporic dialogue (*t*/*d*-for-*th* substitution and *b*-for-*v*/*f* substitution).

One way to further clarify how Brereton constructs his Chinese literary dialect is to compare his Chinese dialogue to his African diasporic dialogue from *The Hero of Panama*. A log-likelihood analysis reveals an interesting pattern (see Figure 7.5). Of

the nine features that have at least moderately significant distributions ($p < 0.01$), only five are highly significant ($p < 0.0001$). Of the three that distinguish his Chinese dialogue, two (*l*-for-*r* substitution and word-final *-ee*/*-y*/*-i* insertion) are the phonological features evident in (13). The third (null subject) is one of the discourse-organization-type features identified with Chinese dialogue by ANOVA (see Figure 4.6). Of the two that distinguish African diasporic dialogue, one (*b*-for-*v*/*f* substitution) is also evident in (13). As it does in the excerpt, the feature occurs in Brereton's Chinese dialogue, just to a significantly lesser degree than it does in his African diasporic dialogue. This is also true of *n*-for-*ng* substitution, which is exceedingly rare in Chinese dialogue. (There are only two other instances outside of Brereton's texts.) The other feature distinguishing African diasporic dialogue is *o*-for-*ou* substitution, which is lexically restricted to *yo* for *you*.

Figure 7.5: Bar plot showing log-likelihood comparisons between Chinese and African diasporic dialogue from *The Hero of Panama* for features where $p < 0.001$. The red lines mark the points at which $p < 0.0001$.



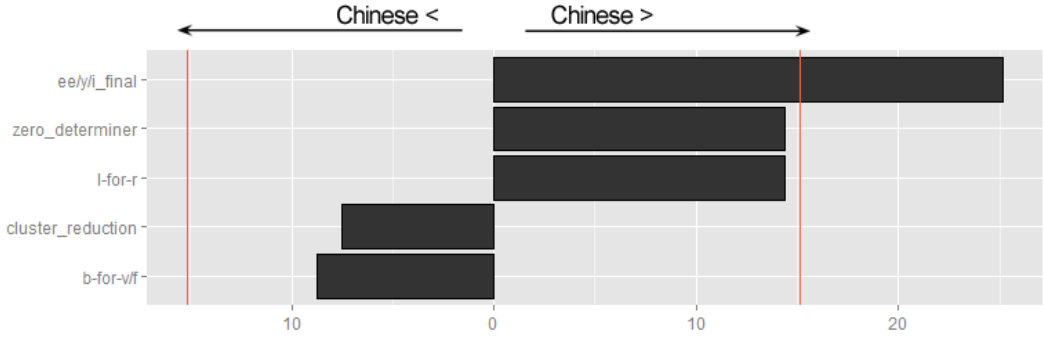
What is perhaps even more telling than the features identified by ANOVA are the features that are absent. Of the twelve morphosyntactic features contained in Figure 4.6 that are relevant to Chinese dialogue, only one (null subject) is present in the log-likelihood analysis illustrated in 7.5. Similarly, realizations of *t*/*d*-for-*th* substitution and address-type features, which we would expect to be skewed in favor of African diasporic dialogue, show no significant comparative distributions. The evidence seems to point to Brereton using African diasporic literary dialect as a kind of prototype onto which he grafts a small set of highly indexical features (*l*-for-*r* substitution, word-final *-ee*/*-y*/*-i* insertion, and null subject) in order to construct his Chinese literary dialect. Remember, too, that Brereton is one of only three authors to employ *massa* as an address form in Chinese dialogue, a fact that further supports this

interpretation. Variants of *YOU* serve as additional markers of differentiation, although the variant does occur once in John Jong's dialogue, as well ("**Yo** hold de light high, so as to shine on de enemy only").

Both from the atypical position his texts occupy on the dendrogram and their clustering together, it is reasonable to conclude that Brereton's Chinese dialogue evidences distinct traces of his individual authorial style. Even his African diasporic dialogue, which is otherwise quite conventional, is marked by its frequency of *o*-for-*ou* substitution. (There are 45 occurrences *The Hero of Panama* and only a total of four in two other texts.) However, stylistic idiosyncrasies are only a partial explanation for the groupings we find on the dendrogram.

One indication that there are other factors at work is the fact that Brereton's Chinese dialogue is not alone in its sub-cluster. It is part of a trifoliate grouping with the Chinese dialogue from Robert Ballantyne's *Under the Waves*. A log-likelihood comparison of Ballantyne's Chinese dialogue in *Under the Waves* with his African diasporic dialogue from *The Middy and Moors* reveals a similar pattern to what was found in the comparison of Brereton's texts (see Figure 7.6). The two most indexical features of Chinese literary dialect (*l*-for-*r* substitution and word-final *-ee*/*-y*/*-i* insertion) appear grafted onto a base structure of conventionally African diasporic features, with a small number of additional differences, but few of the other lexical, morphosyntactic, or phonological variations that typically distinguish the literary dialects. Thus, the patterning of Brereton's Chinese dialogue, while unusual, is not wholly unique.

Figure 7.6: Bar plot showing log-likelihood comparisons between Chinese from *Under the Waves* and African diasporic dialogue from *Middy and the Moors* for features where p < 0.01. The red lines mark the points at which p < 0.0001.

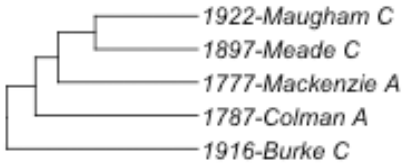


Precisely why Brereton's and Ballantyne's texts intersect in this way is difficult to say. Interestingly, both authors are highly productive writers of juvenile adventure fiction, Ballantyne having written over 100 novels and Brereton over 40. It

is certainly plausible that works produced quickly for an audience not particularly concerned with verisimilitude would be prone to instantiations of generic or marginally modified literary dialect. Whether or not such forces are at work in *The Hero of Panama*, *Under the Chinese Dragon*, or *Under the Waves* is, of course, impossible to determine definitively. In any event, such an explanation is likely only partial as equally prolific authors working in the same genre (like George Alfred Henty and John C. Hutchenson) produce literary dialect that is positioned very differently on the dendrogram.

The motivations – whether conscious or unconscious, whether stylistic or pragmatic – that shape atypical expressions of literary dialect like Brereton's are undoubtedly difficult to isolate. The attitudes and identities that his literary dialect encodes, however, are less obscure, and it is the kind of racial figuring that we have seen before. Both Brereton and Ballantyne are defenders of the ideological underpinnings of the British imperial project in their stories, which use exoticized imperial settings and imperial conflict as occasions for the moral instruction of boys and the promotion of white British masculinity (Kennedy, 2014; Richards, 1989). Thus, Brereton's specific expression of Chinese literary dialect is neither ideologically naïve nor ideologically neutral. Its foundation of African diasporic features suggests his Chinese and African diasporic characters have shared subjectivities. And, in fact, they are characterized in strikingly similar terms – as sometimes comic, but always loyal supporters of the Anglo heroes. John Jong, a Chinese cook in *Under the Chinese Dragon*, is described as a "faithful celestial" and a "faithful Chinaman." Ching Hu, also a cook and laborer, is virtually indistinguishable from the African diasporic characters, Sam and Tom, in *The Hero of Panama*. They are conflated as "these three faithful fellows" who are devoted to Jim, "their youthful master."

Figure 7.7: Pentafoliate grouping from cluster 3C in the full dendrogram, which contains the Chinese dialogue from *East of Suez* (1922), *Under the Dragon Throne* (1897), and *Limehouse Nights* (1916).



The other anomalous cluster – which consists of the Chinese dialogue from W. Somerset Maugham's (1922) *East of Suez*, Elizabeth Meade's (1897) *Under the*

Dragon Throne, and Thomas Burke's (1916) *Limehouse Nights* – raises similar questions of genres, their themes, and their ideological leanings. The sub-cluster in which the texts appear is one that contains a large number of early texts, as well as a grouping of later Indian dialogue. The five-text grouping within cluster 3C that contains the three instances of Chinese dialogue also contains two eighteenth century examples of African diasporic dialogue (see Figure 7.7). Log-likelihood comparisons of the Chinese and African diasporic dialogue from this pentafoliate grouping show again *l*-for-*r* substitution (LL = 66.79) and word-final -*ee*/-*y*/-*i* insertion (LL = 41.90) being significantly more frequent in the Chinese dialogue. These comparisons are aggregations, and we will see how specific texts like Burke's *Limehouse Nights* realize these two features in unusual ways. For now, however, it is enough to note that the Chinese dialogue in this group is distinguished from its African diasporic counterparts by the same indexical markers that differentiate Brereton's and Ballantyne's texts. The structures underlying the resemblances within the overall grouping, however, are different. The heat map from chapter 4 (see Figure 4.17) showed that cluster 3C is characterized by low overall feature frequencies, in contrast to cluster 1A, which is marked by high frequencies of *t*/*d*-for-*th* substitution and cluster reduction among other features.

Unsurprisingly, then, *Under the Dragon Throne* and *East of Suez* have the lowest composite feature frequencies for Chinese dialogue (273.58 and 370.44). The composite frequency for *Limehouse Nights* is a bit higher (427.78), but is still the fifth lowest for Chinese dialogue. The question is what might explain these lower frequencies and their grouping together? One potentially salient factor is that, like the works of Brereton and Ballantyne, these are linked by genre. Whereas *The Hero of Panama*, *Under the Chinese Dragon*, and *Under the Waves* are juvenile adventures, *Under the Dragon Throne*, *East of Suez* and *Limehouse Nights* are romances and domestic melodramas. Additionally, all three contain plots that involve cross-racial romance. In *Under the Dragon Throne*, a young British officer, James Pennant, absconds with Amethyst, who is betrothed to a Chinese official. In *East of Suez*, Daisy, the daughter of a British father and a Chinese mother, attempts to hide her parentage by having her mother pose as her amah, while Daisy pursues her love for the British George Conway, who is the best friend of her husband, Harry Anderson. *Limehouse Nights* actually has a number of related plotlines. In one, the English Lucy

is rescued from unnamed "horrors" by Cheng Huan and made into "the living interpretation of a Chinese lyric."

Of the two nineteenth century texts in the grouping with the Chinese dialogue, Colman's *Inkle and Yarico* treads similar ground. The play follows the romance between a shipwrecked British trader, Inkle, and the Indian woman who saves him, Yarico, as well as a parallel romance between the servant Trudge and the African diasporic Wowski. The only one of the five texts that does not seem to adhere to the pattern is Mackenzie's *Julia de Roubigné*. The literary dialect in the epistolary novel comes from the voicing of Yambu, an enslaved former "prince" who had been "master of them all" (meaning the other plantation slaves). The narrator, Savillon, decides to "free" Yambu by, in effect, making him the plantation foreman or overseer. This affirms Savillon's (and by extension Mackenzie's) moral self-image as a proto-abolitionist, while also preserving the racist order. Additionally, Lilly (2007, p. 662) argues, it also serves to more efficiently marshal the resources of the plantation economy and "control African bodies." As Savillon reports of Yambu after his abolitionist experiment:

> (14) He has, accordingly, ever since had the command of his former subjects, and superintended their work in a particular quarter of the plantation; and, having been declared free, according to the mode prescribed by the laws of the island, has a certain portion of ground allotted him, the produce of which is his property. I have had the satisfaction of observing those men, under the feeling of good treatment, and the idea of liberty, do more than almost double their number subject to the whip of an overseer. I am under no apprehension of desertion or mutiny; they work with the willingness of freedom, yet are mine with more than the obligation of slavery.

Mackenzie's novel, thus, articulates a paternalistic and sentimental vision of cross-racial relationships that echoes the other works in the cluster. All five works express a deep ambivalence toward the charisma of non-white bodies and toward their autonomy, whether economic or erotic. Such ambivalence is evident in Colman's play, which is a reworking of an older story. In its original form Inkle sells Yarico into slavery, but Colman's version has Inkle reconsider his betrayal, a change that Odumosu (2014, p. 132) interprets as "a touch of abolitionist sentiment." Yet, in her analysis of *Inkle and Yarico*, Nussbaum (2003, p. 249) asserts that the "edgy racism" voiced by characters in the play (e.g., Trudge consistently comments on Wowski's complexion, calling her "my dingy dear," "my poor, dear, dingy, wife," etc.) "reflects the unresolved tensions surrounding racial issues as blacks are incorporated within the English economy and culture."

Similar tensions are at work in Burke's figurations of the Chinese-controlled worlds in *Limehouse Nights*. They, too, are marginally incorporated into mainstream British culture. They are charismatic, but also morally and physically perilous to the outsider. Cheng Huan may rescue Lucy, but his attentions ultimately lead to her death at the hands of her enraged father. Witchard (2009, p. 4) calls Burke's Limehouse as a "Chinoiserie" – a place "as alluring as it is forbidding." It is a rendering she views as distinct from Sax Rohmer's more monolithically paranoid and hostile vision. These differing imaginings of Chinese community and culture are reflected in their imaginings of Chinese voices. The literary dialect from Rohmer's novel *Dope* occupies a very different position on the dendrogram from Burke's dialogue. It appears in cluster 2, paired with the dialogue from John C. Hutcheson's *Afloat at Last*. These two texts also have the highest composite frequencies in the corpus (roughly 786 for both).

By log-likelihood comparisons, the most significant difference between Burke's and Rohmer's dialogue is the greater frequency of word-final *-ee*/*-y*/*-i* insertion in *Dope* ($LL = 62.11$, $p < 0.0001$). After 1861, there are, in fact, only two examples in the corpus of Chinese dialogue that does not contain word-final *-ee*/*-y*/*-i* insertion: *Interplay* by Beatrice Harraden (1908) and *Limehouse Nights*. The former is a domestic drama that explores the proposition that the protagonist, Harriet Rivers, is justified in having an affair because of an abusive husband, or, as one review derisively puts it, "did right by violating the seventh commandment." The Chinese dialogue belongs to a servant, Quong, who is described as having "a whole fund of real human kindness in his Chinese heart." The lack of the indexical feature may be an attempt to mitigate the perception of comic stereotype, in accordance with the novel's progressive themes.

Burke similarly manipulates conventional renderings of Chinese identities and voices without necessarily toppling them. As we have seen, stereotypes of Chinese culture and identity calcify at the turn of the century. One strain figures Chinese people as just another iteration in a long line of non-white, solicitous servants whose sole function is to facilitate the progress (physical, economic, military, moral, romantic, etc.) of a white protagonist. A second strain emerges from late nineteenth century imperial conflict and Chinese immigration in North America. It figures Chinese identity as at once indolent and cruel, as emasculated yet posing a sexual danger to white femininity and a demographic threat to Western culture. Witchard

(2009, p. 18) attributes the durability of this latter image to the influence of De Quincey, who she says "is responsible for collating in *China* (1857) the many facets of the stereotype that would gain widespread currency and sustain the Rohmeresque Chinaman into the twentieth century." Burke breaks with these conventions, Witchard argues, by presenting Chinese characters like Lucy's protector, Cheng Huan, as sympathetic.

Indeed, there are clear moments of subversion in *Limehouse Nights*. Many of the most pernicious racial attitudes are given voice by unreliable or unsympathetic characters. Lucy's abusive father believes "yeller" is the "supreme condemnation" because his "birth and education in Shadwell had taught him that of all creeping things that creep upon the earth the most insidious is the Oriental in the West." That this paranoia is embodied by a drunk who flogs his daughter undoubtedly holds it up to ridicule. Yet, Burke also invites the reader to at least partly share in one aspect of the father's fears. As the father seethes in thinking about Cheng Huan with twelve-year-old Lucy, the narration is turned over to his interior monologue: "It was… as you might say… so… kind of … well, wasn't it?" Burke hails his reader with the second person pronoun, and we are called to fill in the ellipses that the father's consciousness cannot quite articulate. The elided information is, of course, the possibility that the relationship between Cheng Huan and Lucy is or might turn sexual. Though it is clear that Cheng Huan is kind and that Lucy does not fear him, Burke subtly clouds the nature of their relationship with the juxtaposition of adverbs describing how Cheng Huan looks at her ("reverently yet passionately") and touches her ("wistfully yet eagerly"). On the one hand, Cheng Huan is figured as feeling love, loss, and pain, a range of emotions not conventionally accorded Chinese characters. On the other, Burke uses the stereotype of a sexually predatory Chinese man to insinuate the salacious possibility that there is more going here than meets the eye, that their relationship was "kind of… well, wasn't it?"

This ambivalence is evident in other ways, as well. The "edgy racism" in *Inkle and Yarico* that Nussbaum critiques has analogous expression in *Limehouse Nights*. There are similar, repeated references complexion (*yellow men*, *yellow hands*, *yellow faces*, etc.) and uses of dysphemisms (*chink*, *chinky*). It is an ambivalence that is also reflected in Burke's literary dialect. In its unusual eschewing word-final *-ee*/*-y*/*-i* insertion, his Chinese dialogue rejects a stereotypical constituent and its enregistered associations. However, his Chinese dialogue does realize the other conventional

constituent: *l*-for-*r* substitution ("Oh, lou'll have **evelything** beautiful"). Too, it is the only dialogue in the corpus to contain *l*-for-*y* substitution, which occurs word-initially (in *lou* for *you* and *les* for *yes*). The effect of this latter feature is to amplify the presence of the already marked nonstandard *l*. Thus, along one dimension, the stereotypical indexes of his literary dialect are attenuated, but along another, they are exaggerated.

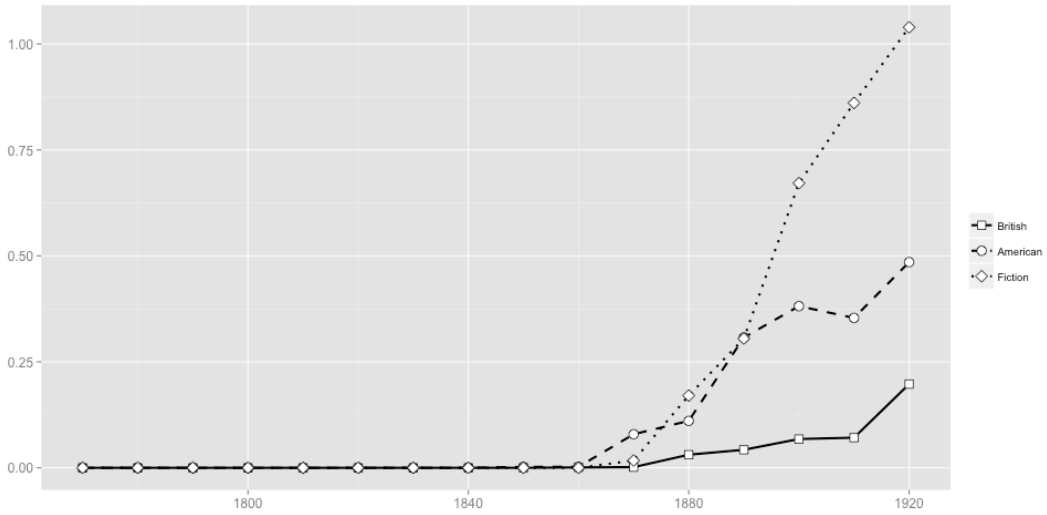
Finally, there is Burke's description of Limehouse, itself. The model for Limehouse as a Chinatown is San Francisco (as it is for other fictionalized Chinatowns in cities like New York and Chicago). Literary descriptions of San Francisco's Chinatown begin in the nineteenth century in novels like Altwell Whitney's (1878) *Almond-eyed; A Story of the Day*. In the source works, San Francisco's Chinatown features in two pieces. *The Shadow of Quong Lung* (1900) takes place entirely in San Francisco. (The author, Charles William Doyle, was living in the city at the time and writing for *The Overland Monthly*.) The novel's titular character is a Chinatown crime boss, and Doyle makes his own attitude clear in novel's prologue, stating that "the best thing to do with Chinatown would be to burn it down." Horace Annesley Vachell's (1912) *Bunch Grass* is similarly set in California. (Like Doyle, Vachell spent some time living there.) Midway through the work, the narrator and his brother, who work on a ranch in Southern California, travel to San Francisco to rescue a friend who has become addicted to opium. There, the protagonists enter "the labyrinth of Chinatown." They pick their way through "an abominable rookery," its foul smells and indecipherable sounds signaling that they are on "unholy ground." In these and similar works, Chinatown serves as a site of mystery, danger, and desire. Burke maps these same qualities onto Limehouse:

> (15) In the Causeway all was secrecy and half tones. The winter's day had died in a wrath of flame and cloud, and now pinpoints of light pricked the curtain of mist. The shuttered gloom of the quarter showed strangely menacing. Every whispering house seemed an abode of dread things. Every window seemed filled with frightful eyes. Every corner, half lit by the bleak light of a naked gas-jet, seemed to harbour unholy things, and a sense of danger hung on every step. The Causeway was just a fog of yellow faces and labial murmurings.

The passages from Burke and Vatchell are remarkably similar in their imagery – their evocations unfamiliarity and peril, their shared use of the word *unholy*. Moreover, the move from Vatchel to Burke, the move from San Francisco to London, is an example of the emergence of Limehouse as a social and linguistic space. This is an important turn in the representations of Chinese people and culture in British

literature. It signals the domestication of both conventions circulating in Anglophone discourse and the attitudes and anxieties that those conventions encode. The Google Books data can again assist in illustrating the broad contours of these changes, this time by looking at frequencies of *CHINATOWN* (see Figure 7.8). One trend that the graph highlights is the earlier adoption of *CHINATOWN* in American as compared to British English. This fits with the evidence from the source works, where the token first occurs in *The Shadow of Quong Lung*, which is not published until 1900.

Figure 7.8: Frequencies (normalized per million words) of lemmatized *CHINATOWN* in the Google Books data tables from 1770-1930.



The turn-of-the-century publication date for Doyle's novel also aligns with the token's spread in fiction in the early twentieth century. Importantly, this is the period that marks the rise of dime novels and detective fiction – genres in which Chinatowns figure prominently as pockets of mysterious and exoticized danger within the domestic metropole. Hoppenstand (1992, p. 283) argues, "Chinatown and the opium den, because of the dime novel's influence, framed a symbol of warning to every Anglo-American who wanted to 'experiment' in foreign Oriental cultures, suggesting that such experimentation could result in drug-induced madness or a hatchet in the back." One only need look at the number of titles in a dime novel series like the *Brady Detectives* to glean the popularity of Chinatown as a setting in these genres. Between 1899 and 1912, the series published 62 titles (such as *The Bradys and the Drug Slaves; or, The Yellow Demons of Chinatown*) set or partially set in a Chinatown, according to the Dime Novel and Story Paper Collection database (Stanford University, 2015).

In British literature, the shift from North American to domestic settings is prefaced not only by works like Doyle's, Vatchell's, and dime novel detective stories, but also by nineteenth century works like Charles Dickens' (1870) *The Mystery of Edwin Drood* and Arthur Conan Doyle's (1891) "The Man with the Twisted Lip," as well as a series of articles in popular publications, all of which depict opium dens in Britain and connect opium distribution with Chinese culture. Many of these fictional and quasi-journalistic accounts frame their narratives as providing readers access to an otherwise inaccessible part of the city and consequently describe the district's Chinese identity as hidden. One such article published in the *Penny Illustrated Paper* ("Opium dens in London," 1910) suggests that Limehouse has a veneer of Britishness that belies its true character: "Outside the houses look thoroughly English in appearance, but go inside No. X or No. Z, and the scene changes from English to Oriental by the simple process of stepping through a doorway." The suggestion of camouflage invites the reader imagine that something alien may lurk behind something familiar. Despite its outwardly British appearance, Limehouse may be, as the author calls it, "the Orient of the metropolis."

It is also a neighborhood, like San Francisco's Chinatown, marked by its language variety. The American author Chester Bailey Fernald (1907, p. 75), whose stories are frequently set in San Francisco, refers to "the Chinatown English dialect, with its vulgar intonations and its slang, drawn from the streets." In (15), the "labial murmurings" that Burke references are as much a part of the enigmatic menace of Limehouse as the "bleak light of a naked gas-jet" or the "fog of yellow faces." In *Bunch Grass*, Vatchell's narrator similarly describes the voices of San Francisco's Chinatown as "mere guttural sounds, that conveyed nothing to the ear." Once again, Burke and Vatchell are striking in their overlap – this time for their linking of sound with space, of the speech of Chinese people with the social and psychological meanings of the geographies they control.

The other source work in which Limehouse figures prominently is Sax Rohmer's *Dope*, and Rohmer, too, draws these connections, though in a different way. In *Dope*, Sin Sin Wa's language is "strange, sibilant speech which is alien from all Western conceptions of oral intercourse." It is "murmured gibberish" and "that weird jargon known as 'pidgin.'" These latter descriptions are offered up as evidence of the "inscrutable mystery" of Sin Sin Wa as an archetype of his culture – "his racial inability to express his thoughts intelligibly in any European tongue" standing in

contrast to his ability to "converse upon many and curious subjects in his own language."

The intersections of the linguistic and the geo-social are not limited to characters whom Rohmer ventriloquizes using literary dialect, like Sin Sin Wa or Shen-Yan (a Chinese barber whose London shop is a front for an opium den in *The Mystery of Dr. Fu-Manchu*). Rohmer's iconic villain, Dr. Fu-Manchu, is "a linguist who speaks with almost equal facility in any of the civilized languages, and in most of the barbaric." However, his "perfect English" functions as slightly flawed guise. The "occasional guttural" betrays the identity that is more outwardly visible upon his body: his "wicked, pock-marked face," his "wolfish fangs," and his "inanimate, dull, inhuman" eyes. It is a vocal trait shared by Chinese master criminals in other works. Doyle's Quong Lung speaks "with a refined English accent," and Shiel's Yen How makes "of himself an epitome of the West," but is betrayed by "his inability […] to pronounce the word 'little,'" instead "still call[ing] it 'lillee.'" These voices are integral parts of a Westernized façade. In this way, they are like the exteriors of Limehouse, itself. In a collection of short stories titled *Tales of Chinatown*, Rohmer (1922, p. 14) writes:

> (16) Unlike its sister colony in New York, there are no show places in Limehouse. The visitor sees nothing but mean streets and dark doorways. The superficial inquirer comes away convinced that the romance of the Asiatic district has no existence outside the imaginations of writers of fiction. Yet here lies a secret quarter, as secret and as strange, in its smaller way, as its parent in China which is called the Purple Forbidden City.

What seems British hides something "secret" and "strange," or, as the article from the *Penny Illustrated Paper* had it, "the scene changes from English to Oriental by the simple process of stepping through a doorway." This constitutes a kind of paranoid distorting of the "mimic men" principle, which was discussed in the previous chapter (§6.1). In the Indian context, mimicry was a sign of imperial success and expansion. In this context, mimicry is posited as a Trojan horse – a mechanism that disguises a threat to the domestic metropole.

7.5 Conclusion

Chinese literary dialect is characterized by a relatively large set of distinguishing features. Many of these, as was discussed in the statistical overview (§4.4.1), are at the intersection of lexicon and syntax (e.g., *piece* as a determiner). That such features are prone to enregisterment may result from their being more

interpretable by readers than other kinds of morphosyntactic features like the zero determiner. The latter may read as disfluent or generically nonstandard, but not necessarily indexical of a particular vocal culture. Alternatively, a feature that is lexically distinctive like *piece* may more readily develop specific associations.

As important as morphosyntactic features are to representations of Chinese voices, the most significant features are phonological: *l*-for-*r* substitution and word-final -*ee*/-*y*/-*i* insertion. The consonant substitution has a particularly long history of association with Chinese vocal culture, with descriptions circulating at least into the eighteenth century. Yet, it, like the others, does not emerge as a convention for voicing Chinese characters in fiction until the mid- to late nineteenth century. The development and spread of these conventions appears to be fueled, in part, by circumstances in the American West and San Francisco in particular. These circumstances include a growing Chinese immigrant population and the concomitant racial paranoia; the rise of new adventure genres that figure Chinatowns as sites of exoticized culture, mystery and crime; and the burgeoning influence of San Francisco as a center for literary output, which attracts visitors and imitators alike.

The specific form of Chinese literary dialect is shaped by Chinese Pidgin English as it developed in Canton in the eighteenth and nineteenth centuries. Not only do descriptions of the variety circulate globally prior to their adoption into literature, but the speakers themselves are also an important part of the transpacific migration into and through San Francisco (Chen, 2000). The literary dialect that ultimately emerges, then, quickly calcifies into a fairly stable set of conventions. That stability, in combination with the greater number of distinguishing features, creates a robust signal that generates the relatively homogeneous clustering that we saw in the dendrogram (see Figure 7.4).

In many ways, Chinese literary dialect clearly arises from historical contexts that are distinct from either African diasporic or Indian literary dialect. The two varieties that have been analyzed previously appear contemporaneously, and they both exhibit early similarities before individuating. While the very first example of Chinese dialogue in the corpus realizes features that are indexical of African diasporic representations (e.g., *massa* as an address form), the structures of Chinese literary dialect are more consistently distinctive, likely the product of its mimicry of a specific language variety and of its development during the height of literary dialect's popularity.

In spite of such differences, there are overlaps that may be less apparent, such as highlighting the importance of circulation to all three types of literary dialect. In the introduction (§1.5), a 1747 letter supposedly written by a slave, Toby, to his "Masser Frankee" was presented (see Figure 1.7) in order to illustrate the circulation of representational practices before they are taken up in fiction. The connection between the Toby letter and a specific work like *The Padlock* is entirely circumstantial, of course. We have no evidence to suggest Bickerstaff or any of his collaborators were directly influenced by it. The inclusion of the Norval parody in Jacob's *Something New, or, Tales for the Times* (§7.4), however, confirms that these kinds of circulating artifacts directly inspire at least some authors, in addition to contributing to the more general processes of enregisterment.

The contexts surrounding Chinese literary dialect and its emergence also underscore the salience of the social and cultural conditions of empire in informing the representations of peoples – whether materially subjugated, aspirationally subjugated, or adversarial. In earlier chapters, we have seen how ideas of empire can intersect with the imaginings of identity in order to rationalize imperial authority. This chapter has demonstrated similar patterns. It has shown that more aggressively racist representations of Chinese vocal culture developed just as Britain and China clashed over Britain's access to China's markets. Moreover, as much as this chapter has emphasized San Francisco's role in propagating stereotypes and racist paranoia, Britain's fascinations and fears were primed by its own contact and conflict with China. Thus, the British imagination is receptive to tropes of Chinatown, Chinese immorality, and demographic apocalypse at the turn of the century.

Chapter 8
Conclusion

8.1 Introduction

In this last chapter, the major findings of the study are summarized – findings related to diachronic and synchronic patterns in the literary dialect used to represent African diasporic, Indian, and Chinese speakers (§8.2). That is followed by a discussion of the study's implications for research at the intersection of literature and linguistics (§8.3), its limitations (§8.4), and some potential avenues for future inquiry (§8.5). The concluding remarks (§8.6) examine how some of the patterns that this study has exposed project not just into the past, but also into the present and not just into imaginary worlds, but into the worlds of real speech communities.

8.2 Summary of major findings

This study set out to investigate the following, overarching research question: how is literary dialect used as an imaginative tool to represent the language of African diasporic, Chinese, and Indian speakers? That question was then broken down into the following more specific operable questions:

- What are the patterns of lexical, morphosyntactic, orthographic and phonological features that distinguish specific, imagined language varieties?
- In what ways, if any, do such patterns evolve over time?
- To what extent and in what ways are there any shared patterns of features between or among varieties?
- How are patterns of linguistic representation implicated in evolving understandings of race, culture, and empire?

The first three questions were addressed sequentially in each of chapters 5, 6, and 7. In addressing those questions, the study employed a number of computational techniques. Deviation of proportions and normalized composite frequencies were used in the analysis of constituent patterns. Regression analysis was used to model changes over time in both composite frequencies and diversity indices, while resemblances were modeled using hierarchical cluster analysis. Of these, diversity indices have been least frequently implemented in corpus research, and their use here suggests that they may have wider application as a measure of linguistic complexity (syntactic complexity, phonological complexity, etc.) Although the other quantitative techniques

are not new to corpus research, their specific combination as an approach to modeling language variation and change along multiple dimensions shows how such techniques might productively complement other types of analysis.

The quantitative analysis identified a number of significant patterns, including but not limited to the following.

- For African diasporic dialogue:
    - the salience of address and three phonological features (*t*/d-for-*th* substitution, *b*-for-*v*/*f* substitution, and cluster reduction) as indexical markers;
    - an increasing frequency and complexity in phonological features over time; and
    - stronger resemblances among later representations of African diasporic vocal culture than among earlier ones, though still with some outliers.
- For Indian dialogue:
    - only two distinguishing features, both of which are lexical (address and code-mixing);
    - a decreasing complexity in its structure over time but with lexical features increasing into the middle of the nineteenth century before declining; and
    - the weakest resemblances, but with a substantial number of later texts similar in their low frequencies or low diversity indices.
- For Chinese dialogue:
    - a high number of distinguishing features, including two significant phonological features (*l*-for-*r* substitution and word-final *-ee*/*-y*/*-i* insertion) and a large number of morphosyntactic features (e.g., *piece* as a determiner, *-man* as a nominal suffix, *much* as an intensifier, *belong* as a copular verb, and *heap* as an intensifier); and
    - the strongest resemblances, which is partly driven by its greater overall marking.

Addressing the fourth research question, then, necessitated explicating these patterns. To do so, the analysis incorporated qualitative data from the imperial archive.

Quantitative patterns have been examined as they relate to social, ideological, aesthetic, and material forces: global circulations of people and texts, developments in printing technologies and consumption, the waxing and waning in the popularity of genres, as well as the changing social, economic, and political circumstances of empire. In the case of African diasporic dialogue, the late nineteenth century craze for accent imitation and calcifying racist tropes partly drive the increase in phonological features. The opposing trends in Indian dialogue are affected by changes in the imagining of racialized identities and newly circulating Anglo-Indian lexicons. And

demographic paranoia together with the catalytic conditions of the American West inform Chinese literary dialect as a relatively late emerging and stable convention.

8.3 Implications for the field

This thesis makes a number of contributions to the linguistic analysis of literature. For one, as a study of literary dialect, it is unusually large in both its scope and size. That said, using some of the techniques set out in this thesis, research into literary dialect might be greatly expanded by additionally incorporating advances in statistical learning (a possibility that I discuss in a later section). This study has also demonstrated how computational tools that are less commonly applied in corpus analysis can be marshaled in the identification of synchronic and diachronic patterns.

In addition to its purely statistical contributions, the analysis has sought to bridge the quantitative and qualitative commitments of a number of disciplines. The study, for example, engages with a number of research areas in sociolinguistics and literary studies that have more qualitatively oriented traditions, like enregisterment research and colonial discourse studies. The results demonstrate how such traditions of textual analysis might be supplemented by quantitative approaches that zoom out to expose patterns that are otherwise difficult to discern. Alternatively, the study also engages with research areas that are computationally oriented, like the digital humanities. These traditions already emphasize the perspective that is gained by zooming out; in some cases, they are defined by their "distance" from data, in opposition to older, established ways of "close reading" (e.g., Moretti, 2005). While the results certainly affirm the explanatory value of quantitative, "distant" analysis, they also suggest how qualitative, "close" analysis can be productively allied with the former. By zooming in and out, by working with data at different levels of resolution, we can produce robust accounts of synchronic and diachronic patterns.

As important as these kinds of contributions are, the study has sought to expose more than the power of a method. It has sought to expose the power of the patterns themselves. The combination of data has shown how enduring some features have been as indexes of vocal cultures. In all three cases, there is evidence of features circulating in other genres and text-types, sometimes for centuries, before they are taken up in literature. Once they are taken up, some constituents of literary dialect (like *v*-for-*w*/*wh* substitution) are subject to shifting conventions. Others (like *t*/*d*-for-*th* substitution, *l*-for-*r* substitution, and forms of address), however, maintain strong

associations. Those associations are not simply neutral linkages of signs and speakers, convenient resources for writers to summon accents or communities. They carry with them social valuations, and while individual texts may challenge or subvert orthodox ideologies, the overall pattern is for these associations to uphold the racial order and rationalize asymmetries of power. It is this last point that I emphasize in my concluding remarks, as the coda to this study.

8.4 Limitations of the study

In the statistical overview (§4.2.3), I noted some of the complications regarding p-values and significance in this study. In particular, that brief discussion suggested the contingencies that are introduced when working with digital archives. These archives are not constructed to be statistically representative of periods, genres, or regions. They are built from artifacts that are on hand at a single institution or small set of cooperating institutions. There is, therefore, a degree of arbitrariness to the data that this study draws from, adding uncertainty to the claims that are founded on that data. That uncertainty is further augmented by sampling methods that require relatively inefficient searches and sorting to prevent introducing various kinds of bias (§3.2.1).

Like the data collection, the data coding also presented challenges that attenuate the strength of the findings. In the chapter on research methods, I described some of the decisions that I faced when developing the coding taxonomy (§3.5.1 and §3.5.2). Those decisions led to a continual reassessment and revising of the taxonomy, even into the period when the analysis was being drafted. One of the strengths of the resulting scheme is that it accounts for most features in a way that limits researcher bias. By putting the focus on respellings rather than the phonemic inferences of those respellings, for example, the scheme reduced the likelihood that the researcher's own internal sense of "standardness" would substantially influence the assigning of particular codes. Yet, the potential for such bias proved more difficult to mitigate for other categories. For instance, general vocabulary designates words that distinguish a speaker's dialogue from other dialogue or the narration. Because the coding was done by hand, determinations of what is distinctive were susceptible to unconscious selectivity. Although the preference in the study toward general conservativeness in the assigning of codes means that any error likely skews toward under-coding rather than over-coding, there are methods that could improve both how codes are assigned,

as well as how data are collected and sorted. These are addressed in the following section.

8.5 Directions for future research

As I observed previously, the study is unusually large for literary dialect research. For corpus analysis, however, the data set would be considered rather specialized and small. One reason for this is human limitation. As all of the data were collected and coded by hand, even at its current scale, those processes were labor intensive. But now that the coded data exists, there are new possibilities for extending the study's scope. One way to do this would be to keep the current time and speaker parameters, but to automate the data collection and coding. This could be accomplished by scraping the data from digital archives. Then, a classifier trained on the current data could identify works with speakers of interest. Dialogue could then be extracted, sorted, and coded, again using code trained on the current data set. Such processes could bootstrap this study's output using advances in statistical learning.

There are other ways this study could be expanded. A corpus of literary dialect produced by members of the same communities that are being voiced might produce contrasts to the one compiled for this study. It might also be interesting to use these or similar techniques to analyze how regional or national varieties have been historically represented. In carrying out the research, I encountered numerous representations of Irish, cockney, and French speakers, for example. Each of these, I would guess, would have very different historical trajectories. Like the ones analyzed here, those trajectories, too, are likely informed by the social and cultural conditions that mediate perceptions of specific vocal cultures and the symbols that are used to impersonate them.

8.6 Concluding remarks

In the introduction to this thesis, I referenced Blake (1981) and his assertion in *Non-Standard Language in English Literature* that the defining characteristic of literary dialect is the power of its social signaling (§1.6). This study's accumulated quantitative and qualitative evidence has shown just how perniciously and enduringly literary dialect can tap into and activate historical patterns of associations. Though the study stops at 1930, those patterns do not cease their propagation, of course. They continue on into the present.

They can be detected in a novel like Kathryn Stockett's (2009) *The Help* and the controversy surrounding her depiction African American domestic workers almost a century after the study concludes. The novel, set in the American South during the 1960s, follows the efforts of Eugenia "Skeeter" Phelan, the daughter of a wealthy white family, to compile the stories of the African American women who work in the homes of friends and neighbors. Stockett writes in three narrative voices: Eugenia's and those of two African American women, Aibileen Clark and Minny Jackson. In both Aibileen's and Minny's dialogue and their narration, Stockett makes use of literary dialect as in (6):

(6) "Law, Miss Hilly gone be here in five minutes. She better put that fire out fast." It feel crazy that we rooting for her. It's confusing in my mind.

Although Stockett eschews what is arguably the most indexical feature of African diasporic literary dialect (*t*/*d*-for-*th* substitution), a number of others are apparent: zero copula (*we rooting*), invariant present tense (*it feel*), and the religiously related insert (*law* for *lord*). Among its features, the novel's literary dialect also realizes *ain't* as a negator, invariant *be*, consonant cluster reduction, and syllable deletion.

These literary dialect practices prompted both praise and criticism (see Ruzich & Blake, 2015). On the one hand, there were plaudits for supposed naturalism, and on the other condemnation for evocations of stereotype. The Association of Black Women Historians (Jones, Berry, Gill, Gross, & Sumler-Edmond, 2011), for example, denounced the novel for "misrepresent[ing] African American speech and culture," specifically in its use of "child-like, over-exaggerated 'black' dialect." What makes *The Help* particularly compelling is not just this debate, but what that debate reveals about the tension between Stockett's stated purpose in writing the novel and its reception. In an afterword, Stockett asserts that one of her overriding motivations was to humanize and de-marginalize her novel's subjects. Yet, in giving voice to her African American characters, she taps into a centuries-old system of representation that undermines that very purpose. As the critique from the Association of Black Women Historians observes, Stockett's ventriloquizing evokes tropes of a contented and infantilized black servant class – tropes evident at least as far back as the 1743 Toby letter that was presented in the introduction (§1.5). However much Stockett may intend to counter stereotypes, her linguistic characterizations function to perpetuate them. And *The Help* is hardly unique in this regard. A range of contemporary media, from novels to films to television, participate in what Bloomquist (2015) terms "the

minstrel legacy" by rearticulating many of the patterns that this study has described. Neither is modern mimicry directed only at African diasporic speech communities. It continues in representations of Chinese (see, e.g., Chun, 2004; Chung, 2013) and Indian speakers (see, e.g., Davé, 2005; Gottschlich, 2011), as well.

Moreover, not only do these routines of mimicry project beyond the historical confines of this study, their consequences are similarly more than literary. These consequences are helpfully framed by Ribeiro's (2001, p. 166) critique of post-colonial studies in which he questions research that uses fiction as a proxy for people:

> También me llama la atención el uso acrítico de la literatura y la ficción (en general basado en el poder hermenéutico de las metáforas) como sustitutos de la realidad social y de investigaciones teóricas y metodológicas densas de las ciencias sociales. Esto levanta la cuestión de la posible existencia de ciencias sociales sin cientistas sociales, una problemática bastante complicada pues involucra factores epistemológicos, históricos y de poder interno de la academia.
>
> *I was also struck by the uncritical use of literature and fiction (usually based on the hermeneutic power of metaphor) as substitutes for the social reality and dense theoretical and methodological research in the social sciences. This raises the question of the possible existence of social sciences without social scientists, a rather complicated problem because it involves epistemological and historical factors internal to the academy.*

There is, I believe, an inverse danger in a study like this one that purposefully approaches literary dialect as a representational system without focusing on its accuracy. It can be easy to view these systems as abstractions without any material connections to communities in the world, to view them as self-contained fictions. Yet, such a view would run counter to one of this study's goals, which has been to expose how representations of vocal cultures are connected to political, economic, and social conditions – how they are implicated in the rationalizing of chattel slavery, the imperial conquest of India, and legal movements against the Chinese, for example. That dialect mimicry has effects on individuals and communities is made evident in the letter written by James Hewlett (§5.4). In that letter, Hewlett castigates Charles Mathews for his mockery of Hewlett and his colleagues at the African Theatre in New York. He closes by asking Mathews – and by implication the audiences who were laughing along with him – "Was this well for a brother actor?"

From distances, historical and methodological, literary dialects can seem like aesthetic curiosities: interesting to study, perhaps, but ultimately operating only in the fictional worlds for which they were created or the quantitative models that are used to explain them. However, part of "the minstrel legacy," Bloomquist (2015) points out, are the corrosive effects that historical routines of imitation continue to have on the communities they impersonate. In the United States, they shaped the response to

the Oakland Ebonics proposal that sought to treat African American English as a distinct language for educational purposes. The overwhelmingly negative reaction was informed by the representational patterns we have been examining, patterns that figure African diasporic vocal culture as disfluent and pathological (Baugh, 2000). In Singapore, a researcher published a description of Singapore English including *l*-for-*r* substitution, which he exemplified with the phrase "flied lice" for "fried rice" (Forbes, 1993). This prompted a reply from the faculty and students at the National University of Singapore in which they noted that the feature is not part of the phonological inventory of Singapore English, but is rather "an ancient stereotype of Chinese people speaking English" (Gupta, 1994). It certainly seems that the original author was neither malicious nor purposefully dishonest, but was so influenced by conventions of representing Chinese vocal culture that he perceived them, whether or not they were actually articulated. Examples like these speak to the enduring power of these patterns. In light of that power, we might approach James Hewlett's letter not as some distant artifact. We might hear it, instead, as entreaty to us all, as immediate and as relevant as the day it was written.

References

Corpus source works

Addison, H. R. (1858). *Traits and stories of Anglo-Indian life*. London: Smith, Elder & Co.

Ainsworth, W. H. (1839). *Jack Sheppard: A romance*. London: Richard Bentley, Printed by Samuel Bentley.

Almar, G. (1833). *The knights of St. John, or, the fire banner! A grand melo-drama*. London: J. Duncombe.

Anonymous. (1828). *Marly, or, a planter's life in Jamaica*. London: Hunt and Clarke.

Bage, R. (1792). *Man as he is. A novel. In four volumes*. London: printed for William Lane, at the Minerva Press.

Baillie, J. (1836). *The alienated manor*. London: Printed for Longman, Rees, Orme.

Bainbridge, M. (1843). *Rose of Woodlee*. London: Edward Bull.

Ballantyne, R. M. (1861). *The Golden Dream; or, adventures in the far west*. London: J.F. Shaw and Co.

Ballantyne, R. M. (1876). *Under the waves, or, diving in deep waters: A tale*. London: J. Nisbet & Co.

Ballantyne, R. M. (1888). *The Middy and the Moors: An Algerine story*. London: James Nisbet.

Banks, I. (1882). *Through the night, tales of shades and shadows*. London: Simpkin, Marshall, and Company.

Barker, L. D. S. (1874). *With a stout heart*. London: George Routledge and Sons.

Barker, M. H. (1839). *Hamilton King; or, the smuggler and the dwarf, by the old sailor*. London: Richard Bentley.

Bayly, A. E. (1899). *The house of strange secrets*. London: Sands & Co.

Besant, W. (1876a). *The case of Mr. Lucraft; and other tales*. London: Sampson Low, Marston, Searle & Rivington.

Besant, W. (1876b). *The golden butterfly*. London: Tinsley Bros.

Bickerstaff, I. (1768). *The padlock: A comic opera: As it is perform'd by his majesty's servants, at the Theatre-Royal in Drury-Lane*. London: printed for W. Griffin.

Boucicault, D. (1859). *The octoroon, or, life in Louisiana: A play in four acts*. London: Thomas Hailes Lacy, 89, Strand.

Brereton, F. S. (1912a). *The hero of Panama*. London: Blackie & Son.

Brereton, F. S. (1912b). *Under the Chinese dragon*. London: Blackie and Son.

Burke, T. (1916). *Limehouse nights: Tales of chinatown*. London: Grant Richards Ltd.

Burroughs, E. R. (1929). *The monster men*. Chicago: A. C. McClurg & Co.

Cobb, J. (1788). *Love in the east; or, adventures of twelve hours: A comic opera, in three acts. Written by the author of the strangers at home. As performed at the Theatre-Royal, Drury-Lane*. London: printed for W. Lowndes.

Collingwood, H. (1898). *A pirate of the Caribbees*. London: Griffith, Farran, Browne & Co.

Collingwood, H. (1915). *A Chinese command; a story of adventures in eastern seas* London: Blackie & Son.

Colman, G. (1787). *Inkle and yarico: An opera, in three acts. As performed at the Theatre-Royal in the Hay-Market, on Saturday, August 11th, 1787. Written by George Colman, junior*. London: printed for G. G. J. and J. Robinson.

Colman, G. (1808). *The Africans, or, war, love, and duty: A play, in three acts*. London: J. Cumberland.

Conrad, J. (1915). *Victory: An island tale*. London: Methuen & Co.
Croker, B. M. (1902). *The cat's paw*. London: Chatto & Windus.
Cupples, G. (1850). *The green hand: A "short" yarn*. New York: Harper & Bros.
Dallas, R. C. (1809). *Not at home: A dramatic entertainment*. London: Printed by J. & R. Hodson for B. Crosby.
Dell, E. M. (1919). *The lamp in the desert*. London: Hutchinson & Co.
Dibdin, C. (1779). *The mirror; or, harlequin every-where. A pantomimical burletta, in three parts. As it is performed at the Theatre-Royal in Covent-Garden*. London: printed for G. Kearsly.
Dimond, W. (1820). *The lady and the devil, a musical drama*. London: Printed for R.S. Kirby.
Donovan, D. (1900). *The adventures of Tyler Tatlock, private detective*. London: Chatto & Windus.
Dorling, H. T. (1916). *Stand by! Naval sketches and stories*. London: Pearson.
Doyle, A. C. (1890). *The captain of the polestar; and other tales*. London: Longmans & Co.
Doyle, C. W. (1900). *The shadow of Quong Lung*. London: J.B. Lippincott Company.
Fenn, G. M. (1893). *Blue jackets, or, the log of the Teaser*. London: Griffith Farran & Co.
Fitzball, E. (1829). *The flying dutchman; or, the phantom ship: A nautical drama*. London: John Cumberland.
Foote, S. (1778). *The cozeners, a comedy, of three acts, as it was performed at the Theatre Royal in the Hay-Market*. London: Printed for J. Wheble.
Foskett, S. (1915). *The temple in the tope*. London: Hodder and Stoughton.
Frith, H. (1898). *In the yellow sea*. London: Griffith, Farran, Browne & Co.
Galsworthy, J. (1924). *The forest: A drama in four acts*. London: Duckworth.
Garrick, D. (1772). *The Irish widow. In two acts. As it is performed at the Theatre Royal in Drury-Lane*. London: printed for T. Becket.
Giles, E. W. E. (1897). *China coast tales*. Singapore: Kelly and Walsh.
Grant, J. (1868). *First love & last love: A tale of the Indian mutiny*. London: George Routledge & Sons.
Harraden, B. (1908). *Interplay*. London: Methuen & Co.
Harrison, A. S. (1867). Chota sabib Charlie. In E. Walford (Ed.), *Once a week* (Vol. 4, pp. 721-729, 751-758). London: Bradbury and Evans.
Hemyng, B. (1900). *Jack harkaway's boy tinker among the turks*. Chicago: M.A. Donohue.
Henty, G. A. (1889). *Tales of daring and danger*. London: Blackie & Son.
Herbert, H. W. (1843). *My shooting box*. Philadelphia: T. B. Peterson.
Hoare, P. (1793). *The prize or, 2, 5, 3, 8, a musical farce in two acts, as performed by his majesty's company. Written by Prince Hoare, esq. The music by storace. Correctly taken from the manager's book*. Dublin: printed for F. Farquhar.
Hockley, W. B. (1828). *The English in India, by the author of 'pandurang hari'*. London: Printed for W. Simpkin and R. Marshall.
Hofland, B. (1816). *Matilda, or, the Barbadoes girl a tale for young people*. London: Printed at the Minerva Press for A.K. Newman and Co.
Holman, J. G. (1799). *The votary of wealth; a comedy, in five acts. As performed at the Theatre-Royal, Covent-Garden*. London: T. N. Longman and O. Rees.
Hook, T. E. (1840). *Precepts and practice*. London: H. Colburn.
Howard, E. (1836). *Rattlin, the Reefer*. London: Richard Bentley.
Hume, F. (1898). *For the defense*. Chicago: Rand, McNally & Co.

Hutcheson, J. C. (1889). *The black man's ghost*. London: Ward, Lock, and Co.
Hutcheson, J. C. (1890). *Afloat at last: A sailor boy's log of his life at sea*. London: Blackie & Son.
Hyne, C. J. C. W. (1899). *Further adventures of Captain Kettle*. London: Arthur Pearson.
Inchbald, E. (1805). *To marry or not to marry, a comedy*. London: Longman, Hurst, Rees, and Orme.
Jacob, E. W. (1863). *Something new; or, tales for the times*. London: Emily Faithfull.
Jenkins, J. E. (1877). *Lutchmee and Dilloo, a study of West Indian life*. London: W. Mullan & Son.
Jerome, J. K. (1900). *Three men on the bummel*. Bristol: Arrowsmith.
Jones, H. A. (1899). *Carnac Sahib; an original play in four acts*. London: Macmillan Co.
K. (1878). The Indian famine. In C. M. Yonge (Ed.), *The monthly packet of evening readings for members of the English church* (Vol. 25, pp. 60-65, 145-151, 260-268, 348-354, 458-466, 551-555). London: Mozley and Smith.
Kaye, J. W. (1844). *Peregrine Pultuney; or, life in India* (Vol. 2). London: John Mortimer.
Kelly, H. (1774). *The romance of an hour, a comedy of two acts, as it is performed, with universal applause, at the Theatre Royal in Covent Garden*. London,: G. Kearsley.
Kingston, W. H. G. (1876). *Twice lost a story of shipwreck, and of adventure in the wilds of Australia*. London: T. Nelson and Sons.
Kipling, R. (1901). *Kim*. London: Macmillan and Co.
Le Queux, W. T. (1917). *The secrets of potsdam* (2nd impr. ed.). London: London Mail.
Leakey, C. W. (1859). *The broad arrow, by oliné keese*. London: Richard Bentley.
Levett-Yeats, S. K. (1897). *A galahad of the creeks, and other stories*. London: Griffith Farran & Co.
Lewis, M. G. (1808). *Romantic tales*. London: Printed by D.N. Shury for Longman, Hurst, Rees, and Orme.
Little, A. H. N. (1902). *Out in China!* London: Treherne & Co.
Mackenzie, H. (1777). *Julia de Roubigné, a tale. In a series of letters. Published by the author of the man of feeling, and the man of the world*. London: printed for W. Strahan; and T. Cadell.
Macready, W. (1793). *The Irishman in London; or, the happy African. A farce. In two acts. Performed at the Theatre-Royal, Covent-Garden*. London: printed by W. Woodfall, and sold by T. N. Longman.
Marryat, F. (1836). *Mr. Midshipman Easy*. London: Saunders and Otley.
Marryat, F. (1874). *Sybil's friend and how she found him*. London: G. Routledge.
Maugham, W. S. (1922). *East of Suez: A play in seven scenes*. London: William Heinemann.
Meade, E. T. (1897). *Under the dragon throne*. London: Gardner, Darton & Co.
Middleton, L. F. F. M. (1922). *The happy adventurers*. London: Blackie & Son.
Milner, C., & Stirling, E. (1837). *Don Juan: A musical drama in three acts*. London: W. Strange.
Moncrieff, W. T. (1826). *Tom and Jerry; or life in London. An operatic extravaganza. With a key, vocabulary, &c*. London: W.T. Moncrieff.
Moore, H. C. (1906). *Afloat on the Dogger Bank; a story of adventure in the North Sea and in China*. Boston: D. Estes & Company.

Morton, J. M. (1845). *The mother and child are doing well. A farce*. London: The National Acting Drama Office.
Mundy, T. (1926). *The devil's guard*. Toronto: McClelland & Stewart.
Murray, W. (n.d.). Obi, or, three-fingered jack a melo-drama, in two acts. London: Thomas Hailes Lacy.
Neale, W. J. (1833). *The port admiral: A tale of the war*. London: Cochrane and M'Crone.
Nesbit, E. (1904). *New treasure seekers*. London: E. Benn.
O'Keeffe, J. (1783). *The dead alive: A comic opera. In two acts. As it is performed at the theatres in London and dublin*. Dublin: Sold by the Booksellers.
O'Keeffe, J. (1789). *The highland reel: A comic opera. In three acts. As it is performed at the theatres-royal in London and dublin.* Dublin: printed by T. M'Donnel.
Oppenheim, E. P. (1915). *The black box*. New York: Grosset & Dunlap.
Oxenford, J. (1837). *No followers, a burletta*. London: Published for the proprietor by W. Strange.
Pardoe, J. S. H. (1858). *The poor relation*. London: Hurst and Blackett.
Paull, S. M. (1879). *Levelsie manor*. London: Hodder and Stoughton.
Payn, J. (1867). *Carlyon's year: A novel*. New York: Harper.
Peake, R. B. (1824). *Jonathan in England (Americans abroad)*. Lord Chamberlain's Plays. Vol. IV. Aug.-Oct. 1824, (42868). British Library, London.
Penny, F. E. (1905). *Dilys; an Indian romance*. London: Chatto & Windus.
Phillips, E. C. (1882). *Peeps into China: Or, the missionary's children*. London: Cassell, Petter, Galpin & Co.
Pilon, F. (1779). *The Liverpool prize; a farce: In two acts. As it is performed at the Theatre-Royal, in Covent-Garden, with universal applause. Written by f. Pilon*. London: printed for T. Evans.
Pocock, I. (1817). *Robinson Crusoe; or, the bold bucaniers: A romantic melo-drama. Produced, for the first time, at the Theatre-Royal, Covent-Garden, easter monday, 1817*. London: J. Miller.
Pratt, S. (1790). *The new cosmetic or the triumph of beauty, a comedy*. London: printed for the author; and sold by Cadell; Egerton; Harlow; Richardson; Bew; and Trueman and Son, Exeter.
Rae, M. (1912). *A bottle in the smoke: A tale of Anglo-Indian life*. London: Hodder and Stoughton.
Rafter, M. (1855). *Percy Blake; or, the young rifleman*. London: Hurst and Blackett.
Reade, C. (1863). *Hard cash: A matter-of-fact romance*. London: Sampson Low, Son, & Marston.
Rohmer, S. (1913). *The mystery of Dr. Fu-Manchu*. London: Methuen & Co.
Rohmer, S. (1919). *Dope: A story of chinatown and the drug traffic*. London: Cassell.
Russell, W. C. (1887). *The frozen pirate*. London: Low, Marston, Searle, and Rivington.
Russell, W. C. (1891). *My Danish sweetheart*. London: Methuen & Co.
Scott, M. (1834). *Tom Cringle's log* (Second edition ed.). Edinburgh: William Blackwood.
Shiel, M. P. (1898). *The yellow danger*. London: Grant Richards.
Shipp, J. (1832). *The k'haunie kineh-walla; or, eastern story-teller: A collection of tales*. London: printed for Longman & Co.
Somerset, C. A. (1828). *The sea!* London: John Cumberland.

Stables, W. G. (1883). *Wild adventures round the pole; or, the cruise of the Snowbird crew in the Arrandoon*. London: Hodder and Stoughton.

Starke, M. (1788). *The sword of peace; or, a voyage of love; a comedy, in five acts. First performed at the Theatre Royal in the Hay Market*. London: printed for J. Debrett.

Strang, H. (1912). *The flying boat: A story of adventure and misadventure*. London: Henry Frowde.

Taylor, T. (1860). *Up at the hills: An original comedy of Indian life, in two acts*. London: Thomas Hailes Lacy.

Tracy, L. (1915). *Number seventeen*. New York: Edward J. Clode.

Trusler, J. (1793). *Life; or, the adventures of William Ramble, esq. With three frontispieces, designed by Ibbetson, ... And two new and beautiful songs, with the music by Pleyel and Sterkel. By the author of modern times; or, the adventures of Gabriel Outcast. In three volumes*. London: printed for Dr. Trusler, and sold at the Literary Press.

Vachell, H. A. (1912). *Bunch grass: A chronicle of life on a catte ranch*. London: John Murray.

Wallace, E. (1916). *The tomb of Ts'in*. London: Ward, Lock & Co.

Westerman, P. F. (1922). *The wireless officer*. London: Blackie & Son.

Wren, P. C. (1920). *Cupid in Africa, or the making of Bertram in love and war*. London: Heath Cranton.

Archival sources

African amusements. (1821, September 21). *National Advocate,* p. 2.
Africans. (1821, August 3). *National Advocate,* p. 2.
Almon, J. (Ed.) (1769). *The political register, and impartial review of new books*. London: Printed for Henry Beevor.
An Anglo-Indian. (1907, September 14). Baboo English. *The Spectator*.
Anderson, A. (1795). *A narrative of the British embassy to China in the years 1792, 1793, and 1794; containing the various circumstances of the embassy, with accounts of customs and manners of the Chinese*. London: printed for J. Debrett.
Banim, J. (1825). *Tales, by the O'Hara family*. London: Printed for W. Simpkin and R. Marshall.
Barnard, E. A. (1882). *Maple range; a frontier romance*. Chicago: H. A. Sumner.
Basu, B. D. (1922). *History of education in India under the rule of the East India Company*. Calcutta: The Modern Review Office.
Bell, H. H. (1897). His highness Prince Kwakoo. In J. K. Jerome (Ed.), *The idler magazine* (Vol. 10, pp. 685-696). London: Chatto & Windus.
Boucicault, D. (1861a, November 20). The Octoroon. *The Times,* p. 5.
Boucicault, D. (1861b). Unpublished note. London: Theatre Museum.
Bower, A. (1732). *Historia litteraria: Or, an exact and early account of the most valuable books published in the several parts of Europe. ... With a compleat alphabetical index* (Vol. 3). London: printed for N. Prevost, over-against Southampton-street, in the Strand; and E[dward]. Symon, in Cornhill.
Bowker, R. R. (1900, January 27). Fiction. *The publishers' weekly,* 193-195.
Bridgman, E. C. (Ed.) (1836). *The Chinese repository*. Canton: Printed for the proprietors.
Broughton, H. (1864, May). Letters from a competition wallah. Letter XII. And last. -- education of India since 1835 (with a hiterto unpublished minute of Lord Macaulay). *Macmillan's Magazine, 55,* 1-17.
Brown, T. A. (1870). *History of the American stage. Containing biographical sketches of nearly every member of the profession that has appeared on the American stage, from 1733 to 1870*. New York: Dick & Fitzgerald.
Buckstone, J. B. (1828). *A new Don Juan!: An operatical, satirical, poetical, egotistical, melo-dramatical, extravaganzacal, but strictly moral burletta, in two acts (founded on Lord Byron's celebrated poem)*. London: T. Richardson.
Burke, W. (1896). The Anglo-Irish dialect. In J. F. Hogan (Ed.), *The Irish ecclesiastical record* (pp. 694-704). Dublin: Browne & Nolan.
Chambers, W., & Chambers, R. (Eds.). (1880). *Chamber's encyclopedia: A dictionary of universal knowledge for the people, vol. 8* (Fifteenth American ed.). London: W. and R. Chambers.
Colburn, H. (Ed.) (1833). *The united service journal and naval and military magazine*. London: Richard Bentley.
Cooke, G. W. (1858). *China: Being "the times" special correspondence from China in the years 1857-58*. London: G. Routledge.
de Leon, T. C. (1897, November). The day of dialect. *Lippincott's Monthly Magazine,* 679-683.
De Quincey, T. (1857). *China*. Edinburgh: J. Hogg.
Delano, A. (1857). *A live woman in the mines; or, Pike County ahead!* New York: S. French.

Dickens, C. (1837). *The posthumous papers of the pickwick club*. London: Chapman and Hall.

Dicks, J. T. (1884). *List of Dicks' Standard Plays and free acting drama*. London: John Dicks Press Ltd.

Domestic news. (1747, December 5). *Jacobite's Journal*.

Downing, C. T. (1838). *The fan-qui in China, in 1836-7*. London: Henry Colburn.

Doyle, C. W. (1899). *The taming of the jungle*. Westminster: A. Constable.

Drury-Lane Theatre. (1828, February 20). *The Morning Chronicle,* p. 3.

Dyche, T. (1740). *A new general English dictionary; peculiarly calculated for the use and improvement of such as are unacquainted with the learned languages. ... Originally begun by the late reverend Mr. Thomas Dyche, ... And now finish'd by William Pardon, gent* (The third edition, with the addition of the several market towns .. ed.). London: printed for Richard Ware.

The Englishman in China. (1860). London: Sounders, Otley & Co.

A fast-day at Foxden. (1864, June). *The Atlantic Monthly, 13,* 676-693.

Fernald, C. B. (1907). *John Kendry's idea*. New York: Outing Publishing Co.

For massatuse pie. (1782, January 10). *Massachusetts Spy,* p. 2.

Gibbes, P. (1789). *Hartly house, Calcutta*. Dublin: William Jones.

Hadley, G. (1796). *A compendious grammar of the current corrupt dialect of the jargon of Hindostan, (commonly called moors); with a vocabulary, English and moors, moors and English. ... By George Hadley* (The fourth edition corrected and much enlarged. ed.). London: printed for J. Sewell.

Hamilton, W. (1872, December 14). "Pidgin" English. *Pro and Con: A Journal for Literary Investigation,* 2.

Harrison, J. A. (1892). Negro-English. *Modern Language Notes, 7*(2), 62.

Hardy, R. S. (1863). *The sacred books of the Buddhists compared with history and modern science*. Columbo: Wesleyan Mission Press.

Henderson, H. B. (1829). *The Bengalee: Or, sketches of society and manners in the east*. London: Smith, Elder.

Home, J. (1757). *Douglas: A tragedy. As it is acted at the Theatre-Royal in Covent-Garden*. Edinburgh: printed for G. Hamilton & J. Balfour, W. Gray & W. Peter.

Hutton, L. (1889). The Negro on the stage. *Harper's New Monthly Magazine, 79,* 131-145.

J. V. P. (1861, December). The dramatic and musical world of London. *Baily's magazine of sports and pastimes,* 50-54.

Jenkins, E. (1871). *The coolie, his rights and wrongs: Notes of a journey to British Guiana, with a review of the system and of the recent commission of inquiry*. London: Strahan & co.

Jerdan, W. (Ed.) (1821). *The literary gazette and journal of belles lettres, arts, sciences, &c*. London: W.A. Scripps.

Kingdon, G. R. (1895). Some notes on pronunciation. *The Irish Monthly, 23*(261), 145-156.

Kitchiner, W. (1823). *The sea songs of Charles Dibdin: With a memoir of his life and writings*. London: Printed for G. and W. B. Whittaker.

Knight, C. (Ed.) (1838). *The penny magazine of the society for the diffusion of useful knowledge* (Vol. 7). London: Charles Knight.

Kingston, M. H. (1987). *Tripmaster monkey: His fake book*. New York: Vintage International.

Leman, W. M. (1886). *Memories of an old actor*. San Francisco: A. Roman Co.

Lentzner, K. A. (1891). *Colonial English: A glossary of Australian, Anglo-Indian, pidgin English, West Indian, and South African words*. London: Kegan Paul, Trench, Trübner & Co., Ltd.

A letter from Cuffee to the printer, relative to the Negro-bill which did not pass. (1785, March 31). *New-York Packet,* p. 3.

Literature. (1836, September 8). *The Morning Post*.

Lockhart, J. G. (1838). *Memoirs of the life of Sir Walter Scott*. Paris: published by A. and W. Galignani and Co.

Lockhart, J. G. (Ed.) (1830). *The quarterly review*. London: John Murray.

Lord Chamberlain's plays. Vol. IV. Aug.-Oct. 1824. (1824, Aug. 1824-Oct. 1824). Retrieved from Nineteenth Century Collections Online database (42868). British Library.

Lord Chamberlain's plays. Vol. XXV. Jan.-March 1828. (1828, Jan. 1828-Mar. 1828). Retrieved from Nineteenth Century Collections Online database (42889). British Library.

Lung, W. (1893). A chinaman's beekeeping. *British Bee Journal, 21*(551), 29-30.

Lynn, T. (1821). *Star tables, for more readily ascertaing the latitude and longitude at sea during the night*. London: Printed for the Author.

Mackay, C. (1865, April 22). The London theatre. *The London review and weekly journal of politics, society, literature, art, & science,* 431-432.

Markby, W. (1907, September 21). Baboo English. *The Spectator,* 393-394.

Marshall, J. (1812). *Newcastle songster; being a choice collection of songs, descriptive of the language and manners of the common people of Newcastle upon tyne and the neighbourhood* Retrieved from http://www.asaplive.com/archive/detail.asp?id=N0601402

Mathews, C. (1824). *The London Mathews; containing an account of this celebrated comedian's trip to America*. London: Hodgson.

Matthews. (1824, May 8). Editorial. *The National Advocate*.

Milner, H. M., & Reeve, G. W. (1828). *The songs, duets, trios, glees, chorusses, &c. In the new operatic entertainment, founded on the first six cantos of Lord Byron's celebrated poem of Don Juan, and called juan's early days*. London: Printed for John Lowndes.

Morton, T. (1797). *A cure for the heart-ache: A comedy in five acts: As performed at the Theatre-Royal, Covent Garden*. London: Printed for T.N. Longman.

Mowry, W. A. (1887). Education, a monthly magazine devoted to the science, art, philosophy and literature of education (pp. v.). Boston: Palmer Company.

Murphy, A., & Du Halde, J. B. (1759). *The orphan of China: A tragedy, as it is perform'd at the Theatre-Royal, in Drury-Lane*. London: P. Vaillant.

Murphy, J. R. (1899). The survival of African music in America. *Appleton's Popular Science Monthly, 55,* 660-672.

A Negro, from the coast of coromandel. (1792, October 10). *Columbian Centinel*.

Old Footlights. (1861, December 1). The green room. *Saunders, Otley & Co.'s literary budget for England, India, China, Australia and the colonies,* 76-77.

Opium dens in London. By an ex-member of the London police force. (1910, July 16). *Penny Illustrated Paper,* p. 90.

Parker, A. A. (1835). Trip to the west and texas comprising a joruney of eight thousand miles, through New York, Michigan, Illinois, Missouri, Louisiana and Texas, in the autumn and winter of 1834-5. Interspersed with anecdotes, incidents and observations. Concord: White & Fisher.

Parrish, R. (1908). *Prisoners of chance; the story of what befell Geoffrey Benteen, borderman, through his love for a lady of France*. Chicago: A.C. McClurg & Co.

Parrish, R. (1918). *Wolves of the sea; being a tale of the colonies from the manuscript of one Geoffry Carlyle, seaman, narrating certain strange adventures which befell him aboard the pirate craft "Namur"*. Chicago: A. C. McClurg.

Phillips, R. (Ed.) (1837). *The monthly magazine of politics, literature, art, science, and the belle lettres*. London: Sherwood, Gilbert, and Piper.

Planché, J. R. (1848). *The king of the peacocks: An original fairy extravaganza in two acts*. London: Thomas Hailes Lacy.

Pollock, A. W. A. (1863). The united service magazine. London: Hurst and Blackett, Publishers.

Ramsay, A. (1725). *The gentle shepherd; a Scots pastoral comedy. By Allan Ramsay*. Edinburgh: printed by Mr. Tho. Ruddiman, for the author, and by Mr. Thomas Longman, and Mr. James McEwin, London, and by Mr. Alexander Carmichael in Glasgow.

Ramsden, R. (1841). *The triumphs of truth; or, facts displaying the value and power of the word of god, more especially ... In the operations of the British and foreign bible society* (3rd. ed.). Lond.

Roberts, E. (1830). Indian scenes: Shopping. *The Athenaeum*, 570-571.

Rockwell, C. (1842). *Sketches of foreign travel, and life at sea; including a cruise on board a man-of-war, as also a visit to Spain, Portugal, the south of France, italy*. Boston: Tappan and Dennet.

Rogers, J. W. F. (1883). *Grammar and logic in the nineteenth century: As seen in a syntactical analysis of the English language*. London: Trübner and Co.

Rohmer, S. (1922). *Tales of chinatown*. London: Cassell.

Ruxton, G. F. A. (1849). *Life in the far west*. Edinburgh: W. Blackwood.

Saunders, J., & Marston, W. (1862). The national magazine. London: W. Tweedie.

Short notes on the West Indies. (1845). *Chambers's Edinburgh Journal, 53,* 3-6.

Snowden, W. W., Sigourney, L. H., & Embury, E. C. (1840). Ladies companion and literary expositor, a monthly magazine embracing every department of literature. New York: W. W. Snowden.

Stanford University. (2015). Dime novel and story paper collection. Retrieved March 15, 2014, from Stanford University http://library.duke.edu/digitalcollections/hasm/

Stockett, K. (2009). *The help*. New York: Amy Einhorn Books.

Stocqueler, J. H. (1844). *The hand-book of India, a guide to the stranger and the traveller, and a companion to the resident*. London: Wm. H. Allen & Co.

Strand theatre. (1837, September 6). Article. *The Standard,* p. 1. Retrieved from http://tinyurl.galegroup.com/tinyurl/f9mk3

Terry, E. (1655). *A voyage to East-India: Wherein some things are taken notice of in our passage thither, but many more in our abode there, within that rich and most spacious empire of the the great mogol*. London: Printed by T.W. for J. Martin, and J. Allestrye, ...

The London Missionary Society. (1847). The evangelical magazine and missionary chronicle (Vol. 25). London: Ward and Co.

Townsend, M., & Hutton, R. H. (1861, November 23). A new sensation drama. *The spectator: a weekly review of politics, literature, theology, and art,* 1284.

Tuckerman, H. T. (1854). *A month in England*. London: Richard Bentley.

Twain, M. (1872). *Roughing it*. Hartford: American Publishing Company.

Weiss, J. (1863). The horrors of Santo Domingo. *The Atlantic Monthly, 11,* 289-306.

Wheble, J. (Ed.) (1799). *The sporting magazine; or monthly calendar of the transactions of the turf, the chace, and every other diversion interesting to the man of pleasure and enterprize*. London England: printed for the proprietors, and sold by J. Wheble, no. 18. Warwick Square, Warwick Lane, near St. Paul's.

Whitney, A. (1878). *Almond-eyed; a story of the day*. San Francisco: A. L. Bancroft & Company.

Wied, H., & The British and Foreign Bible Society. (1829). *Da njoe testament va wi masra en helpiman Jesus Christus*. London: W. M'Dowall, printer.

Yule, H., & Burnell, A. C. (1886). *Hobson-Jobson: Being a glossary of Anglo-Indian colloquial words and phrases, and of kindred terms; etymological, historical, geographical, and discursive*. London: John Murray.

Secondary sources

Abberley, W. (2015). *English fiction and the evolution of language, 1850-1914*. Cambridge: Cambridge University Press.

Agha, A. (2003). The social life of cultural value. *Language & Communication, 23*(3–4), 231-273.

Agha, A. (2007). *Language and social relations*. Cambridge, UK: Cambridge University Press.

Alim, H. S. (2006). *Roc the mic right: The language of hip hop culture*. New York: Routledge.

Altick, R. D. (1958). From Aldine to Everyman: Cheap reprint series of the English classics 1830-1906. *Studies in Bibliography, 11*, 3-24.

Anand, A. S. (2011). Cosmopolitanism in Hobson-Jobson: Remaking imperial subjects. *Comparative Studies of South Asia, Africa and the Middle East, 31*(2), 521-537.

Anand, D. (2007). Western colonial representations of the other: The case of exotica Tibet. *New Political Science, 29*(1), 23-42.

Appel, J. J. (1957). Jewish literary dialect. *American Speech, 32*(4), 313-314.

Ardis, A. L., & Collier, P. (2008). *Transatlantic print culture, 1880-1940: Emerging media, emerging modernisms*. Basingstoke: Palgrave Macmillan.

Bailey, G., Tiller, J., & Andres, C. (2005). Some effects of transcribers on data in dialectology. *American Speech, 80*(1), 3-21.

Bailey, R. W. (1991). *Images of English: A cultural history of the language*. Ann Arbor: University of Michigan Press.

Baker, P. (2014). Considering context when analysing representations of gender and sexuality: A case study. In J. Flowerdew (Ed.), *Discourse in context: Contemporary applied linguistics, volume 3* (pp. 27-48). London: Bloomsbury Academic.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273-306.

Barry, B. (2001). 'It's hard fuh me to understand what you mean, de way you tell it': Representing language in Zora Neale Hurston's Their Eyes were Watching God. *Language and Literature, 10*(2), 171-186.

Bataille, R. R. (2000). *The writing life of Hugh Kelly: Politics, journalism, and theater in late-eighteenth-century London*. Carbondale: Southern Illinois University Press.

Baugh, J. (2000). *Beyond Ebonics: Linguistic pride and racial prejudice*. Oxford: Oxford University Press.

Beal, J. C. (2000). From Geordie Ridley to Viz: Popular literature in Tyneside English. *Language and Literature, 9*(4), 343-359.

Bell, A., & Gibson, A. (2011). Staging language: An introduction to the sociolinguistics of performance. *Journal of Sociolinguistics, 15*(5), 555-572.

Bhabha, H. (1984). Of mimicry and man: The ambivalence of colonial discourse. *October, 28*, 125-133.

Bhattacharya, N. (2006). *Slavery, colonialism, and connoisseurship: Gender and eighteenth-century literary transnationalism*. Aldershot, Hants, England ; Burlington, Vt.: Ashgate.

Biber, D., & Burges, J. (2000). Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics, 28*(1), 21-37.

Birnbaum, M. (1991). Dark dialects: Scientific and literary realism in Joel Chandler Harris's Uncle Remus series. *The New Orleans review., 18*(1), 36.

Blake, N. F. (1981). *Non-standard language in English literature*. London: André Deutsch.

Blank, P. (1996). *Broken English: Dialects and the politics of language in renaissance writings*. London: Routledge.

Bloomquist, J. (2015). The minstrel legacy: African American English and the historical construction of "black" identities in entertainment. *Journal of African American Studies, 19*(4), 410-425.

Bolton, K. (2000). Language and hybridization: Pidgin tales from the China coast. *Interventions, 2*(1), 35-52.

Bolton, K. (2002). Chinese Englishes: From Canton jargon to global English. *World Englishes, 21*(2), 181-199.

Bolton, K. (2003). *Chinese Englishes: A sociolinguistic history*. Cambridge: Cambridge University Press.

Boskin, J. (1986). *Sambo: The rise & demise of an American jester*. New York: Oxford University Press.

Brantlinger, P. (1988). *Rule of darkness: British literature and imperialism, 1830-1914*. Ithaca: Cornell University Press.

Bratton, J. S. (1981). English Ethiopians: British audiences and black-face acts, 1835-1865. *The Yearbook of English Studies, 11*, 127-142.

Brody, J. D. (1998). *Impossible purities: Blackness, femininity, and Victorian culture*. Durham: Duke University Press.

Brook, T., & Wakabayashi, B. T. (Eds.). (2000). *Opium regimes: China, Britain, and Japan, 1839-1952*. Berkeley: University of California Press.

Brown, D. W. (2005). *Public discourse and debate about African American English*. Paper presented at the Modern Language Association, Washington, DC.

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics, 32*(10), 1439-1465.

Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics, 7*(3), 398-416.

Bucholtz, M., & Lopez, Q. (2011). Performing blackness, forming whiteness: Linguistic minstrelsy in Hollywood film. *Journal of Sociolinguistics, 15*(5), 680-706.

Burkette, A. (2001). The use of literary dialect in Uncle Tom's Cabin. *Language and Literature, 10*(2), 158-170.

Buzelin, H., & Winer, L. (2008). Literary representations of creole languages: Cross-linguistic perspectives from the Caribbean. In S. Kouwenberg & J. V. Singler (Eds.), *The handbook of pidgin and creole studies* (pp. 637-665). Oxford: Blackwell Publishing.

Caldas-Coulthard, C. R., & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society, 21*(2), 99-133.

Caldwell, D. (1971). The negroization of the Chinese stereotype in California. *Southern California Quarterly, 53*(2), 123-131.

Carkeet, D. (1979). The dialects in Huckleberry Finn. *American Literature, 51*(3), 315-332.

Carlson, J. A. (2007). New lows in eighteenth-century theater: The rise of Mungo. *European Romantic Review, 18*(2), 139-147.
Chakravarty, G. (2005). *The Indian mutiny and the British imagination*. Cambridge: Cambridge University Press.
Chen, Y. (2000). *Chinese San Francisco, 1850-1943: A trans-pacific community*. Stanford: Stanford University Press.
Chun, E. (2004). Ideologies of legitimate mockery: Margaret Cho's revoicings of Mock Asian. *Pragmatics, 14*(2-3), 263-289.
Chung, H. S. (2013). From 'me so horny' to 'i'm so ronery': Asian images and yellow voices in American cinema. In J. Jaeckle (Ed.), *Film dialogue* (pp. 172-191). New York: Columbia University Press.
Cooley, M. (1997). An early representation of African American English. In C. G. Bernstein, T. Nunnally, & R. Sabino (Eds.), *Language variety in the South revisited* (pp. 51-58). Tuscaloosa: University of Alabama Press.
Cooper, P. (2013). *Enregisterment in historical contexts: A framework.* (PhD), University of Sheffield., Sheffield.
Coupland, N. (2003). Sociolinguistic authenticities. *Journal of Sociolinguistics, 7*(3), 417-431.
Coupland, N. (2007). *Style: Language variation and identity*. Cambridge: Cambridge University Press.
Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29-59.
Cutler, C. L. (1994). *O brave new words!: Native American loanwords in current English*. Norman: University of Oklahoma Press.
Darwin, J. (1999). A third British empire? The dominion idea in imperial politics. In J. M. Brown & W. R. Louis (Eds.), *The Oxford history of the British empire v 4* (Vol. The twentieth century, pp. 64-87). Oxford: Oxford University Press,.
Davé, S. (2005). Apu's brown voice: Cultural inflection and South Asian accents. In S. Davé, N. LeiLani, & T. Oren (Eds.), *East main street: Asian American popular culture* (pp. 313-336). New York: New York University Press.
Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora, 7*(2), 121-157.
Davies, M. (2013). *The 155 billion word Google books 'corpus': Can it be used for serious research on diachronic syntax?* Paper presented at the Studies in the History of the English Language, Provo, Utah.
Davis, T. C. (2011). Acting black, 1824: Charles Mathews's trip to America. *Theatre Journal, 63*(2), 163-189.
de Cillia, R., Reisigl, M., & Wodak, R. (1999). The discursive construction of national identities. *Discourse & Society, 10*(2), 149-173.
De Haan, P. (1996). More on the language of dialogue in fiction. *ICAME Journal, 20*, 23-40.
De Smet, H. (2005). A corpus of late modern English texts. *ICAME Journal, 29*, 69-82.
Dennison, S. (1982). *Scandalize my name: Black imagery in American popular music*. New York: Garland Publishing.
Dicks, G. (2006). *The John Dicks Press*. U.S.: Lulu.com.
Earle, W., & Aravamudan, S. (2005). *Obi, or, the history of Three-Fingered Jack*. Peterborough: Broadview Press.
Eckert, P. (2003). Elephants in the room. *Journal of Sociolinguistics, 7*(3), 392-397.

Elliott, R. W. V. (1974). *Chaucer's English*. London: Deutsch.
Ellis, M. (1994). Literary dialect as linguistic evidence: Subject-verb concord in nineteenth-century southern literature. *American Speech, 69*(2), 128-144.
Erickson, L. (1996). *The economy of literary form: English literature and the industrialization of publishing, 1800-1850*. Baltimore: Johns Hopkins University Press.
Erll, A. (2006). Re-writing as re-visioning. *European Journal of English Studies, 10*(2), 163-185.
Errington, J. J. (2008). *Linguistics in a colonial world: A story of language, meaning, and power*. Oxford: Blackwell.
Fairclough, N., Mulderrig, J., & Wodak, R. (2011). Critical discourse analysis. In T. A. van Dijk (Ed.), *Discourse studies: A multidisciplinary introduction* (pp. 357-378). London: Sage Publications.
Feather, J. (2006). *A history of British publishing* (2nd ed.). London: Routledge.
Fenno, C. R. (1983). Nineteenth-century Illinois dialect: Robert casey. *American Speech, 58*(3), 244-254.
Ferguson, S. L. (1998). Drawing fictional lines: Dialect and narrative in the Victorian novel. *Style, 32*(1), 1-17.
Fisher, M. H. (2011). Making London's "oriental quarter". In G. Pandey (Ed.), *Subalternity and difference: Investigations from the north and the south* (pp. 79-96). London: Routledge.
Fitzgerald, J. (2007). *Big white lie: Chinese Australians in white Australia*. Sydney: University of New South Wales Press.
Forbes, D. (1993). Singlish. *English Today, 9*(02), 18-22.
Forman, R. G. (2013). *China and the Victorian imagination: Empires entwined*. Cambridge: Cambridge University Press.
Foucault, M. (1971). *The order of things: An archaeology of the human sciences* ([1st American ed.). New York: Pantheon Books.
Foucault, M. (1972). *The archaeology of knowledge* (1st American ed.). New York: Pantheon Books.
Foucault, M. (1977). *Discipline and punish: The birth of the prison* (1st American ed.). New York: Pantheon Books.
Fulford, T., & Kitson, P. (1998). *Romanticism and colonialism: Writing and empire, 1780-1830*. Cambridge: Cambridge University Press.
García-Bermejo Giner, M. F., & Montgomery, M. (2001). Yorkshire English two hundred years ago. *Journal of English Linguistics, 29*(4), 346-362.
Gibbs, J. M. (2014). *Performing the temple of liberty slavery, theater, and popular culture in London and Philadelphia, 1760-1850*. Baltimore: Johns Hopkins University Press.
Gottschlich, P. (2011). *Apu, Neela, and Amita stereotypes of Indian Americans in mainstream tv shows in the United States.* Paper presented at the Internationales Asien Forum. International Quarterly for Asian Studies.
Green, L. J. (2002). *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.
Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437.
Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. *Language and Computers, 71*(1), 197-212.
Gries, S. T., & Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora, 3*(1), 59-81.

Griffiths, A., Robinson, L. A., & Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation, 40*(3), 175-205.

Gupta, A. F. (1994). A Singlish stereotype? *English Today, 10*(01), 63.

Gupta, A. F. (2000). Marketing the voice of authenticity: A comparison of Ming Cher and Rex Shelley. *Language and Literature, 9*(2), 150-169.

Gyory, A. (1998). *Closing the gate: Race, politics, and the Chinese exclusion act*. Chapel Hill: University of North Carolina Press.

Hakala, T. (2010). A great man in clogs: Performing authenticity in Victorian Lancashire. *Victorian Studies, 52*(3), 387-412.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word, 17*(3), 241-292.

Hasan, R. (1987). The grammarian's dream: Lexis as most delicate grammar. In M. A. K. Halliday, R. P. Fawcett, & D. J. Young (Eds.), *New developments in systemic linguistics* (pp. 184-212). London: Frances Pinter.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models* (1st ed.). London: Chapman and Hall.

Hay, S. A. (1994). *African American theatre: An historical and critical analysis*. Cambridge: Cambridge University Press.

Herzog, D. (1998). *Poisoning the minds of the lower orders*. Princeton: Princeton University Press.

Hill, J. (1995). Junk Spanish, covert racism, and the (leaky) boundary between public and private spheres. *Pragmatics, 5*(2), 195-212.

Hilpert, M., & Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing, 24*(4), 385-401.

Hischak, T. S. (2012). *American literature on stage and screen: 525 works and their adaptations*. Jefferson, N.C.: McFarland.

Hobsbawm, E. J. (1987). *The age of empire, 1875-1914*. London: Weidenfeld & Nicolson.

Hodson, J. (2014). *Dialect in film and literature*. Basingstoke: Palgrave Macmillan.

Hodson, J., & Broadhead, A. (2013). Developments in literary dialect representation in British fiction 1800–1836. *Language and Literature, 22*(4), 315-332.

Holm, J. (1984). Variability of the copula in Black English and its creole kin. *American Speech, 59*(4), 291-309.

Holohan, M. (2013). British illustrated editions of Uncle Tom's Cabin: Race, working-class literacy, and transatlantic reprinting in the 1850s. *Resources for American Literary Study, 36*(1), 27-65.

Honeybone, P., & Watson, K. (2013). Salience and the sociolinguistics of Scouse spelling: Exploring the phonology of the contemporary humorous localised dialect literature of liverpool. *English World Wide, 34*(3), 305-340.

Hoppenstand, G. (1992). Yellow devil doctors and opium dens: The yellow peril stereotype in mass media entertainment. In J. G. Nachbar & K. Lausé (Eds.), *Popular culture: An introductory text* (pp. 277-291). Bowling Green: Bowling Green State University Popular Press.

Hutton, C. (2000). Race and language: Ties of 'blood and speech', fictive identity and empire in the writings of Henry Maine and Edward Freeman. *Interventions, 2*(1), 53-72.

Huzzey, R. (2012). *Freedom burning: Anti-slavery and empire in Victorian Britain*. Ithaca: Cornell University Press.

Ives, S. (1950). A theory of literary dialect. *Tulane Studies in English, 2*, 137-182.
Ives, S. (1955). Dialect differentiation in the stories of Joel Chandler Harris. *American Literature, 27*(1), 88-96.
Jaffe, A. (2000). Introduction: Non-standard orthography and non-standard speech. *Journal of Sociolinguistics, 4*(4), 497-513.
Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87-106.
Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Ubana: University of Illinois Press.
Johnstone, B. (2009). Pittsburghese shirts: Commodification and the enregisterment of an urban dialect. *American Speech, 84*(2), 157-175.
Johnstone, B. (2011). Dialect enregisterment in performance. *Journal of Sociolinguistics, 15*(5), 657-679.
Jones, G. R. (1999). *Strange talk: The politics of dialect literature in Gilded Age America*. Berkeley, CA: University of California Press.
Jones, I. E., Berry, D. R., Gill, T. M., Gross, K. N., & Sumler-Edmond, J. (2011). Association of black women historians: Open letter to fans of "The Help". Retrieved from http://newamericamedia.org/2011/08/association-of-black-women-historians-open-letter-to-fans-of-the-help.php
Jortner, M. L. (2009). Throwing insults across the ocean: Charles Matthews and the staging of "the American" in 1824. In K. J. Wetmore (Ed.), *Portrayals of Americans on the world stage: Critical essays* (pp. 26-49). Jefferson, N.C.: McFarland.
Juilland, A. G., Brodin, D. R., & Davidovitch, C. (1970). *Frequency dictionary of french words*. The Hague: Mouton.
Juola, P. (2013). Using the Google n-gram corpus to measure cultural complexity. *Literary and Linguistic Computing, 28*(4), 668-675.
Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Hoboken, N.J.: Wiley.
Kennedy, D. (2002). *Britain and empire, 1880-1945*. Harlow: Longman.
Kennedy, R. (2014). *The children's war: Britain, 1914-1918*. Basingstoke: Palgrave Macmillan.
Kent, E. (2014). *Corporate character: Representing imperial power in British India, 1786-1901*. Toronto: University of Toronto Press.
Kersten, H. (2000). The creative potential of dialect writing in later-nineteenth-century America. *Nineteenth-Century Literature, 55*(1), 92-117.
Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.
Kortmann, B., & Szmrecsanyi, B. (2004). Global synopsis: Morphological and syntactic variation in English. In E. W. Schneider & B. Kortmann (Eds.), *A handbook of varieties of English* (Vol. 2, pp. 1142-1202). New York: Mouton de Gruyter.
Koteyko, N. (2006). Corpus linguistics and the study of meaning in discourse. *Linguistics Journal, 1*(2), 132-157.
Kouwenberg, S., & Singler, J. V. (2008). *The handbook of pidgin and creole studies*. Chichester, West Sussex ; Malden, MA: Wiley-Blackwell Pub.
Kretzschmar, W. A. (2001). Literary dialect analysis with computer assistance: An introduction. *Language and Literature, 10*(2), 99-110.
Kumar, K. (2000). Nation and empire: English and British national identity in comparative perspective. *Theory and Society, 29*(5), 575-608.

Lalla, B., & D'Costa, J. (1990). *Language in exile: Three hundred years of Jamaican creole*. Tuscaloosa: University of Alabama Press.
Lauterbach, E. S., & Davis, W. E. (1973). *The transitional age; British literature, 1880-1920*. Troy: Whitston Pub. Co.
Léglise, I., & Migge, B. (2007). Le "taki-taki", une langue parlée en guyane? Fantasmes et réalités (socio) linguistiques. *Pratiques et représentations linguistiques en Guyane: regards croisés*, 133-157.
Leigh, P. (2011). *A game of confidence: Literary dialect, linguistics, and authenticity.* (PhD), The University of Texas at Austin, Austin, TX.
Lewis, I. (1991). *Sahibs, nabobs and boxwallahs: A dictionary of the words of Anglo-India*. Oxford: Oxford University Press.
Li, J. (2004). Pidgin and code-switching: Linguistic identity and multicultural consciousness in Maxine Hong Kingston's Tripmaster Monkey. *Language and Literature, 13*(3), 269-287.
Lilley, J. D. (2007). Henry Mackenzie's ruined feelings: Romance, race, and the afterlife of sentimental exchange. *New Literary History, 38*(4), 649-666.
Long, A. C. (2014). *Reading Arabia: British orientalism in the age of mass publication, 1880-1930*
Lyne, A. A. (1986). In praise of Juilland's D. *Methodes Quantitatives et Informatiques das l'Etude des Textes, 2*, 589-595.
Magee, G. B., & Thompson, A. S. (2010). *Empire and globalisation: Networks of people, goods and capital in the British world, c.1850-1914*. Cambridge: Cambridge University Press.
Magurran, A. E. (1988). *Ecological diversity and its measurement*. Princeton: Princeton University Press.
Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. New York: Routledge.
Mair, C. (1992). A methodological framework for research on the use of nonstandard language in fiction. *Arbeiten aus Anglistik und Amerikanistik, 17*(1), 103-123.
Mar, L. R. (2010). *Brokering belonging: Chinese in canada's exclusion era, 1885-1945*. New York: Oxford University Press.
Marshall, W. (2011). An eisteddfod for Yorkshire? Professor Moorman and the uses of dialect. *Yorkshire Archaeological Journal, 83*(1), 199-217.
Mautner, G. (2001). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (pp. 122-143). London: Sage.
Mautner, G. (2005). Time to get wired: Using web-based corpora in critical discourse analysis. *Discourse & Society, 16*(6), 809-828.
McAllister, M. E. (2003). *White people do not know how to behave at entertainments designed for ladies & gentlemen of colour: William Brown's African & American theater*. Chapel Hill: University of North Carolina Press.
McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
McMahon, J. G., & Smith, F. J. (1996). Improving statistical language model performance with automatically generated word hierarchies. *Compututational Linguistics, 22*(2), 217-247.
Meek, B. A. (2006). And the Injun goes "how!": Representations of American Indian English in white public space. *Language in Society, 35*(01), 93-128.
Meer, S. (2009). Boucicault's misdirections: Race, transatlantic theatre and social position in the Octoroon. *Atlantic Studies, 6*(1), 81-95.

Melchers, G. (2010). Southern English in writing. In R. Hickey (Ed.), *Varieties of English in writing: The written word as linguistic evidence* (pp. 81-98). Amsterdam: John Benjamins Pub. Co.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.

Miller, M. L. (2009). *Slaves to fashion: Black dandyism and the styling of black diasporic identity*. Durham: Duke University Press.

Minnick, L. C. (2001). Jim's language and the issue of race in Huckleberry Finn. *Language and Literature, 10*(2), 111-128.

Minnick, L. C. (2007). *Dialect and dichotomy: Literary representations of African American speech*. Tuscaloosa, AL: University of Alabama Press.

Montgomery, M. (1999). Eighteenth-century Sierra Leone English: Another exported variety of African American English. *English World-Wide, 20*(1), 1-34.

Moore, C. (2015). Histories of talking about talk: Quethen, quoth, quote. In J. Arendholz, W. Bublitz, & M. Kirner-Ludwig (Eds.), *The pragmatics of quoting now and then* (pp. 255-270). Berlin: De Gruyter Mouton.

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.

Moskal, J. (2000). English national identity in Mariana Starke's "The Sword of Peace": India, abolition, and the rights of women. In C. B. Burroughs (Ed.), *Women in British romantic theatre: Drama, performance, and society, 1790-1840* (pp. 102-131). Cambridge: Cambridge University Press.

Mulderrig, J. (2011). Manufacturing consent: A corpus-based critical discourse analysis of new labour's educational governance. *Educational Philosophy and Theory, 43*(6), 562-578.

Mulderrig, J. (2012). The hegemony of inclusion: A corpus-based critical discourse analysis of deixis in education policy. *Discourse & Society, 23*(6), 701-728.

Munn, C. (2013). *Anglo-China: Chinese people and British rule in Hong Kong, 1841-1880*. London: Routledge.

Murphy, S. (2015). I will proclaim myself what i am: Corpus stylistics and the language of Shakespeare's soliloquies. *Language and Literature, 24*(4), 338-354.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal, 26*(4), 354-359.

Muthiah, K. (2012). Performing Bombay and displaying stances: Stylized Indian English in fiction. *English World Wide, 33*(3), 264-292.

Nechtman, T. W. (2010). *Nabobs: Empire and identity in eighteenth-century Britain*. Cambridge: Cambridge University Press.

Nettels, E. (1988). *Language, race, and social class in Howells's America*. Lexington, Ky.: University Press of Kentucky.

Nickell, J. (1984). Hillbilly talk: Southern Appalachian speech as literary dialect in the writings of Mary Noailles Murfree. *Appalachian Heritage, 12*(3), 37-45.

Nielsen, H. F. (2005). *From dialect to standard: English in England 1154-1776*. Odense: University Press of Southern Denmark.

North, M. (1994). *The dialect of modernism: Race, language, and twentieth-century literature*. New York: Oxford University Press.

Nussbaum, F. (2003). *The limits of the human: Fictions of anomaly, race, and gender in the long eighteenth century*. Cambridge: Cambridge University Press.

Nussbaum, F. (2004). The theatre of empire: Racial counterfeit, racial realism. In K. Wilson (Ed.), *A new imperial history: Culture, identity, and modernity in Britain and the empire, 1660-1840* (pp. 71-90). Cambridge: Cambridge University Press.

O'Quinn, D. (2005). *Staging governance: Theatrical imperialism in London, 1770-1800*. Baltimore: Johns Hopkins University Press.

O'Rourke, J. (2006). The revision of Obi; or, three-finger'd Jack and the Jacobin repudiation of sentimentality. *Nineteenth-Century Contexts, 28*(4), 285-303.

O'Donnell, M. (2009). The UAM corpustool: Software for corpus annotation and exploration. In C. M. B. Callejas, J. F. F. Sánchez, J. R. I. Ibáñez, M. E. G. Sánchez, M. E. C. de los Ríos, S. S. Ramiro, M. S. C. Martínez, N. P. Honeyman, & B. C. Márquez (Eds.), *Applied linguistics now: Understanding language and mind/la lingüística aplicada actual: Comprendiendo el lenguaje y la mente* (pp. 1433-1447). Almería: Universidad de Almería.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Odumosu, T. (2014). In bad taste? Slavery and the African presence in the subversive mockery of royalty. In A. Kremers & E. Reich (Eds.), *Loyal subversion?: Caricatures from the personal union between England and Hanover (1714-1837)* (pp. 122-139). Göttingen: Vandenhoeck & Ruprecht.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G. L., Solymos, P., . . . Wagner, H. (2008). Vegan: Community ecology package. R package version.

Olender, M. (1992). *The languages of paradise: Race, religion, and philology in the nineteenth century*. Cambridge: Harvard University Press.

Oostdijk, N. (1990). The language of dialogue in fiction. *Literary and Linguistic Computing, 5*(3), 235-241.

Page, N. (1973). *Speech in the English novel*. London: Longman.

Paradis, E., Claude, J., & Strimmer, K. (2004). Ape: Analyses of phylogenetics and evolution in R language. *Bioinformatics, 20*(2), 289-290.

Pederson, L. (1967). Mark Twain's Missouri dialects: Marion county phonemics. *American Speech, 42*(4), 261-278.

Pederson, L. (1985). Language in the Uncle Remus tales. *Modern Philology, 82*(3), 292-298.

Poussa, P. (1999). Dickens and sociolinguist: Dialect in David Copperfield. In I. Taavitsainen, G. Melchers, & P. Pahta (Eds.), *Writing in nonstandard English,* (pp. 27-44). Amsterdam: John Benjamins Pub. Co.

Prakash, B. (1994). *Indian themes in English fiction: A socio-literary study*. New Dehli: Mittal Publications.

Pratt, L. (2002). Dialect writing and simultaneity in the American historical romance. *differences: A Journal of Feminist Cultural Studies, 13*(3), 121-142.

Preston, D. R. (1985). The Li'l Abner syndrome: Written representations of speech. *American Speech, 60*(4), 328-336.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ragussis, M. (2010). *Theatrical nation: Jews and other outlandish Englishmen in Georgian Britain*. Philadelphia: University of Pennsylvania Press.

Ribeiro, G. L. (2001). Post-imperialismo: Para una discusión después del post-colonialismo y del multiculturalismo. In D. Mato (Ed.), *Estudios latinoamericanos sobre cultura y transformaciones sociales en tiempos de*

globalización (pp. 161-183). Buenos Aires: Consejo Latinoamericano de Ciencias Sociales.
Richards, J. (1989). *Imperialism and juvenile literature*. Manchester: Manchester University Press.
Richards, J. H. (1997). *Early American drama*. New York: Penguin Books.
Richards, T. (1993). *The imperial archive: Knowledge and the fantasy of empire*. London: Verso.
Rickford, J. R. (1998). The creole origins of African-American vernacular English: Evidence from copula absence. In S. S. Mufwene, J. R. Rickford, G. Bailey, & J. Baugh (Eds.), (pp. 154-200). London: Routledge.
Rickford, J. R., & Rickford, R. J. (2000). *Spoken soul: The story of black English*. New York: Wiley.
Robinson, C. J. (2001). The inventions of the Negro. *Social Identities, 7*(3), 329-361.
Rugemer, E. B. (2004). The southern response to British abolitionism: The maturation of proslavery apologetics. *The Journal of Southern History, 70*(2), 221-248.
Ruzich, C., & Blake, J. (2015). Ain't nothing like the real thing: Dialect, race, and identity in Stockett's novel the help. *The Journal of Popular Culture, 48*(3), 534-547.
Rzepka, C. (Ed.) (2002). *Obi: A romantic circles praxis volume*.
Said, E. W. (1994). *Orientalism*. New York: Vintage Books.
Schneider, E. W. (1993). Africanisms in the grammar of Afro-American English: Weighing the evidence. In S. S. Mufwene & N. Condon (Eds.), *Africanisms in Afro-American language varieties* (pp. 209-221). Athens: University of Georgia Press.
Schneider, E. W., & Wagner, C. (2006). The variability of literary dialect in Jamaican creole: Thelwell's The Harder They Come. *Journal of Pidgin and Creole Languages, 21*(1), 45-96.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656.
Sharma, D. (2011). Return of the native: Hindi in British English. In R. Kothari & R. Snell (Eds.), *Chutnefying English: The phenomenon of Hinglish* (pp. 1-21). New Delhi: Penguin Books.
Sinclair, J. (2004). Lexical grammar. In J. Sinclair & R. Carter (Eds.), *Trust the text: Language, corpus and discourse* (pp. 164-176). London: Routledge.
Smylitopoulos, C. (2011). *A nabob's progress: Rowlandson and Combe's "the grand master", a tale of British imperial excess, 1770--1830.* (Ph.D.), McGill University, Montreal.
Stierstorfer, K. (1996). *John Oxenford (1812-1877) as farceur and critic of comedy*. New York: Peter Lang.
Strand, A. D. (2009). *Language, gender, and citizenship in American literature, 1789-1919*. New York: Routledge.
Sullivan, J. P. (1980). The validity of literary dialect: Evidence from the theatrical portrayal of Hiberno-English forms. *Language in Society, 9*(02), 195-219.
Swaminathan, S. (2009). *Debating the slave trade: Rhetoric of British national identity, 1759-1815*. Farnham: Ashgate.
Tamasi, S. (2001). Huck doesn't sound like himself: Consistency in the literary dialect of Mark Twain. *Language and Literature, 10*(2), 129-144.
Thompson, A. S. (2000). *Imperial Britain: The empire in British politics, c. 1880-1932*. Harlow: Longman.

Tidwell, J. N. (1942). Mark Twain's representation of Negro speech. *American Speech, 17*(3), 174-176.
Traugott, E. C. (1981). The sociostylistics of minority dialect in literary prose *Proceedings of the seventh annual meeting of the Berkeley Linguistics Society* (pp. 308-316).
Troike, R. C. (2010). Assessing the authenticity of Joel Chandler Harris's use of gullah. *American Speech, 85*(3), 287-314.
Trumpener, K. (1997). *Bardic nationalism: The romantic novel and the British empire*. Princeton: Princeton University Press.
Underwood, G. N. (1970). Linguistic realism in "Roderick Random". *The Journal of English and Germanic Philology, 69*(1), 32-40.
Utsumi, A. (2005). The role of feature emergence in metaphor appreciation. *Metaphor and Symbol, 20*(3), 151-172.
Utsumi, A. (2007). Interpretive diversity explains metaphor–simile distinction. *Metaphor and Symbol, 22*(4), 291-312.
Van Dyke, P. A. (2005). *The Canton trade: Life and enterprise on the China coast, 1700-1845*. Hong Kong: Hong Kong University Press.
Waegner, C. C. (2014). Blackface minstrelsy and ethnic identity as globalized market commodities. In G. Pultar (Ed.), *Imagined identities: Identity formation in the age of globalization* (pp. 124-138). Syracuse: Syracuse University Press.
Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics, 27*(4), 299-309.
Wales, K. (2006). *Northern English: A social and cultural history*. Cambridge: Cambridge University Press.
Ward, S. (2001). *British culture and the end of empire*. Manchester: Manchester University Press.
Waters, H. (2009). Jacks and diamonds — some aspects of race on the London Victorian stage. *Race & Class, 50*(3), 77-89.
Wheeler, R. (2000). *The complexion of race: Categories of difference in eighteenth-century British culture*. Philadelphia: University of Pennsylvania Press.
Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: Springer.
Widdowson, H. G. (1995). Discourse analysis: A critical view. *Language and Literature, 4*(3), 157-172.
Widdowson, H. G. (2004). *Text, context, pretext critical issues in discourse analysis*. Malden: Blackwell.
Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management, 24*(5), 577-597.
Witchard, A. V. (2009). *Thomas Burke's dark chinoiserie: Limehouse nights and the queer spell of chinatown*. Farnham: Ashgate.
Wodak, R., de Cillia, R., Reisigl, M., & Liebhart, K. (2009). *The discursive construction of national identity* (2nd ed.). Edinburgh: Edinburgh University Press.
Wolfram, W. (2000). Issues in reconstructing earlier African-American English. *World Englishes, 19*(1), 39-58.
Wong, J. Y. (1998). *Deadly dreams: Opium, imperialism, and the Arrow War (1856-1860) in China*. Cambridge: Cambridge University Press.

Yang, C.-M. (2011). *Performing China: Virtue, commerce, and orientalism in eighteenth-century England, 1660-1760*. Baltimore: Johns Hopkins University Press.

Young, A. R. (2012). John Dicks's illustrated edition of "Shakspere for the millions". *The Papers of the Bibliographical Society of America, 106*(3), 285-310.

Zanger, J. (1966). Literary dialect and social change. *Midcontinent American Studies Journal, 7*(2), 40-48.

Appendix A

Corpus Composition

Source Work			Words of Dialogue				Words in Work
Year	Author	Title	Af. d.	Chi.	Ind.	Total	
1768	Bickerstaff, Isaac	The Padlock: A Comic Opera	838	0	0	838	7336
1772	Garrick, David	The Irish Widow	17	0	0	17	11383
1774	Kelly, Hugh	The Romance of an Hour	0	0	1039	1039	10169
1777	Mackenzie, Henry	Julia De Roubigné, a Tale	95	0	0	95	48861
1778	Foote, Samuel	The Cozeners	33	0	0	33	17180
1779	Dibdin, Charles	The Mirror	66	0	0	66	6730
1779	Pilon, Frederick	The Liverpool Prize	0	0	44	44	10993
1783	O'Keeffe, John	The Dead Alive	31	0	0	31	8093
1787	Colman, George	Inkle and Yarico	287	0	0	287	16245
1788	Cobb, James	Love in the East	0	0	597	597	17224
1788	Starke, Mariana	The Sword of Peace	554	0	144	698	16661
1789	O'Keeffe, John	The Highland Reel	181	0	0	181	18417
1790	Pratt, Samuel	The New Cosmetic	140	0	0	140	13187
1792	Bage, Robert	Man as He Is	1007	0	0	1007	36893
1793	Hoare, Prince	The Prize or, 2, 5, 3, 8	200	0	0	200	7806
1793	Macready, William	The Irishman in London	340	0	0	340	8119
1793	Trusler, John	Life	131	0	0	131	19992
1799	Holman, J. G	The Votary of Wealth	0	0	556	556	21856
1805	Inchbald, Elizabeth	To Marry or Not to Marry	414	0	0	414	16580
1808	Colman, George	The Africans	264	0	0	264	17222
1808	Lewis, Matthew	Romantic Tales	0	0	47	47	18845
1809	Dallas, Robert Charles	Not at Home	232	0	0	232	9132
1816	Hofland, Barbara	Matilda, or, the Barbadoes Girl	470	0	0	470	36421
1817	Pocock, Isaac	Robinson Crusoe	54	0	0	54	11262
1820	Dimond, William	The Lady and the Devil	293	0	0	293	9728
1824	Peake, Richard	Americans Abroad	396	0	0	396	12055
1826	Moncrieff, William	Tom and Jerry; or Life in London	414	0	0	414	22382
1828	Anonymous	Marly	1007	0	0	1007	127207
1828	Hockley, William	The English in India	0	0	640	640	66911
1828	Somerset, Charles A.	The Sea!	152	0	0	152	11869
1829	Fitzball, Edward	The Flying Dutchman	404	0	0	404	14630
1830	Murray, William	Obi, or, Three-Fingered Jack	380	0	0	380	6096
1832	Shipp, John	The K'haunie Kineh-Walla	0	0	24	24	110053
1833	Almar, George	The Knights of St. John	891	0	0	891	12922
1833	Neale, W. Johnson	The Port Admiral	0	0	235	235	61899
1834	Scott, Michael	Tom Cringle's Log	1065	0	0	1065	236421
1836	Baillie, Joanna	The Alienated Manor	258	0	0	258	21153
1836	Howard, Edward	Rattlin, the Reefer	230	0	0	230	148955
1836	Marryat, Frederick	Mr. Midshipman Easy	1046	0	0	1046	140960
1837	Milner, Charles	Don Juan	423	0	0	423	7319

Source Work			Words of Dialogue				Words in Work
Year	Author	Title	Af. d.	Chi.	Ind.	Total	
1837	Oxenford, John	No Followers	810	0	0	810	6647
1839	Ainsworth, William H.	Jack Sheppard: A Romance	46	0	0	46	153863
1839	Barker, Matthew	Hamilton King	1032	0	0	1032	59216
1840	Hook, Theodore	Precepts and Practice	282	0	0	282	171808
1843	Bainbridge, Maria	Rose of Woodlee	0	0	45	45	48312
1843	Herbert, Henry	My Shooting Box	60	0	0	60	52485
1844	Kaye, John William	Peregrine Pultuney	0	0	568	568	80169
1845	Morton, John	The Mother and Child Are Doing Well	161	0	0	161	7116
1850	Cupples, George	The Green Hand: A "Short" Yarn	32	0	191	223	198507
1855	Rafter, Michael	Percy Blake	0	0	784	784	72325
1858	Addison, Henry	Traits and Stories of Anglo-Indian Life	0	20	728	748	72387
1858	Pardoe, Julia S. H	The Poor Relation	0	0	282	282	48728
1859	Boucicault, Dion	The Octoroon, or, Life in Louisiana	949	0	0	949	14666
1859	Leakey, Caroline	The Broad Arrow	0	88	0	88	131503
1860	Taylor, Tom	Up at the Hills	0	0	704	704	19743
1861	Ballantyne, Robert M.	The Golden Dream	0	31	0	31	99050
1863	Jacob, Eustace W.	Something New	0	25	0	25	15189
1863	Reade, Charles	Hard Cash	664	0	0	664	264749
1867	Harrison, Archibald S.	Chota Sabib Charlie	0	0	1048	1048	13895
1867	Payn, James	Carlyon's Year	0	0	221	221	84782
1868	Grant, James	First Love & Last Love	0	0	126	126	58765
1874	Barker, Lucy D. Sale	With a Stout Heart	798	0	349	1147	70917
1874	Marryat, Florence	Sybil's Friend & How She Found Him	0	0	947	947	66071
1876	Ballantyne, Robert M.	Under the Waves	0	180	0	180	93037
1876	Besant, Walter	The Case of Mr. Lucraft	471	0	0	471	21188
1876	Besant, Walter	The Golden Butterfly	0	51	0	51	176673
1876	Kingston, William H. G.	Twice Lost a Story of Shipwreck	172	0	0	172	98036
1877	Jenkins, John Edward	Lutchmee and Dilloo	500	0	207	707	69739
1878	K.	The Indian Famine	0	0	163	163	17375
1879	Paull, Susannah Mary	Levelsie Manor	0	0	195	195	17928
1882	Banks, Isabella	Through the Night	394	0	0	394	126198
1882	Phillips, Edith Caroline	Peeps into China	0	60	0	60	36255
1883	Stables, William	Wild Adventures Round the Pole	1028	0	0	1028	102273
1887	Russell, William Clark	The Frozen Pirate	221	0	0	221	109080
1888	Ballantyne, Robert M.	The Middy and the Moors	905	0	0	905	59275
1889	Henty, George Alfred	Tales of Daring and Danger	0	151	48	199	32651
1889	Hutcheson, John C.	The Black Man's Ghost	963	0	0	963	68686
1890	Doyle, Arthur Conan	The Captain of the Polestar	205	0	0	205	84959
1890	Hutcheson, John C.	Afloat at Last	0	336	0	336	75819
1891	Russell, William Clark	My Danish Sweetheart	0	0	172	172	52111
1893	Fenn, George Manville	Blue Jackets	0	995	0	995	119425
1897	Giles, Elise	China Coast Tales	0	112	0	112	69446
1897	Levett-Yeats, Sidney	A Galahad of the Creeks	0	19	471	490	55421
1897	Meade, Elizabeth	Under the Dragon Throne	0	212	0	212	80340
1898	Collingwood, Harry	A Pirate of the Caribbees	960	0	0	960	86071
1898	Frith, Henry	In the Yellow Sea	0	71	0	71	59664

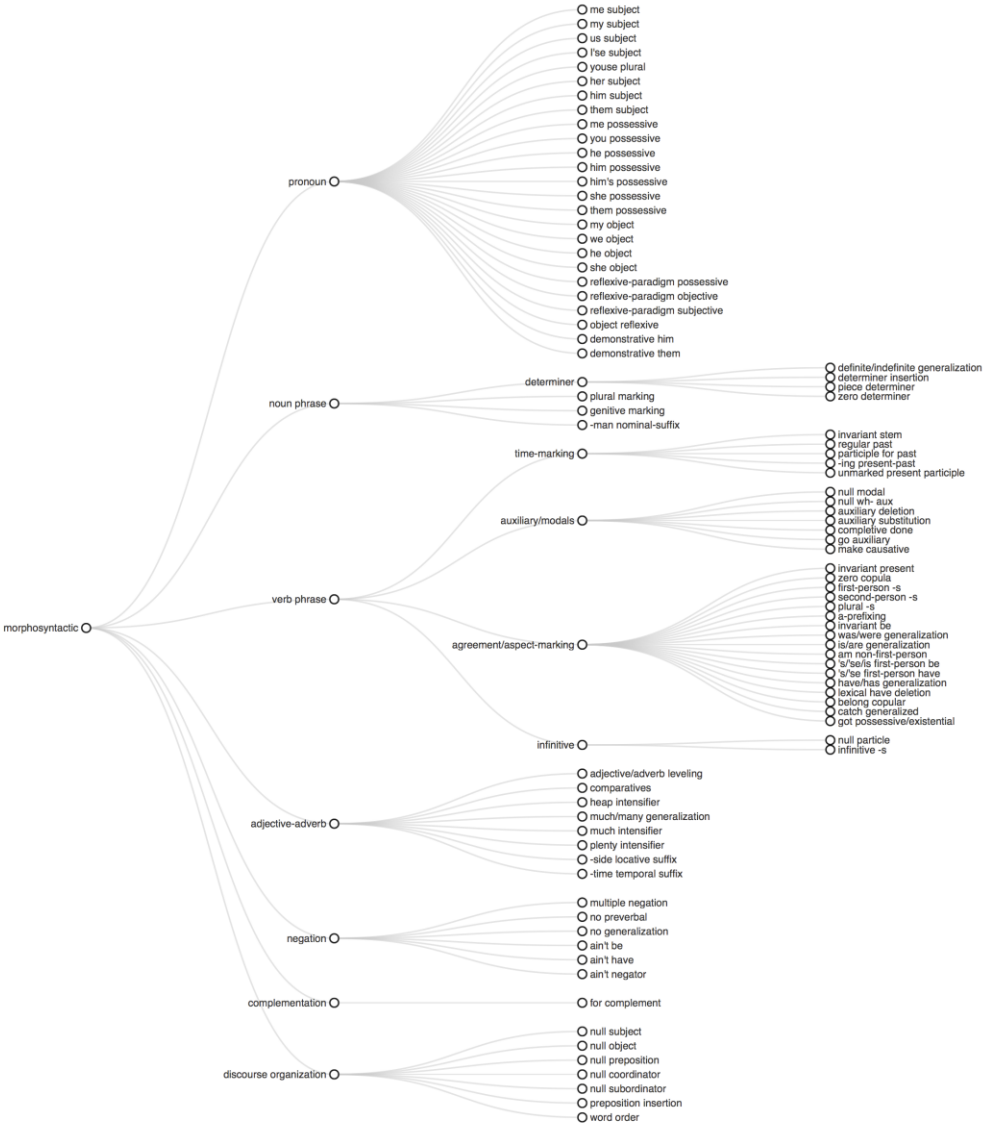
Source Work			Words of Dialogue				Words in Work
Year	Author	Title	Af. d.	Chi.	Ind.	Total	
1898	Hume, Fergus	For the Defense	784	0	0	784	56847
1898	Shiel, Matthew Phipps	The Yellow Danger	0	37	0	37	121490
1899	Bayly, A. Eric	The House of Strange Secrets	0	11	0	11	52917
1899	Hyne, Charles	Further Adventures of Captain Kettle	17	0	0	17	92333
1899	Jones, Henry Arthur	Carnac Sahib	0	0	713	713	25231
1900	Donovan, Dick	The Adventures of Tyler Tatlock	5	18	0	23	99524
1900	Doyle, Charles William	The Shadow of Quong Lung	0	374	0	374	34465
1900	Hemyng, Bracebridge	Boy Tinker among the Turks	516	0	0	516	84112
1900	Jerome, Jerome K	Three Men on the Bummel	132	0	0	132	67820
1901	Kipling, Rudyard	Kim	0	0	1020	1020	107494
1902	Croker, Bithia M.	The Cat's Paw	0	0	1165	1165	99750
1902	Little, Alicia Helen N.	Out in China!	0	153	0	153	38416
1904	Nesbit, Edith	New Treasure Seekers	0	101	0	101	68243
1905	Penny, Fanny Emily	Dilys; an Indian Romance	0	0	108	108	81702
1906	Moore, Henry Charles	Afloat on the Dogger Bank	0	142	0	142	47042
1908	Harraden, Beatrice	Interplay	0	166	0	166	128338
1912	Brereton, Frederick S.	The Hero of Panama	774	209	0	983	104438
1912	Brereton, Frederick S.	Under the Chinese Dragon	0	700	0	700	94916
1912	Rae, Milne	A Bottle in the Smoke	0	0	1013	1013	104214
1912	Strang, Herbert	The Flying Boat	0	1109	0	1109	56800
1912	Vachell, Horace	Bunch Grass	0	69	0	69	85134
1913	Rohmer, Sax	The Mystery of Dr. Fu-Manchu	0	26	0	26	74216
1915	Collingwood, Harry	A Chinese Command	0	391	0	391	104111
1915	Conrad, Joseph	Victory: An Island Tale	0	70	0	70	115488
1915	Foskett, Samuel	The Temple in the Tope	0	0	150	150	101288
1915	Oppenheim, Edward	The Black Box	0	35	0	35	94599
1915	Tracy, Louis	Number Seventeen	0	10	0	10	70245
1916	Burke, Thomas	Limehouse Nights	0	180	0	180	53433
1916	Dorling, Henry Taprell	Stand By! Naval Sketches and Stories	0	44	0	44	27414
1916	Wallace, Edgar	The Tomb of Ts'in	0	28	0	28	52582
1917	Le Queux, William	The Secrets of Potsdam	0	21	0	21	56753
1919	Dell, Ethel May	The Lamp in the Desert	0	0	690	690	121533
1919	Rohmer, Sax	Dope	0	500	0	500	88916
1920	Wren, Percival	Cupid in Africa	973	0	69	1042	88822
1922	Maugham, W. S.	East of Suez	0	1042	0	1042	32354
1922	Middleton, Lydia	The Happy Adventurers	0	29	0	29	64602
1922	Westerman, Percy	The Wireless Officer	0	0	107	107	77748
1924	Galsworthy, John	The Forest	374	0	0	374	19556
1926	Mundy, Talbot	The Devil's Guard	0	0	1029	1029	59835
1929	Burroughs, Edgar Rice	The Monster Men	0	155	0	155	57968
		Totals	**26541**	**7971**	**16639**	**51151**	**7952399**

Appendix B

Coding Taxonomy

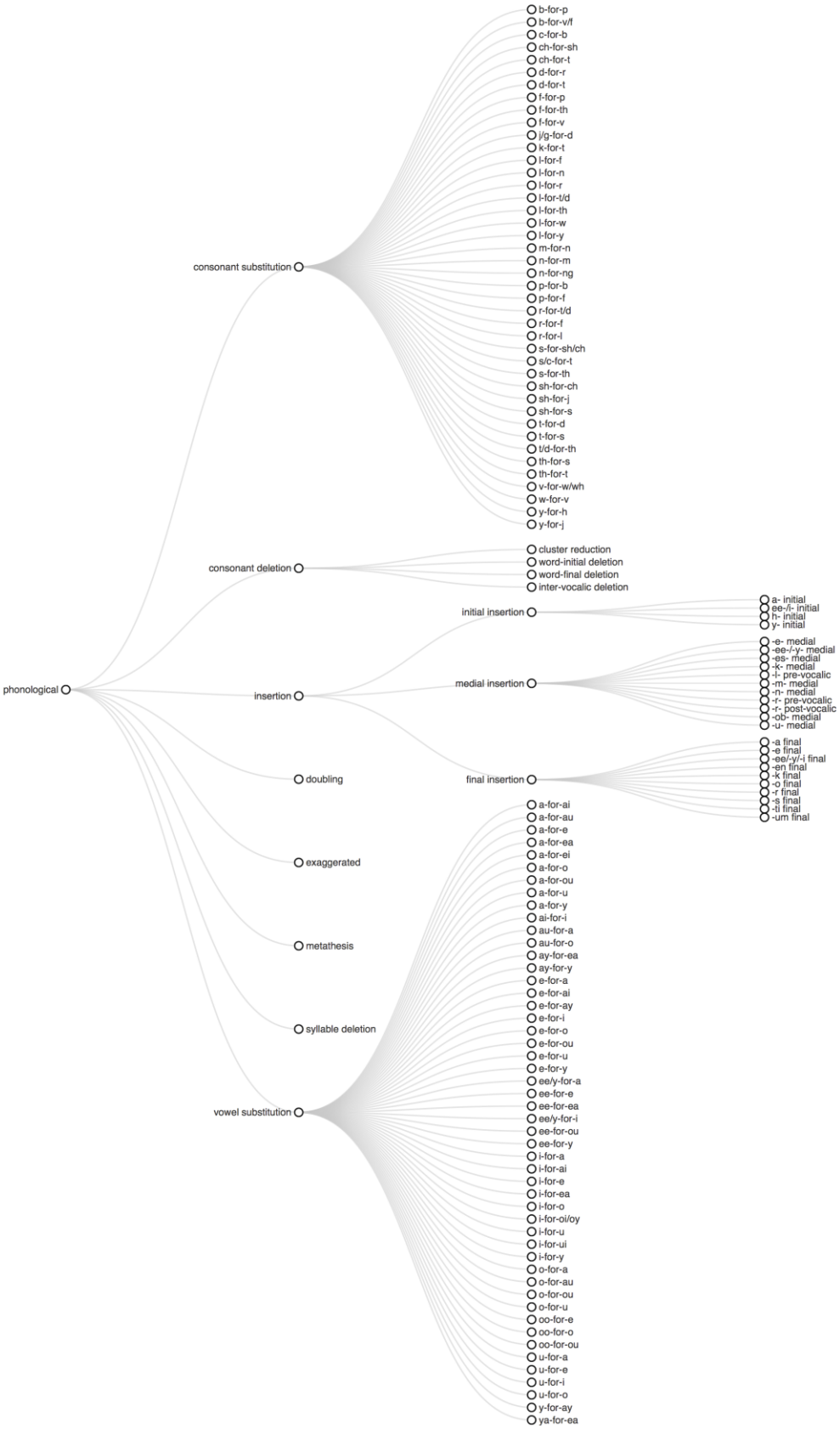
Lexical:

Morphosyntactic:



Orthographic:

Phonological:



Appendix C

Coding Category Descriptions

Lexical Feature Categories	Description	Example
general	This subcategory is for general vocabulary. It is one of the trickier ones, as it is sometimes difficult to determine whether a word or phrase is socially marked, particularly from a historical distance. As with all features, potential items are checked against the dialogue of other speakers within a work. In addition, words or phrases are examined in contemporaneous discourse to try to determine their historical valences.	I can see de yellow bear's moder and two **piccaninnies** on de ice.
address	Forms of address include the appellations or honorifics used to identify a speaker's interlocutor. Following the general principles of coding for this project, these must be forms that distinguish one group of speakers from another. Thus, words like *massa* and *sahib* would most likely be coded, but not *sir*. Also note that this subcategory treats almost all respellings realized in address forms like *massa* as lexicalized. Although the respellings may have a phonological origin, they become so widespread that they can be understood as distinct words with their own, distinct significations. This treatment is adopted here in order to prevent over-assigning codes.	Yes, **missie**, I make plenty barl-dresses for judge's lady, collector's lady, captain's lady.
self-address	Self-address identifies the words or phrases characters use to refer to themselves. This can include their own names, as well as signifiers like *poor negro*.	Me follow her all the world over – Missa be every ting to **poor Cubba**.
inserts	This subcategory covers inserts including interjections like *golly* and *god almighty*, as well as other kinds of particles and connectives.	**Cluck**! dat massa can do if massa like.
wh- word	Variations in *wh-* words usually include the addition of the preposition *for* as in *what for* for *why*. Note that his change is considered a phrasal unit and, thus, would not call for an additional code for preposition insertion.	**What for** he make such fuss?
class shifting	Class shifting identifies a word being used as a different part-of-speech: verbs as nouns, nouns as adjectives, adjectives as verbs, etc.	Oh! it **joy** my heart to hear.
neologism	Neologisms are invented words or malapropisms. Word invention is often accomplished though the unusual joining of free and bound morphemes as in *desufficient* or *responsify*.	Missa Bella **responsify**, 'No matter 'bout de jelly, it keep de ice warm.'
lexical substitution	This subcategory describes the substitution of word or phrase with a closely related word or phrase. The substitution realizes a marked semantic change or collocational pattern. Note that this code can be applied to lexical classes like nouns or verbs, as well as functional classes like prepositions.	I **giving** for seven rupees less. (In this case, *giving* is substituted for a verb like *selling*.)
reduplication	Reduplication codes the repetition of words or phrases. Note that it does not include instances where such repetition has been lexicalized like *chop chop*.	As for me, missee, poor black man, me niber tink not at all; it enough for him to **workee, workee**, when cross old massa make him;
code-mixing	This subcategory is used for the inclusion non-English words and phrases in English language dialogue.	Judge sahib **burra burra buhadoorkea**! – ver' great man!

Morphosyntactic Feature Categories		Description	Example
pronoun	*me* subject *my* subject *us* subject *I'se* subject *youse* plural *her* subject *him* subject *them* subject *me* possessive *you* possessive *he* possessive *him* possessive *him's* possessive *them* possessive *my* object *we* object *he* object *she* object reflexive paradigm possessive reflexive paradigm objective reflexive paradigm subjective object reflexive demonstrative *him* demonstrative *them*	Most of the features in this subcategory are related to personal pronouns and the relationship between their cases their grammatical functions. In these instances, their names are fairly self-explanatory. Less clear may be the reflexive features, which describe paradigmatic variations: the possessive paradigm would include *hisself,* the object paradigm *meself,* and so on. Object reflexives include object pronouns without *-self* being used reflexively. The two demonstrative features indicate object pronouns being as determiners as in *them things*. Note that most of the features in this category are relatively unambiguous and easily identifiable. One exception is *I'se* as a pronominal subject. There are similar codes for agreement in the verb phrase subcategory. Those describe instances where *'se* is a contracted form of *is* or *has.* Here, the entire form is functioning as a noun phrase not as noun + verb. This feature has been identified, for example, by Schneider (1989: 180).	**demonstrative *them*:** We'll jest take the floor togeder and show **dem** female gals what de poetry of motion is like. ***him* subject:** **Him** very fine man.

Morphosyntactic Feature Categories			Description	Example
noun phrase	plural marking genitive marking *-man* nominal suffix		This subcategory includes inflectional variations in nouns, as well as variations in constituents of noun phrases like determiners. Plural marking most often indicates zero inflections, but also includes regularized plurals (like *gentlemans*). Genitive marking indicates zero inflections exclusively. The determiner code is used for both the zero determiner and for the interchange of definite (*the*) and indefinite articles (*a, an*). The *-man* nominal suffix occurs in nouns that describe human actors (*soldier*) with the affix applied word-finally (*soldierman*).	**plural marking:**When I was love at Tanjapour, me was ready to do thousand mad **action** for Balsora – **zero determiner**:Take care not fire **[Ø]** pistol.
	determiner	definite/indefinite generalization *piece* determiner zero determiner determiner insertion		

Morphosyntactic Feature Categories			Description	Example
verb phrase	tense marking	invariant stem regular past participle past *-ing* marking present-past unmarked present-participle	This subcategory codes variation in verb morphology related to time marking. Invariant stem features include regular verbs with no *-ed* inflection to mark the past tense, as well as irregular verbs for which the present and past tense forms are the same (*catch* used for the past tense). Regular past describes an irregular verb that is regularized in its past tense (*catched*). Participle past is for a past participle used in the simple past tense (*seen*), and *-ing* present-past indicates a present participle being used in the simple present or past tense.	**invariant stem:** I **see** Seymour sahib's dubashee last night **participle past:** I **seen** you out of one corner of my eye admirationing my rings
	auxiliary/modals	null modal null *wh*- aux auxiliary deletion auxiliary substitution completive *done* *go* auxiliary *make* causative	This subcategory includes variations in auxiliary and modal verbs. Note that null modal verbs are sometimes difficult to determine and often require contextual information from other speakers. Out of context, the example could express the simple present tense. However, the context makes clear that what is being discussed is a possible future. Also note that the precise modal verb being elided is often impossible to determine. In the example, it could be any number of volition/prediction modals like *will, shall,* or *would.*	**null modal:** Trudge. No? why what shall I do, if I get in their paws? Wows. I **[Ø]** fight for you! Trudge. Will you? **null *wh*- auxiliary:** What **[Ø]** you want? ***go* auxiliary:** Bengallee, he get sick too – bad place, all men **go** die. ***make* causative:** He say one little word; mistress too much sorry because Resident sahib **make** quarrel with his master.

Morphosyntactic Feature Categories			Description	Example
verb phrase	agreement/aspect-marking	invariant present zero copula first-person *-s* second-person *-s* plural *-s* *a-* prefixing invariant *be* *was/were* generalization *is/are* generalization *am* non-first-person *-s/'se/is* first-person *be* *-s/'se* first-person *have* *have/has* generalization lexical *have* deletion *got* possessive/existential	This subcategory includes variation in verb forms as they relate to person and number, as well as variation in aspectual marking. Invariant present indicates verbs that do not change their inflections according to person or number (*he* + *go*). Many of the remaining features describe verbal inflections in the present tense. Note that the zero have in this subcategory is lexical and is distinct from zero have when it is an auxiliary as part of a verb phrase. Also note that the zero copula here includes the absence of be in S + V + SP constructions (*he* + Ø + *old*), as well as in progressives (*sh*e + Ø + *running*) – a definition that follows Labov (1969).	**zero copula:** You **[Ø]** bad man, massa – **a- prefixing:** They're at it again, **a-hammerin'** away **invariant *be*:** Ah, no, you **be** too good; ***was/were* generalization:** True, me tink you nebber **was** cut out as a gentleman in your life. ***got* possessive/existential:** Plenty eestuff **got.**
	infinitive	null particle infinitive *-s*	This subcategory covers two features related to the infinitive phrase. The first is the absence of the infinitive particle *to.* The second is the addition of the *-s* inflection to the infinitive.	**null particle:** – but me no want you **[Ø]** love me –

Morphosyntactic Feature Categories		Description	Example
adjective/adverb	adjective/adverb leveling comparatives *much*/*many* generalization *much* intensifier *plenty* intensifier *-side* locative suffix *-time* temporal suffix	The features in this subcategory are related to modification. Adjective/adverb leveling identifies the collapsing of distinctions between adjectives and adverbs either through inflection (the absence the *-ly* affix in adverbs) or use (good as an adverb). Comparatives code variations in such constructions primarily through the addition of more. Much/many generalization identifies the application of either quantifier to both count and non-count nouns. As to the intensifier features, note that plenty as an intensifier is distinct from plenty as a quantifier when of is elided. Thus, plenty sorry and plenty matches would be coded differently.	**comparatives:** **More better** go nother load. ***much/many* generalization:** Too **muchee** men all wait ***plenty* intensifier:** Pilates, sah; **plenty** bad fellas. ***-side* locative suffix:** Supposey you catchee tlouble, what my tellum boss **Shanghai side**? ***-time* temporal suffix:** Went out to ride, **gun-fire time**, with Seymour sahib.
negation	multiple negation *no* preverbal *no* generalization *ain't* as *be* *ain't* as *have* *ain't* as negator	This subcategory identifies variations in negation. Note that preverbal *no* (as in the example) is concomitant with the elision of the auxiliary do. Because these are mutually dependent features, it is only coded once as preverbal *no*.	***no* preverbal:** mistress **no** like me too much ***no* generalization:** Me **no** afraid now, massa. ***ain't* as have:** Yo **ain't** no right ter order us away.

Morphosyntactic Feature Categories		Description	Example
complementation	*for* complement	This identifies infinitival purpose clauses that are headed by *for to*.	soon as ebber you leave me I begin **for to watch de ham**
discourse organization	null subject null object null preposition null coordinator null subordinator preposition insertion word order	Word order includes the elision of phrasal and lexical grammatical constituents that are not covered by other subcategories, as well as alterations in the ordering of grammatical constituents. Note that null subjects are particularly difficult to identify. Ellipsis is common in the dialogue of all speakers. Thus, it can be challenging to locate cases where elision does not follow normative patterns. As it is with other codes, the practice here is to be conservative. Not all null subjects are coded.	**null subject:** must do what **[Ø]** tink most proper **null subordinator:** Missy, me naughty, same **[Ø]** you used to be **word order:** I **Inglitch can is-peek**

Orthographic Feature Categories	Description	Example
simple	Simple features are respellings that are phonologically unmotivated, or what is sometimes called "eye dialect." Note that the example contains two different orthographic alterations, one to the vowel and one to the consonant. Despite that, the example would take only one code. The vowel change actually follows necessarily from the consonant change. Left in its standard form, the quality of the vowel would become ambiguous. The vowel changes that accompany consonant changes are similarly important when the consonant changes are phonologically motivated. Consider *lub* from *love*. The vowel change is necessary to disambiguate *lub* from *lob*. In that case, the word would not be coded as orthographic at all, but rather as phonological.	'Tis a'most **enuff**, sah, to make a gem'lam turn pale, sah!
ambiguous	This subcategory identifies respellings where the salience is ambiguous. The orthography may or may not be phonologically motivated. Note that it is most often applied to changes in or deletions of vowels.	Spose, massa, no shoot em black dog, **p'rhaps** he shoot em blue monkey

Phonological Feature Categories	Description	Example
metathesis	Metathesis describes the rearrangement of sounds or syllables. Note that this rearrangement may be indicated by the movement of letters or letter groupings or by the substitution of letters.	Me no savee Massa, you never **ax** me before. Ebery one neger, massa want to be **kirstened** in de buckra fashion, massa. (*christened*)
doubling	This identifies the repetition of a consonant or vowel.	Oh, as to **thatt**, I'm not particular anxious to face those business dens.
syllable deletion	Syllable deletions may occur word-initially, word-medially, or word-finally. Note that they can include single letter deletions (like the deletion of a vowel) or vowel + consonant units, as in the example.	Massa him gib you thirty pounds a month, and you spend it all in **'temperate** courses.
exaggerated	Exaggerated features are amalgamated respellings that do not follow any established pattern and signal particularly extreme or disfluent pronunciations. Rather than delineating each of the individual components, this subcategory assigns a single, collective code in order to indicate their overall effect.	Sahib likee **mazinloree**? *(mother-in-law*)

Phonological Feature Categories		Description	Example
consonant substitution	*b*-for-*p* *b*-for-*v*/*f* *c*-for-*b* *ch*-for-*sh* *ch*-for-*t* *d*-for-*r* *d*-for-*t* *f*-for-*p* *f*-for-*th* *f*-for-*v* *j*/*g*-for-*d* *k*-for-*t* *l*-for-*f* *l*-for-*n* *l*-for-*r* *l*-for-*t*/*d* *l*-for-*th* *l*-for-*w* *m*-for-*n* *n*-for-*m* *n*-for-*ng* *p*-for-*b* *p*-for-*f* *r*-for-*t*/*d*	This subcategory describes phonologically motivated substitutions of consonants. The category names identify the specific substitutions. Note that concomitant respellings (like the doubling of the *r* in the *b*-for-*v* example) are not normally assigned an additional code, unless such respellings signal an additional and distinct phonological feature.	***b*-for-*v*:** I should be **berry** glad if you could visit my gentleman ***n*-for-*ng*:** I stop and do de **cookin',** plenty quick

Phonological Feature Categories		Description	Example
consonant substitution, continued	*r*-for-*f* *r*-for-*l* *s*-for-*sh*/*ch* *s*/*c*-for-*t* *s*-for-*th* *sh*-for-*ch* *sh*-for-*s* *t*-for-*d* *t*-for-*s* *t*/*d*-for-*th* *th*-for-*s* *th*-for-*t* *v*-for-*w*/*wh* *w*-for-*v* *y*-for-*h* *y*-for-*j*	This subcategory describes phonologically motivated substitutions of consonants. The category names identify the specific substitutions. Note that concomitant respellings (like the doubling of the *r* in the *b*-for-*v* example) are not normally assigned an additional code, unless such respellings signal an additional and distinct phonological feature.	***t*/*d*-for-*th*:** **De** **oder** day he haul out **de** weather ear-ring, and touch him hat to a midshipman.
consonant deletion	cluster reduction word-initial deletion word-final deletion inter-vocalic deletion	This subcategory describes the deletion of consonants in various positions. Note that cluster reduction includes the reduction of a consonant group both within and across syllabic boundaries.	**cluster reduction:** What a **dreffu**l, **dreffu**l fright dis poor **chile** have got! **word-initial deletion:** **'ope** not dekhe after sahib cook-maid.

Phonological Feature Categories			Description	Example
insertion	initial insertion	*a-* initial *ee-/i-* initial *h-* initial *y-* initial	This subcategory identifies the insertion of vowels or consonants at the beginning of words.	***ee-/i-* initial:** Judge sahib **i-send** Culley Mistree his chupprass
	medial insertion	*-e-* medial *-ee-/-y-* medial *-es-* medial *-k-* medial *-l-* pre-vocalic *-m-* medial *-n-* medial *-r-* pre-vocalic *-r-* post-vocalic *-ob-* medial *-u-* medial	This subcategory identifies the insertion of vowels, consonants, or vowel-consonant groupings in the middle of words.	***-ob-* medial:** I bery **miserabobble,** indeed. ***-r-* pre-vocalic:** much sartainer not to break dan the **brank** of England. ***-r-* post-vocalic** Those ladies never having yem; they liking **farlse** peese.
	final insertion	*-a* final *-e* final *-ee/-y/-i* final *-en* final *-k* final *-o* final *-r* final *-s* final *-um* final	This subcategory identifies the insertion of vowels, consonants, or vowel-consonant groupings at the end of words.	***-ee/-y/-i* final:** Missa, you **frettee** so

Phonological Feature Categories		Description	Example
vowel substitution	*a*-for-*ai* *a*-for-*au* *a*-for-*e* *a*-for-*ea* *a*-for-*ei* *a*-for-*o* *a*-for-*ou* *a*-for-*u* *a*-for-*y* *ai*-for-*i* *au*-for-*a* *au*-for-*o* *ay*-for-*ea* *ay*-for-*y* *e*-for-*a* *e*-for-*ai* *e*-for-*ay* *e*-for-*i* *e*-for-*o* *e*-for-*ou* *e*-for-*u* *e*-for-*y* *ee*/*y*-for-*a* *ee*-for-*e* *ee*-for-*ea*	Vowel substitution presents a somewhat more complicated coding task than consonant substitution. In the latter, spelling and phonology at least roughly correspond (as in the realizing of interdental fricatives as stops being represented by *t/d*-for-*th* substitution), although that correspondence is of course not absolute. Vowels, however, are highly dependent on the phonemes around them. Moreover, they vary by time and by location. Thus, the phonological salience of a respelling may be very different for one writer than it is for another. For these reasons, vowel substitutions are categorized by orthography as they are for consonant substitutions.	***a*-for-*e*:** By **Jasus,** you really tark fine, Massa Easy ***o*-for-*au*:** **Cotch** by pirits.

Phonological Feature Categories		Description	Example
vowel substitution, continued	*ee*/*y*-for-*i* *ee*-for-*ou* *ee*-for-*y* *i*-for-*a* *i*-for-*ai* *i*-for-*e* *i*-for-*ea* *i*-for-*o* *i*-for-*oi*/*oy* *i*-for-*u* *i*-for-*ui* *i*-for-*y* *o*-for-*a* *o*-for-*au* *o*-for-*ou* *o*-for-*u* *oo*-for-*e* *oo*-for-*o* *oo*-for-*ou* *u*-for-*a* *u*-for-*e* *u*-for-*i* *u*-for-*o* *y*-for-*ay* *ya*-for-*ea*	Vowel substitution presents a somewhat more complicated coding task than consonant substitution. In the latter, spelling and phonology at least roughly correspond (as in the realizing of interdental fricatives as stops being represented by *t/d*-for-*th* substitution), although that correspondence is of course not absolute. Vowels, however, are highly dependent on the phonemes around them. Moreover, they vary by time and by location. Thus, the phonological salience of a respelling may be very different for one writer than it is for another. For these reasons, vowel substitutions are categorized by orthography as they are for consonant substitutions.	***u*-for-*a*:** I make ready bring –then **breakfus**.

Appendix D

Feature Tables: All Dialogue

For all features tables (Appendices D-G), the column labels denote the following: *N* is the raw number of occurrences, *% Global* is the percentage a feature or category contributes to all coded features, *% Local* is the percentage a feature or subcategory contributes to the higher order category in which it is embedded, *Freq.* is the normalized frequency of a feature or category (per 1000 words), and *DP* is the deviation of proportions (or dispersion measure).

Feature	N	% Global	% Local	Freq.	DP
FEATURES-TYPE					
TOTAL	**18186**			**355.54**	
lexical	3524	19.38%		68.89	
morphosyntactic	7432	40.87%		145.30	
orthographic	190	1.04%		3.71	
phonological	7040	38.71%		137.63	
LEXICAL-TYPE					
TOTAL	**3524**	**19.38%**		**68.89**	
general vocabulary	494	2.72%	14.02%	9.66	0.54
address	1984	10.91%	56.30%	38.79	0.33
self address	329	1.81%	9.34%	6.43	0.48
inserts	260	1.43%	7.38%	5.08	0.66
wh- word	33	0.18%	0.94%	0.65	0.80
class shifting	38	0.21%	1.08%	0.74	0.71
lexical substitution	126	0.69%	3.58%	2.46	0.58
neologism	21	0.12%	0.60%	0.41	0.92
reduplication	33	0.18%	0.94%	0.65	0.81
code-mixing	206	1.13%	5.85%	4.03	0.74
MORPHOSYNTACTIC-TYPE					
TOTAL	**7432**	**40.87%**		**145.30**	
pronoun	1116	6.14%	15.02%	21.82	
noun phrase	1462	8.04%	19.67%	28.58	
verb phrase	3636	19.99%	48.92%	71.08	
adjective/adverb	172	0.95%	2.31%	3.36	
negation	459	2.52%	6.18%	8.97	
complementation	31	0.17%	0.42%	0.61	
discourse organization	556	3.06%	7.48%	10.87	
PRONOUN-TYPE					
TOTAL	**1116**	**6.14%**		**21.82**	
me subject	469	2.58%	42.03%	9.17	0.64
my subject	43	0.24%	3.85%	0.84	0.95
us subject	1	0.01%	0.09%	0.02	0.98
I'se subject	35	0.19%	3.14%	0.68	0.96
youse plural	2	0.01%	0.18%	0.04	0.98
her subject	8	0.04%	0.72%	0.16	0.96
him subject	212	1.17%	19.00%	4.14	0.65
them subject	18	0.10%	1.61%	0.35	0.84
me possessive	13	0.07%	1.16%	0.25	0.91

Feature	N	% Global	% Local	Freq.	DP
you possessive	30	0.16%	2.69%	0.59	0.82
he possessive	9	0.05%	0.81%	0.18	0.94
him possessive	129	0.71%	11.56%	2.52	0.68
him's possessive	1	0.01%	0.09%	0.02	0.98
she possessive	1	0.01%	0.09%	0.02	1.00
them possessive	5	0.03%	0.45%	0.10	0.95
my object	5	0.03%	0.45%	0.10	0.98
we object	3	0.02%	0.27%	0.06	0.98
he object	7	0.04%	0.63%	0.14	0.92
she object	1	0.01%	0.09%	0.02	0.98
reflexive paradigm possessive	4	0.02%	0.36%	0.08	0.96
reflexive paradigm objective	9	0.05%	0.81%	0.18	0.91
reflexive paradigm subjective	1	0.01%	0.09%	0.02	0.98
object reflexive	4	0.02%	0.36%	0.08	0.97
demonstrative *him*	71	0.39%	6.36%	1.39	0.94
demonstrative *them*	35	0.19%	3.14%	0.68	0.84
NOUN PHRASE-TYPE					
TOTAL	**1462**	**8.04%**		**28.58**	
determiner	1147	6.31%	78.45%	22.42	
plural marking	202	1.11%	13.82%	3.95	0.53
genitive marking	49	0.27%	3.35%	0.96	0.77
-man nominal suffix	64	0.35%	4.38%	1.25	0.78
DETERMINER-TYPE					
TOTAL	**1147**	**6.31%**		**22.42**	
definite/indefinite generalization	3	0.02%	0.26%	0.06	0.96
piece determiner	30	0.16%	2.62%	0.59	0.92
zero determiner	1112	6.11%	96.95%	21.74	0.39
determiner insertion	2	0.01%	0.17%	0.04	0.98
VERB PHRASE-TYPE					
TOTAL	**3636**	**19.99%**		**71.08**	
time marking	759	4.17%	20.87%	14.84	
auxiliary/modals	649	3.57%	17.85%	12.69	
agreement/aspect-marking	2059	11.32%	56.63%	40.25	
infinitive	169	0.93%	4.65%	3.30	
TIME-MARKING-TYPE					
TOTAL	**759**	**4.17%**		**14.84**	
invariant stem	607	3.34%	79.97%	11.87	0.47
regular past	13	0.07%	1.71%	0.25	0.89
participle past	12	0.07%	1.58%	0.23	0.88
-ing present/past	85	0.47%	11.20%	1.66	0.93
unmarked present participle	42	0.23%	5.53%	0.82	0.74
AUXILIARY/MODAL-TYPE					
TOTAL	**649**	**3.57%**		**12.69**	
null modal	350	1.92%	53.93%	6.84	0.47
null *wh*-aux	120	0.66%	18.49%	2.35	0.52
auxiliary deletion	119	0.65%	18.34%	2.33	0.64

Feature	N	% Global	% Local	Freq.	DP
aux substitution	4	0.02%	0.62%	0.08	0.96
done completive	5	0.03%	0.77%	0.10	0.97
go auxiliary	19	0.10%	2.93%	0.37	0.89
make causative	32	0.18%	4.93%	0.63	0.85
AGREEMENT/ASPECT-TYPE					
TOTAL	**2059**	**11.32%**		**40.25**	
invariant present	592	3.26%	28.75%	11.57	0.39
zero copula	977	5.37%	47.45%	19.10	0.34
first-person *-s*	43	0.24%	2.09%	0.84	0.85
second-person *-s*	8	0.04%	0.39%	0.16	0.92
plural *-s*	5	0.03%	0.24%	0.10	0.94
a- prefixing	16	0.09%	0.78%	0.31	0.90
invariant *be*	82	0.45%	3.98%	1.60	0.74
was/were generalization	17	0.09%	0.83%	0.33	0.86
is/are generalization	23	0.13%	1.12%	0.45	0.83
am non-first-person	56	0.31%	2.72%	1.09	0.82
's/'se/is first-person-be	49	0.27%	2.38%	0.96	0.87
's/'se first-person-have	9	0.05%	0.44%	0.18	0.90
have/has generalization	48	0.26%	2.33%	0.94	0.69
lexical *have* deletion	13	0.07%	0.63%	0.25	0.89
belong copular	21	0.12%	1.02%	0.41	0.95
catch generalized	29	0.16%	1.41%	0.57	0.94
got possessive/existential	71	0.39%	3.45%	1.39	0.67
INFINITIVE-TYPE					
TOTAL	**169**	**0.93%**		**3.30**	
null particle	168	0.92%	99.41%	3.28	0.61
infinitive *-s*	1	0.01%	0.59%	0.02	1.00
ADJECTIVE/ADVERB-TYPE					
TOTAL	**172**	**0.95%**		**3.36**	
adjective/adverb leveling	22	0.12%	12.79%	0.43	0.83
comparatives	15	0.08%	8.72%	0.29	0.88
much/many generalization	11	0.06%	6.40%	0.22	0.93
heap intensifier	27	0.15%	15.70%	0.53	0.93
much intensifier	35	0.19%	20.35%	0.68	0.88
plenty intensifier	51	0.28%	29.65%	1.00	0.84
-side locative-suffix	3	0.02%	1.74%	0.06	0.98
-time temporal-suffix	8	0.04%	4.65%	0.16	0.95
NEGATION-TYPE					
TOTAL	**459**	**2.52%**		**8.97**	
multiple negation	37	0.20%	8.06%	0.72	0.75
no preverbal	323	1.78%	70.37%	6.31	0.55
no generalization	72	0.40%	15.69%	1.41	0.67
ain't as *be*	25	0.14%	5.45%	0.49	0.80
ain't as *have*	1	0.01%	0.22%	0.02	0.98
ain't as negator	1	0.01%	0.22%	0.02	0.98

Feature	N	% Global	% Local	Freq.	DP
COMPLEMENTATION-TYPE					
TOTAL	**31**	**0.17%**		**0.61**	
for complement	31	0.17%	100.00%	0.61	0.87
DISCOURSE ORGANIZATION-TYPE					
TOTAL	**556**	**3.06%**		**10.87**	
null subject	257	1.41%	46.22%	5.02	0.61
null object	71	0.39%	12.77%	1.39	0.72
null preposition	153	0.84%	27.52%	2.99	0.62
null coordinator	7	0.04%	1.26%	0.14	0.95
null subordinator	7	0.04%	1.26%	0.14	0.96
preposition insertion	5	0.03%	0.90%	0.10	0.94
word order	56	0.31%	10.07%	1.09	0.80
ORTHOGRAPHIC-TYPE					
TOTAL	**190**	**1.04%**		**3.71**	
eye dialect	99	0.54%	52.11%	1.94	0.61
ambiguous	91	0.50%	47.89%	1.78	0.62
PHONOLOGICAL-TYPE					
TOTAL	**7040**	**38.71%**		**137.63**	
consonant substitution	3957	21.76%	56.21%	77.36	
consonant deletion	837	4.60%	11.89%	16.36	
insertion	1045	5.75%	14.84%	20.43	
vowel substitution	731	4.02%	10.38%	14.29	
metathesis	13	0.07%	0.18%	0.25	0.88
syllable deletion	370	2.03%	5.26%	7.23	0.52
doubling	45	0.25%	0.64%	0.88	0.89
exaggerated	42	0.23%	0.60%	0.82	0.85
CONSONANT SUBSTITUTION-TYPE					
TOTAL	**3957**	**21.76%**		**77.36**	
b-for-*p*	2	0.01%	0.05%	0.04	0.99
b-for-*v*/*f*	654	3.60%	16.53%	12.79	0.59
c-for-*b*	1	0.01%	0.03%	0.02	0.98
ch-for-*sh*	1	0.01%	0.03%	0.02	1.00
ch-for-*t*	39	0.21%	0.99%	0.76	0.93
d-for-*r*	1	0.01%	0.03%	0.02	0.98
d-for-*t*	1	0.01%	0.03%	0.02	0.99
f-for-*p*	1	0.01%	0.03%	0.02	0.99
f-for-*th*	43	0.24%	1.09%	0.84	0.84
f-for-*v*	3	0.02%	0.08%	0.06	1.00
j/*g*-for-*d*	6	0.03%	0.15%	0.12	0.96
k-for-*t*	2	0.01%	0.05%	0.04	0.99
l-for-*f*	1	0.01%	0.03%	0.02	0.99
l-for-*n*	1	0.01%	0.03%	0.02	0.98
l-for-*r*	607	3.34%	15.34%	11.87	0.85
l-for-*t*/*d*	3	0.02%	0.08%	0.06	0.99
l-for-*th*	9	0.05%	0.23%	0.18	0.98
l-for-*w*	9	0.05%	0.23%	0.18	1.00
l-for-*y*	10	0.05%	0.25%	0.20	1.00

Feature	N	% Global	% Local	Freq.	DP
m-for-*n*	1	0.01%	0.03%	0.02	0.99
n-for-*m*	1	0.01%	0.03%	0.02	0.98
n-for-*ng*	107	0.59%	2.70%	2.09	0.77
p-for-*b*	1	0.01%	0.03%	0.02	1.00
p-for-*f*	1	0.01%	0.03%	0.02	1.00
r-for-*t*/*d*	27	0.15%	0.68%	0.53	0.91
r-for-*f*	5	0.03%	0.13%	0.10	0.95
r-for-*l*	19	0.10%	0.48%	0.37	0.86
s-for-*sh*/*ch*	19	0.10%	0.48%	0.37	0.89
s/*c*-for-*t*	1	0.01%	0.03%	0.02	1.00
s-for-*th*	4	0.02%	0.10%	0.08	0.96
sh-for-*ch*	1	0.01%	0.03%	0.02	1.00
sh-for-*j*	2	0.01%	0.05%	0.04	1.00
sh-for-*s*	4	0.02%	0.10%	0.08	0.96
t-for-*d*	11	0.06%	0.28%	0.22	0.97
t-for-*s*	3	0.02%	0.08%	0.06	0.98
t/*d*-for-*th*	2227	12.25%	56.28%	43.54	0.45
th-for-*s*	3	0.02%	0.08%	0.06	0.98
th-for-*t*	1	0.01%	0.03%	0.02	0.98
v-for-*w*/*wh*	75	0.41%	1.90%	1.47	0.91
w-for-*v*	31	0.17%	0.78%	0.61	0.90
y-for-*h*	17	0.09%	0.43%	0.33	0.94
y-for-*j*	2	0.01%	0.05%	0.04	0.98
CONSONANT DELETION-TYPE					
TOTAL	**837**	**4.60%**		**16.36**	
cluster reduction	457	2.51%	54.60%	8.93	0.58
word-initial deletion	225	1.24%	26.88%	4.40	0.67
word-final deletion	149	0.82%	17.80%	2.91	0.66
inter-vocalic deletion	6	0.03%	0.72%	0.12	0.94
INSERTION-TYPE					
TOTAL	**1045**	**5.75%**		**20.43**	
initial insertion	22	0.12%	2.11%	0.43	
medial insertion	76	0.42%	7.27%	1.49	
final insertion	947	5.21%	90.62%	18.51	
INITIAL INSERTION-TYPE					
TOTAL	**22**	**0.12%**		**0.43**	
a- initial	2	0.01%	9.09%	0.04	0.99
ee-/*i*- initial	13	0.07%	59.09%	0.25	0.98
h- initial	5	0.03%	22.73%	0.10	0.98
y- initial	2	0.01%	9.09%	0.04	0.97
MEDIAL INSERTION-TYPE					
TOTAL	**76**	**0.42%**		**1.49**	
-*e*- medial	2	0.01%	2.63%	0.04	1.00
-*ee*-/-*y*- medial	7	0.04%	9.21%	0.14	0.95
-*es*- medial	1	0.01%	1.32%	0.02	0.98
-*k*- medial	1	0.01%	1.32%	0.02	0.99
-*l*- pre-vocalic	23	0.13%	30.26%	0.45	0.97

Feature	N	% Global	% Local	Freq.	DP
-*m*- medial	1	0.01%	1.32%	0.02	1.00
-*n*- medial	7	0.04%	9.21%	0.14	0.99
-*r*- pre-vocalic	6	0.03%	7.89%	0.12	0.98
-*r*- post-vocalic	26	0.14%	34.21%	0.51	0.83
-*ob*- medial	1	0.01%	1.32%	0.02	0.98
-*u*- medial	1	0.01%	1.32%	0.02	0.99
FINAL INSERTION-TYPE					
TOTAL	**947**	**5.21%**		**18.51**	
-*a* final	73	0.40%	7.71%	1.43	0.83
-*e* final	12	0.07%	1.27%	0.23	0.94
-*ee*/-*y*/-*i* final	760	4.18%	80.25%	14.86	0.73
-*en* final	1	0.01%	0.11%	0.02	0.99
-*k* final	5	0.03%	0.53%	0.10	0.95
-*o* final	10	0.05%	1.06%	0.20	0.96
-*r* final	71	0.39%	7.50%	1.39	0.83
-*s* final	6	0.03%	0.63%	0.12	0.97
-*ti* final	1	0.01%	0.11%	0.02	0.98
-*um* final	8	0.04%	0.84%	0.16	0.96
VOWEL SUBSTITUTION-TYPE					
TOTAL	**731**	**4.02%**		**14.29**	
a-for-*ai*	8	0.04%	1.09%	0.16	0.96
a-for-*au*	2	0.01%	0.27%	0.04	0.99
a-for-*e*	90	0.49%	12.31%	1.76	0.72
a-for-*ea*	15	0.08%	2.05%	0.29	0.88
a-for-*ei*	1	0.01%	0.14%	0.02	0.98
a-for-*o*	31	0.17%	4.24%	0.61	0.85
a-for-*ou*	5	0.03%	0.68%	0.10	0.98
a-for-*u*	4	0.02%	0.55%	0.08	0.94
a-for-*y*	2	0.01%	0.27%	0.04	0.98
ai-for-*i*	2	0.01%	0.27%	0.04	0.98
au-for-*a*	4	0.02%	0.55%	0.08	0.99
au-for-*o*	1	0.01%	0.14%	0.02	1.00
ay-for-*ea*	5	0.03%	0.68%	0.10	0.98
ay-for-*y*	1	0.01%	0.14%	0.02	0.98
e-for-*a*	21	0.12%	2.87%	0.41	0.83
e-for-*ai*	5	0.03%	0.68%	0.10	0.95
e-for-*ay*	4	0.02%	0.55%	0.08	0.94
e-for-*i*	7	0.04%	0.96%	0.14	0.93
e-for-*o*	45	0.25%	6.16%	0.88	0.88
e-for-*ou*	41	0.23%	5.61%	0.80	0.87
e-for-*u*	10	0.05%	1.37%	0.20	0.92
e-for-*y*	1	0.01%	0.14%	0.02	1.00
ee/*y*-for-*a*	6	0.03%	0.82%	0.12	0.96
ee-for-*e*	8	0.04%	1.09%	0.16	0.95
ee-for-*ea*	5	0.03%	0.68%	0.10	0.97
ee/*y*-for-*i*	25	0.14%	3.42%	0.49	0.83
ee-for-*ou*	2	0.01%	0.27%	0.04	0.98

Feature	N	% Global	% Local	Freq.	DP
ee-for-*y*	56	0.31%	7.66%	1.09	0.89
i-for-*a*	22	0.12%	3.01%	0.43	0.94
i-for-*ai*	2	0.01%	0.27%	0.04	0.98
i-for-*e*	109	0.60%	14.91%	2.13	0.76
i-for-*ea*	1	0.01%	0.14%	0.02	1.00
i-for-*o*	1	0.01%	0.14%	0.02	1.00
i-for-*oi/oy*	6	0.03%	0.82%	0.12	0.93
i-for-*u*	4	0.02%	0.55%	0.08	0.93
i-for-*ui*	1	0.01%	0.14%	0.02	0.98
i-for-*y*	4	0.02%	0.55%	0.08	0.99
o-for-*a*	10	0.05%	1.37%	0.20	0.90
o-for-*au*	12	0.07%	1.64%	0.23	0.93
o-for-*ou*	57	0.31%	7.80%	1.11	0.92
o-for-*u*	9	0.05%	1.23%	0.18	0.92
oo-for-*e*	2	0.01%	0.27%	0.04	0.98
oo-for-*o*	4	0.02%	0.55%	0.08	0.99
oo-for-*ou*	3	0.02%	0.41%	0.06	1.00
u-for-*a*	5	0.03%	0.68%	0.10	0.96
u-for-*e*	47	0.26%	6.43%	0.92	0.78
u-for-*i*	13	0.07%	1.78%	0.25	0.86
u-for-*o*	8	0.04%	1.09%	0.16	0.92
y-for-*ay*	1	0.01%	0.14%	0.02	1.00
ya-for-*ea*	3	0.02%	0.41%	0.06	0.97

Appendix E

Features Tables: African Diasporic Dialogue

For all features tables (Appendices D-G), the column labels denote the following: *N* is the raw number of occurrences, *% Global* is the percentage a feature or category contributes to all coded features, *% Local* is the percentage a feature or subcategory contributes to the higher order category in which it is embedded, *Freq.* is the normalized frequency of a feature or category (per 1000 words), and *DP* is the deviation of proportions (or dispersion measure).

Feature	N	% Global	% Local	Freq.	DP
FEATURES-TYPE					
TOTAL	**10110**			**380.92**	
lexical	1847	18.27%		69.59	
morphosyntactic	3531	34.93%		133.04	
orthographic	133	1.32%		5.01	
phonological	4599	45.49%		173.28	
LEXICAL-TYPE					
TOTAL	**1847**	**18.27%**		**69.59**	
general vocabulary	201	1.99%	10.88%	7.57	0.47
address	1158	11.45%	62.70%	43.63	0.24
self address	130	1.29%	7.04%	4.90	0.47
inserts	215	2.13%	11.64%	8.10	0.52
wh- word	11	0.11%	0.60%	0.41	0.79
class shifting	21	0.21%	1.14%	0.79	0.72
lexical substitution	59	0.58%	3.19%	2.22	0.64
neologism	19	0.19%	1.03%	0.72	0.88
reduplication	15	0.15%	0.81%	0.57	0.81
code-mixing	18	0.18%	0.97%	0.68	0.90
MORPHOSYNTACTIC-TYPE					
TOTAL	**3531**	**34.93%**		**133.04**	
pronoun	778	7.70%	22.03%	29.31	
noun phrase	545	5.39%	15.43%	20.53	
verb phrase	1869	18.49%	52.93%	70.42	
adjective/adverb	32	0.32%	0.91%	1.21	
negation	195	1.93%	5.52%	7.35	
complementation	25	0.25%	0.71%	0.94	
discourse organization	87	0.86%	2.46%	3.28	
PRONOUN-TYPE					
TOTAL	**778**	**7.70%**		**29.31**	
me subject	304	3.01%	39.07%	11.45	0.57
my subject	0	0.00%	0.00%	0.00	NA
us subject	1	0.01%	0.13%	0.04	0.97
I'se subject	35	0.35%	4.50%	1.32	0.92
youse plural	2	0.02%	0.26%	0.08	0.97
her subject	5	0.05%	0.64%	0.19	0.97
him subject	160	1.58%	20.57%	6.03	0.57
them subject	11	0.11%	1.41%	0.41	0.77
me possessive	8	0.08%	1.03%	0.30	0.89

Feature	N	% Global	% Local	Freq.	DP
you possessive	24	0.24%	3.08%	0.90	0.73
he possessive	7	0.07%	0.90%	0.26	0.92
him possessive	94	0.93%	12.08%	3.54	0.61
him's possessive	1	0.01%	0.13%	0.04	0.96
she possessive	0	0.00%	0.00%	0.00	NA
them possessive	4	0.04%	0.51%	0.15	0.95
my object	0	0.00%	0.00%	0.00	NA
we object	2	0.02%	0.26%	0.08	0.96
he object	1	0.01%	0.13%	0.04	0.96
she object	1	0.01%	0.13%	0.04	0.96
reflexive paradigm possessive	4	0.04%	0.51%	0.15	0.93
reflexive paradigm objective	7	0.07%	0.90%	0.26	0.86
reflexive paradigm subjective	1	0.01%	0.13%	0.04	0.96
object reflexive	3	0.03%	0.39%	0.11	0.95
demonstrative *him*	71	0.70%	9.13%	2.68	0.88
demonstrative *them*	32	0.32%	4.11%	1.21	0.72
NOUN PHRASE-TYPE					
TOTAL	**545**	**5.39%**		**20.53**	
determiner	408	4.04%	74.86%	15.37	
plural marking	101	1.00%	18.53%	3.81	0.49
genitive marking	17	0.17%	3.12%	0.64	0.82
-man nominal suffix	19	0.19%	3.49%	0.72	0.84
DETERMINER-TYPE					
TOTAL	**408**	**4.04%**		**15.37**	
definite/indefinite generalization	2	0.02%	0.49%	0.08	0.97
piece determiner	0	0.00%	0.00%	0.00	NA
zero determiner	404	4.00%	99.02%	15.22	0.38
determiner insertion	2	0.02%	0.49%	0.08	0.96
VERB PHRASE-TYPE					
TOTAL	**1869**	**18.49%**		**70.42**	
time marking	404	4.00%	21.62%	15.22	
auxiliary/modals	279	2.76%	14.93%	10.51	
agreement/aspect-marking	1121	11.09%	59.98%	42.24	
infinitive	65	0.64%	3.48%	2.45	
TIME-MARKING-TYPE					
TOTAL	**404**	**4.00%**		**15.22**	
invariant stem	328	3.24%	81.19%	12.36	0.40
regular past	12	0.12%	2.97%	0.45	0.80
participle past	7	0.07%	1.73%	0.26	0.87
-ing present/past	41	0.41%	10.15%	1.54	0.96
unmarked present participle	16	0.16%	3.96%	0.60	0.72
AUXILIARY/MODALS-TYPE					
TOTAL	**279**	**2.76%**		**10.51**	
null modal	165	1.63%	59.14%	6.22	0.43
null *wh*-aux	67	0.66%	24.01%	2.52	0.50
auxiliary deletion	38	0.38%	13.62%	1.43	0.54

Feature	N	% Global	% Local	Freq.	DP
aux substitution	2	0.02%	0.72%	0.08	0.94
done completive	1	0.01%	0.36%	0.04	0.99
go auxiliary	3	0.03%	1.08%	0.11	0.98
make causative	3	0.03%	1.08%	0.11	0.94
AGREEMENT/ASPECT-TYPE					
TOTAL	**1121**	**11.09%**		**42.24**	
invariant present	275	2.72%	24.53%	10.36	0.35
zero copula	519	5.13%	46.30%	19.55	0.30
first-person *-s*	42	0.42%	3.75%	1.58	0.72
second-person *-s*	8	0.08%	0.71%	0.30	0.85
plural *-s*	4	0.04%	0.36%	0.15	0.90
a- prefixing	16	0.16%	1.43%	0.60	0.80
invariant *be*	60	0.59%	5.35%	2.26	0.64
was/were generalization	14	0.14%	1.25%	0.53	0.77
is/are generalization	18	0.18%	1.61%	0.68	0.76
am non-first-person	54	0.53%	4.82%	2.03	0.69
's/'se/is first-person-be	48	0.47%	4.28%	1.81	0.79
's/'se first-person-have	7	0.07%	0.62%	0.26	0.84
have/has generalization	28	0.28%	2.50%	1.05	0.68
lexical *have* deletion	7	0.07%	0.62%	0.26	0.88
belong copular	0	0.00%	0.00%	0.00	NA
catch generalized	0	0.00%	0.00%	0.00	NA
got possessive/existential	21	0.21%	1.87%	0.79	0.70
INFINITIVE-TYPE					
TOTAL	**65**	**0.64%**		**2.45**	
null particle	65	0.64%	100.00%	2.45	0.59
infinitive *-s*	0	0.00%	0.00%	0.00	NA
ADJECTIVE/ADVERB-TYPE					
TOTAL	**32**	**0.32%**		**1.21**	
adjective/adverb leveling	9	0.09%	28.13%	0.34	0.85
comparatives	3	0.03%	9.38%	0.11	0.90
much/many generalization	3	0.03%	9.38%	0.11	0.95
heap intensifier	1	0.01%	3.13%	0.04	0.96
much intensifier	7	0.07%	21.88%	0.26	0.96
plenty intensifier	8	0.08%	25.00%	0.30	0.93
-side locative-suffix	0	0.00%	0.00%	0.00	NA
-time temporal-suffix	1	0.01%	3.13%	0.04	0.97
NEGATION-TYPE					
TOTAL	**195**	**1.93%**		**7.35**	
multiple negation	29	0.29%	14.87%	1.09	0.68
no preverbal	109	1.08%	55.90%	4.11	0.50
no generalization	32	0.32%	16.41%	1.21	0.64
ain't as *be*	23	0.23%	11.79%	0.87	0.65
ain't as *have*	1	0.01%	0.51%	0.04	0.97
ain't as negator	1	0.01%	0.51%	0.04	0.97

Feature	N	% Global	% Local	Freq.	DP
COMPLEMENTATION-TYPE					
TOTAL	**25**	**0.25%**		**0.94**	
for complement	25	0.25%	100.00%	0.94	0.83
DISCOURSE ORGANIZATION-TYPE					
TOTAL	**87**	**0.86%**		**3.28**	
null subject	40	0.40%	45.98%	1.51	0.65
null object	5	0.05%	5.75%	0.19	0.86
null preposition	33	0.33%	37.93%	1.24	0.54
null coordinator	0	0.00%	0.00%	0.00	NA
null subordinator	4	0.04%	4.60%	0.15	0.97
preposition insertion	0	0.00%	0.00%	0.00	NA
word order	5	0.05%	5.75%	0.19	0.92
ORTHOGRAPHIC-TYPE					
TOTAL	**133**	**1.32%**		**5.01**	
eye dialect	81	0.80%	60.90%	3.05	0.48
ambiguous	52	0.51%	39.10%	1.96	0.55
PHONOLOGICAL-TYPE					
TOTAL	**4599**	**45.49%**		**173.28**	
consonant substitution	2808	27.77%	61.06%	105.80	
consonant deletion	693	6.85%	15.07%	26.11	
insertion	284	2.81%	6.18%	10.70	
vowel substitution	531	5.25%	11.55%	20.01	
metathesis	12	0.12%	0.26%	0.45	0.78
syllable deletion	232	2.29%	5.04%	8.74	0.43
doubling	8	0.08%	0.17%	0.30	0.93
exaggerated	31	0.31%	0.67%	1.17	0.79
CONSONANT SUBSTITUTION-TYPE					
TOTAL	**2808**	**27.77%**		**105.80**	
b-for-*p*	2	0.02%	0.07%	0.08	0.98
b-for-*v*/*f*	595	5.89%	21.19%	22.42	0.42
c-for-*b*	1	0.01%	0.04%	0.04	0.96
ch-for-*sh*	0	0.00%	0.00%	0.00	NA
ch-for-*t*	0	0.00%	0.00%	0.00	NA
d-for-*r*	1	0.01%	0.04%	0.04	0.97
d-for-*t*	1	0.01%	0.04%	0.04	0.98
f-for-*p*	1	0.01%	0.04%	0.04	0.98
f-for-*th*	32	0.32%	1.14%	1.21	0.78
f-for-*v*	0	0.00%	0.00%	0.00	NA
j/*g*-for-*d*	6	0.06%	0.21%	0.23	0.93
k-for-*t*	2	0.02%	0.07%	0.08	0.97
l-for-*f*	0	0.00%	0.00%	0.00	NA
l-for-*n*	1	0.01%	0.04%	0.04	0.96
l-for-*r*	3	0.03%	0.11%	0.11	0.93
l-for-*t*/*d*	0	0.00%	0.00%	0.00	NA
l-for-*th*	0	0.00%	0.00%	0.00	NA
l-for-*w*	0	0.00%	0.00%	0.00	NA
l-for-*y*	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
m-for-*n*	1	0.01%	0.04%	0.04	0.99
n-for-*m*	1	0.01%	0.04%	0.04	0.96
n-for-*ng*	96	0.95%	3.42%	3.62	0.67
p-for-*b*	1	0.01%	0.04%	0.04	0.99
p-for-*f*	0	0.00%	0.00%	0.00	NA
r-for-*t*/*d*	26	0.26%	0.93%	0.98	0.83
r-for-*f*	5	0.05%	0.18%	0.19	0.91
r-for-*l*	15	0.15%	0.53%	0.57	0.76
s-for-*sh*/*ch*	3	0.03%	0.11%	0.11	0.91
s/*c*-for-*t*	0	0.00%	0.00%	0.00	NA
s-for-*th*	1	0.01%	0.04%	0.04	0.97
sh-for-*ch*	1	0.01%	0.04%	0.04	1.00
sh-for-*j*	2	0.02%	0.07%	0.08	1.00
sh-for-*s*	3	0.03%	0.11%	0.11	0.92
t-for-*d*	9	0.09%	0.32%	0.34	0.94
t-for-*s*	0	0.00%	0.00%	0.00	NA
t/*d*-for-*th*	1903	18.82%	67.77%	71.70	0.22
th-for-*s*	3	0.03%	0.11%	0.11	0.96
th-for-*t*	1	0.01%	0.04%	0.04	0.97
v-for-*w*/*wh*	61	0.60%	2.17%	2.30	0.86
w-for-*v*	15	0.15%	0.53%	0.57	0.86
y-for-*h*	14	0.14%	0.50%	0.53	0.89
y-for-*j*	2	0.02%	0.07%	0.08	0.96
CONSONANT DELETION-TYPE					
TOTAL	**693**	**6.85%**		**26.11**	
cluster reduction	392	3.88%	56.57%	14.77	0.46
word-initial deletion	185	1.83%	26.70%	6.97	0.61
word-final deletion	112	1.11%	16.16%	4.22	0.57
inter-vocalic deletion	4	0.04%	0.58%	0.15	0.92
INSERTION-TYPE					
TOTAL	**284**	**2.81%**		**10.70**	
initial insertion	5	0.05%	1.76%	0.19	
medial insertion	41	0.41%	14.44%	1.54	
final insertion	238	2.35%	83.80%	8.97	
INITIAL INSERTION-TYPE					
TOTAL	**5**	**0.05%**		**0.19**	
a- initial	0	0.00%	0.00%	0.00	NA
ee-/*i*- initial	0	0.00%	0.00%	0.00	NA
h- initial	4	0.04%	80.00%	0.15	0.96
y- initial	1	0.01%	20.00%	0.04	0.99
MEDIAL INSERTION-TYPE					
TOTAL	**41**	**0.41%**		**1.54**	
-*e*- medial	0	0.00%	0.00%	0.00	NA
-*ee*-/-*y*- medial	3	0.03%	7.32%	0.11	0.95
-*es*- medial	1	0.01%	2.44%	0.04	0.96
-*k*- medial	1	0.01%	2.44%	0.04	0.97
-*l*- pre-vocalic	2	0.02%	4.88%	0.08	1.00

Feature	N	% Global	% Local	Freq.	DP
-m- medial	0	0.00%	0.00%	0.00	NA
-n- medial	7	0.07%	17.07%	0.26	0.98
-r- pre-vocalic	6	0.06%	14.63%	0.23	0.96
-r- post-vocalic	20	0.20%	48.78%	0.75	0.73
-ob- medial	1	0.01%	2.44%	0.04	0.97
-u- medial	0	0.00%	0.00%	0.00	NA
FINAL INSERTION-TYPE					
TOTAL	**238**	**2.35%**		**8.97**	
-a final	39	0.39%	16.39%	1.47	0.78
-e final	1	0.01%	0.42%	0.04	0.96
-ee/-y/-i final	120	1.19%	50.42%	4.52	0.63
-en final	1	0.01%	0.42%	0.04	0.98
-k final	3	0.03%	1.26%	0.11	0.94
-o final	0	0.00%	0.00%	0.00	NA
-r final	66	0.65%	27.73%	2.49	0.76
-s final	4	0.04%	1.68%	0.15	0.95
-ti final	0	0.00%	0.00%	0.00	NA
-um final	4	0.04%	1.68%	0.15	0.98
VOWEL SUBSTITUTION-TYPE					
TOTAL	**531**	**5.25%**		**20.01**	
a-for-*ai*	6	0.06%	1.13%	0.23	0.95
a-for-*au*	0	0.00%	0.00%	0.00	NA
a-for-*e*	83	0.82%	15.63%	3.13	0.57
a-for-*ea*	15	0.15%	2.82%	0.57	0.76
a-for-*ei*	1	0.01%	0.19%	0.04	0.97
a-for-*o*	20	0.20%	3.77%	0.75	0.80
a-for-*ou*	3	0.03%	0.56%	0.11	0.97
a-for-*u*	3	0.03%	0.56%	0.11	0.90
a-for-*y*	1	0.01%	0.19%	0.04	0.99
ai-for-*i*	0	0.00%	0.00%	0.00	NA
au-for-*a*	0	0.00%	0.00%	0.00	NA
au-for-*o*	0	0.00%	0.00%	0.00	NA
ay-for-*ea*	5	0.05%	0.94%	0.19	0.96
ay-for-*y*	1	0.01%	0.19%	0.04	0.97
e-for-*a*	15	0.15%	2.82%	0.57	0.75
e-for-*ai*	4	0.04%	0.75%	0.15	0.91
e-for-*ay*	3	0.03%	0.56%	0.11	0.89
e-for-*i*	7	0.07%	1.32%	0.26	0.86
e-for-*o*	39	0.39%	7.34%	1.47	0.82
e-for-*ou*	41	0.41%	7.72%	1.54	0.77
e-for-*u*	8	0.08%	1.51%	0.30	0.86
e-for-*y*	0	0.00%	0.00%	0.00	NA
ee/y-for-*a*	5	0.05%	0.94%	0.19	0.92
ee-for-*e*	4	0.04%	0.75%	0.15	0.96
ee-for-*ea*	5	0.05%	0.94%	0.19	0.94
ee/y-for-*i*	13	0.13%	2.45%	0.49	0.80
ee-for-*ou*	2	0.02%	0.38%	0.08	0.96

Feature	N	% Global	% Local	Freq.	DP
ee-for-*y*	1	0.01%	0.19%	0.04	0.99
i-for-*a*	12	0.12%	2.26%	0.45	0.91
i-for-*ai*	2	0.02%	0.38%	0.08	0.96
i-for-*e*	89	0.88%	16.76%	3.35	0.67
i-for-*ea*	0	0.00%	0.00%	0.00	NA
i-for-*o*	0	0.00%	0.00%	0.00	NA
i-for-*oi*/*oy*	6	0.06%	1.13%	0.23	0.87
i-for-*u*	4	0.04%	0.75%	0.15	0.87
i-for-*ui*	0	0.00%	0.00%	0.00	NA
i-for-*y*	0	0.00%	0.00%	0.00	NA
o-for-*a*	8	0.08%	1.51%	0.30	0.86
o-for-*au*	8	0.08%	1.51%	0.30	0.92
o-for-*ou*	49	0.48%	9.23%	1.85	0.91
o-for-*u*	6	0.06%	1.13%	0.23	0.85
oo-for-*e*	1	0.01%	0.19%	0.04	0.96
oo-for-*o*	0	0.00%	0.00%	0.00	NA
oo-for-*ou*	0	0.00%	0.00%	0.00	NA
u-for-*a*	4	0.04%	0.75%	0.15	0.95
u-for-*e*	33	0.33%	6.21%	1.24	0.69
u-for-*i*	12	0.12%	2.26%	0.45	0.78
u-for-*o*	8	0.08%	1.51%	0.30	0.85
y-for-*ay*	1	0.01%	0.19%	0.04	0.99
ya-for-*ea*	3	0.03%	0.56%	0.11	0.95

Appendix F

Features Tables: Indian Dialogue

For all features tables (Appendices D-G), the column labels denote the following: *N* is the raw number of occurrences, *% Global* is the percentage a feature or category contributes to all coded features, *% Local* is the percentage a feature or subcategory contributes to the higher order category in which it is embedded, *Freq.* is the normalized frequency of a feature or category (per 1000 words), and *DP* is the deviation of proportions (or dispersion measure).

Feature	N	% Global	% Local	Freq.	DP
FEATURES-TYPE					
TOTAL	**3710**			**222.97**	
lexical	1148	30.94%		68.99	
morphosyntactic	1970	53.10%		118.40	
orthographic	17	0.46%		1.02	
phonological	575	15.50%		34.56	
LEXICAL-TYPE					
TOTAL	**1148**	**30.94%**		**68.99**	
general vocabulary	84	2.26%	7.32%	5.05	0.71
address	698	18.81%	60.80%	41.95	0.39
self address	123	3.32%	10.71%	7.39	0.49
inserts	7	0.19%	0.61%	0.42	0.92
wh- word	9	0.24%	0.78%	0.54	0.89
class shifting	12	0.32%	1.05%	0.72	0.72
lexical substitution	44	1.19%	3.83%	2.64	0.54
neologism	2	0.05%	0.17%	0.12	0.94
reduplication	2	0.05%	0.17%	0.12	0.97
code mixing	167	4.50%	14.55%	10.04	0.56
MORPHOSYNTACTIC-TYPE					
TOTAL	**1970**	**53.10%**		**118.40**	
pronoun	146	3.94%	7.41%	8.77	
noun phrase	526	14.18%	26.70%	31.61	
verb phrase	971	26.17%	49.29%	58.36	
adjective/adverb	58	1.56%	2.94%	3.49	
negation	68	1.83%	3.45%	4.09	
complementation	2	0.05%	0.10%	0.12	
discourse organization	199	5.36%	10.10%	11.96	
PRONOUN-TYPE					
TOTAL	**146**	**3.94%**		**8.77**	
me subject	73	1.97%	50.00%	4.39	0.72
my subject	0	0.00%	0.00%	0.00	NA
us subject	0	0.00%	0.00%	0.00	NA
I'se subject	0	0.00%	0.00%	0.00	NA
youse plural	0	0.00%	0.00%	0.00	NA
her subject	3	0.08%	2.05%	0.18	0.94
him subject	26	0.70%	17.81%	1.56	0.77
them subject	6	0.16%	4.11%	0.36	0.94
me possessive	2	0.05%	1.37%	0.12	0.95

Feature	N	% Global	% Local	Freq.	DP
you possessive	2	0.05%	1.37%	0.12	0.98
he possessive	0	0.00%	0.00%	0.00	NA
him possessive	32	0.86%	21.92%	1.92	0.75
him's possessive	0	0.00%	0.00%	0.00	NA
she possessive	0	0.00%	0.00%	0.00	NA
them possessive	1	0.03%	0.68%	0.06	0.94
my object	0	0.00%	0.00%	0.00	NA
we object	0	0.00%	0.00%	0.00	NA
he object	1	0.03%	0.68%	0.06	0.94
she object	0	0.00%	0.00%	0.00	NA
reflexive paradigm possessive	0	0.00%	0.00%	0.00	NA
reflexive paradigm objective	0	0.00%	0.00%	0.00	NA
reflexive paradigm subjective	0	0.00%	0.00%	0.00	NA
object reflexive	0	0.00%	0.00%	0.00	NA
demonstrative *him*	0	0.00%	0.00%	0.00	NA
demonstrative *them*	0	0.00%	0.00%	0.00	NA
NOUN PHRASE-TYPE					
TOTAL	**526**	**14.18%**		**31.61**	
determiner	452	12.18%	85.93%	27.17	
plural marking	45	1.21%	8.56%	2.70	0.57
genitive marking	18	0.49%	3.42%	1.08	0.73
-man nominal suffix	11	0.30%	2.09%	0.66	0.80
DETERMINER-TYPE					
TOTAL	**452**	**12.18%**		**27.17**	
definite/indefinite generalization	1	0.03%	0.22%	0.06	0.94
piece determiner	0	0.00%	0.00%	0.00	NA
zero determiner	451	12.16%	99.78%	27.10	0.42
determiner insertion	0	0.00%	0.00%	0.00	NA
VERB PHRASE-TYPE					
TOTAL	**971**	**26.17%**		**58.36**	
time marking	206	5.55%	21.22%	12.38	
auxiliary/modals	197	5.31%	20.29%	11.84	
agreement/aspect-marking	537	14.47%	55.30%	32.27	
infinitive	31	0.84%	3.19%	1.86	
TIME-MARKING-TYPE					
TOTAL	**206**	**5.55%**		**12.38**	
invariant stem	142	3.83%	68.93%	8.53	0.64
regular past	0	0.00%	0.00%	0.00	NA
participle past	2	0.05%	0.97%	0.12	0.88
-ing present/past	44	1.19%	21.36%	2.64	0.84
unmarked present participle	18	0.49%	8.74%	1.08	0.82
AUXILIARY/MODAL-TYPE					
TOTAL	**197**	**5.31%**		**11.84**	
null modal	80	2.16%	40.61%	4.81	0.58
null *wh*-aux	25	0.67%	12.69%	1.50	0.66
auxiliary deletion	68	1.83%	34.52%	4.09	0.68

Feature	N	% Global	% Local	Freq.	DP
aux substitution	1	0.03%	0.51%	0.06	0.98
done completive	4	0.11%	2.03%	0.24	0.93
go auxiliary	8	0.22%	4.06%	0.48	0.90
make causative	11	0.30%	5.58%	0.66	0.81
AGREEMENT/ASPECT-TYPE					
TOTAL	**537**	**14.47%**		**32.27**	
invariant present	211	5.69%	39.29%	12.68	0.48
zero copula	271	7.30%	50.47%	16.29	0.47
first-person *-s*	0	0.00%	0.00%	0.00	NA
second-person *-s*	0	0.00%	0.00%	0.00	NA
plural *-s*	0	0.00%	0.00%	0.00	NA
a- prefixing	0	0.00%	0.00%	0.00	NA
invariant *be*	20	0.54%	3.72%	1.20	0.83
was/were generalization	1	0.03%	0.19%	0.06	0.94
is/are generalization	4	0.11%	0.74%	0.24	0.88
am non-first-person	0	0.00%	0.00%	0.00	NA
's/'se/is first-person-be	0	0.00%	0.00%	0.00	NA
's/'se first-person-have	0	0.00%	0.00%	0.00	NA
have/has generalization	5	0.13%	0.93%	0.30	0.78
lexical *have* deletion	2	0.05%	0.37%	0.12	0.93
belong copular	0	0.00%	0.00%	0.00	NA
catch generalized	0	0.00%	0.00%	0.00	NA
got possessive/existential	23	0.62%	4.28%	1.38	0.70
INFINITIVE-TYPE					
TOTAL	**31**	**0.84%**		**1.86**	
null particle	31	0.84%	100.00%	1.86	0.68
infinitive *-s*	0	0.00%	0.00%	0.00	NA
ADJECTIVE/ADVERB-TYPE					
TOTAL	**58**	**1.56%**		**3.49**	
adjective/adverb leveling	12	0.32%	20.69%	0.72	0.73
comparatives	4	0.11%	6.90%	0.24	0.87
much/many generalization	4	0.11%	6.90%	0.24	0.92
heap intensifier	1	0.03%	1.72%	0.06	0.94
much intensifier	11	0.30%	18.97%	0.66	0.89
plenty intensifier	22	0.59%	37.93%	1.32	0.82
-side locative-suffix	0	0.00%	0.00%	0.00	NA
-time temporal-suffix	4	0.11%	6.90%	0.24	0.96
NEGATION-TYPE					
TOTAL	**68**	**1.83%**		**4.09**	
multiple negation	2	0.05%	2.94%	0.12	0.94
no preverbal	46	1.24%	67.65%	2.76	0.62
no generalization	18	0.49%	26.47%	1.08	0.80
ain't as *be*	2	0.05%	2.94%	0.12	0.94
ain't as *have*	0	0.00%	0.00%	0.00	NA
ain't as negator	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
COMPLEMENTATION-TYPE					
TOTAL	**2**	**0.05%**		**0.12**	
for complement	2	0.05%	100.00%	0.12	0.90
DISCOURSE ORGANIZATION-TYPE					
TOTAL	**199**	**5.36%**		**11.96**	
null subject	84	2.26%	42.21%	5.05	0.52
null object	26	0.70%	13.07%	1.56	0.64
null preposition	55	1.48%	27.64%	3.31	0.57
null coordinator	3	0.08%	1.51%	0.18	0.92
null subordinator	0	0.00%	0.00%	0.00	NA
preposition insertion	3	0.08%	1.51%	0.18	0.90
word order	28	0.75%	14.07%	1.68	0.71
ORTHOGRAPHIC-TYPE					
TOTAL	**17**	**0.46%**		**1.02**	
eye dialect	7	0.19%	41.18%	0.42	0.81
ambiguous	10	0.27%	58.82%	0.60	0.80
PHONOLOGICAL-TYPE					
TOTAL	**575**	**15.50%**		**34.56**	
consonant substitution	253	6.82%	44.00%	15.21	
consonant deletion	60	1.62%	10.43%	3.61	
insertion	72	1.94%	12.52%	4.33	
vowel substitution	113	3.05%	19.65%	6.79	
metathesis	0	0.00%	0.00%	0.00	NA
syllable deletion	42	1.13%	7.30%	2.52	0.71
doubling	35	0.94%	6.09%	2.10	0.79
exaggerated	0	0.00%	0.00%	0.00	NA
CONSONANT SUBSTITUTION-TYPE					
TOTAL	**253**	**6.82%**		**15.21**	
b-for-*p*	0	0.00%	0.00%	0.00	NA
b-for-*v*/*f*	23	0.62%	9.09%	1.38	0.82
c-for-*b*	0	0.00%	0.00%	0.00	NA
ch-for-*sh*	1	0.03%	0.40%	0.06	0.99
ch-for-*t*	1	0.03%	0.40%	0.06	0.99
d-for-*r*	0	0.00%	0.00%	0.00	NA
d-for-*t*	0	0.00%	0.00%	0.00	NA
f-for-*p*	0	0.00%	0.00%	0.00	NA
f-for-*th*	0	0.00%	0.00%	0.00	NA
f-for-*v*	0	0.00%	0.00%	0.00	NA
j/*g*-for-*d*	0	0.00%	0.00%	0.00	NA
k-for-*t*	0	0.00%	0.00%	0.00	NA
l-for-*f*	0	0.00%	0.00%	0.00	NA
l-for-*n*	0	0.00%	0.00%	0.00	NA
l-for-*r*	0	0.00%	0.00%	0.00	NA
l-for-*t*/*d*	0	0.00%	0.00%	0.00	NA
l-for-*th*	0	0.00%	0.00%	0.00	NA
l-for-*w*	0	0.00%	0.00%	0.00	NA
l-for-*y*	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
m-for-*n*	0	0.00%	0.00%	0.00	NA
n-for-*m*	0	0.00%	0.00%	0.00	NA
n-for-*ng*	7	0.19%	2.77%	0.42	0.88
p-for-*b*	0	0.00%	0.00%	0.00	NA
p-for-*f*	1	0.03%	0.40%	0.06	0.99
r-for-*t*/*d*	1	0.03%	0.40%	0.06	0.99
r-for-*f*	0	0.00%	0.00%	0.00	NA
r-for-*l*	1	0.03%	0.40%	0.06	0.99
s-for-*sh*/*ch*	9	0.24%	3.56%	0.54	0.94
s/*c*-for-*t*	0	0.00%	0.00%	0.00	NA
s-for-*th*	2	0.05%	0.79%	0.12	0.94
sh-for-*ch*	0	0.00%	0.00%	0.00	NA
sh-for-*j*	0	0.00%	0.00%	0.00	NA
sh-for-*s*	1	0.03%	0.40%	0.06	0.99
t-for-*d*	2	0.05%	0.79%	0.12	1.00
t-for-*s*	1	0.03%	0.40%	0.06	0.99
t/*d*-for-*th*	184	4.96%	72.73%	11.06	0.64
th-for-*s*	0	0.00%	0.00%	0.00	NA
th-for-*t*	0	0.00%	0.00%	0.00	NA
v-for-*w*/*wh*	14	0.38%	5.53%	0.84	0.95
w-for-*v*	2	0.05%	0.79%	0.12	0.93
y-for-*h*	3	0.08%	1.19%	0.18	0.99
y-for-*j*	0	0.00%	0.00%	0.00	NA
CONSONANT DELETION-TYPE					
TOTAL	**60**	**1.62%**		**3.61**	
cluster reduction	25	0.67%	41.67%	1.50	0.76
word-initial deletion	21	0.57%	35.00%	1.26	0.84
word-final deletion	14	0.38%	23.33%	0.84	0.84
inter-vocalic deletion	0	0.00%	0.00%	0.00	NA
INSERTION-TYPE					
TOTAL	**72**	**1.94%**		**4.33**	
initial insertion	13	0.35%	18.06%	0.78	
medial insertion	7	0.19%	9.72%	0.42	
final insertion	52	1.40%	72.22%	3.13	
INITIAL INSERTION-TYPE					
TOTAL	**13**	**0.35%**		**0.78**	
a- initial	0	0.00%	0.00%	0.00	NA
ee-/*i*- initial	13	0.35%	100.00%	0.78	0.95
h- initial	0	0.00%	0.00%	0.00	NA
y- initial	0	0.00%	0.00%	0.00	NA
MEDIAL INSERTION-TYPE					
TOTAL	**7**	**0.19%**		**0.42**	
-*e*- medial	2	0.05%	28.57%	0.12	0.99
-*ee*-/-*y*- medial	0	0.00%	0.00%	0.00	NA
-*es*- medial	0	0.00%	0.00%	0.00	NA
-*k*- medial	0	0.00%	0.00%	0.00	NA
-*l*- pre-vocalic	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
-*m*- medial	1	0.03%	14.29%	0.06	0.99
-*n*- medial	0	0.00%	0.00%	0.00	NA
-*r*- pre-vocalic	0	0.00%	0.00%	0.00	NA
-*r*- post-vocalic	4	0.11%	57.14%	0.24	0.92
-*ob*- medial	0	0.00%	0.00%	0.00	NA
-*u*- medial	0	0.00%	0.00%	0.00	NA
FINAL INSERTION-TYPE					
TOTAL	**52**	**1.40%**		**3.13**	
-*a* final	19	0.51%	36.54%	1.14	0.93
-*e* final	6	0.16%	11.54%	0.36	0.92
-*ee*/-*y*/-*i* final	24	0.65%	46.15%	1.44	0.78
-*en* final	0	0.00%	0.00%	0.00	NA
-*k* final	0	0.00%	0.00%	0.00	NA
-*o* final	1	0.03%	1.92%	0.06	0.94
-*r* final	0	0.00%	0.00%	0.00	NA
-*s* final	1	0.03%	1.92%	0.06	1.00
-*ti* final	1	0.03%	1.92%	0.06	0.94
-*um* final	0	0.00%	0.00%	0.00	NA
VOWEL SUBSTITUTION-TYPE					
TOTAL	**113**	**3.05%**		**6.79**	
a-for-*ai*	2	0.05%	1.77%	0.12	0.97
a-for-*au*	0	0.00%	0.00%	0.00	NA
a-for-*e*	5	0.13%	4.42%	0.30	0.91
a-for-*ea*	0	0.00%	0.00%	0.00	NA
a-for-*ei*	0	0.00%	0.00%	0.00	NA
a-for-*o*	2	0.05%	1.77%	0.12	1.00
a-for-*ou*	0	0.00%	0.00%	0.00	NA
a-for-*u*	1	0.03%	0.88%	0.06	0.99
a-for-*y*	0	0.00%	0.00%	0.00	NA
ai-for-*i*	2	0.05%	1.77%	0.12	0.94
au-for-*a*	4	0.11%	3.54%	0.24	0.98
au-for-*o*	1	0.03%	0.88%	0.06	0.99
ay-for-*ea*	0	0.00%	0.00%	0.00	NA
ay-for-*y*	0	0.00%	0.00%	0.00	NA
e-for-*a*	5	0.13%	4.42%	0.30	0.89
e-for-*ai*	0	0.00%	0.00%	0.00	NA
e-for-*ay*	0	0.00%	0.00%	0.00	NA
e-for-*i*	0	0.00%	0.00%	0.00	NA
e-for-*o*	0	0.00%	0.00%	0.00	NA
e-for-*ou*	0	0.00%	0.00%	0.00	NA
e-for-*u*	0	0.00%	0.00%	0.00	NA
e-for-*y*	1	0.03%	0.88%	0.06	0.99
ee/*y*-for-*a*	0	0.00%	0.00%	0.00	NA
ee-for-*e*	4	0.11%	3.54%	0.24	0.92
ee-for-*ea*	0	0.00%	0.00%	0.00	NA
ee/*y*-for-*i*	11	0.30%	9.73%	0.66	0.82
ee-for-*ou*	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
ee-for-*y*	42	1.13%	37.17%	2.52	0.76
i-for-*a*	0	0.00%	0.00%	0.00	NA
i-for-*ai*	0	0.00%	0.00%	0.00	NA
i-for-*e*	13	0.35%	11.50%	0.78	0.92
i-for-*ea*	1	0.03%	0.88%	0.06	0.99
i-for-*o*	0	0.00%	0.00%	0.00	NA
i-for-*oi*/*oy*	0	0.00%	0.00%	0.00	NA
i-for-*u*	0	0.00%	0.00%	0.00	NA
i-for-*ui*	0	0.00%	0.00%	0.00	NA
i-for-*y*	1	0.03%	0.88%	0.06	0.99
o-for-*a*	2	0.05%	1.77%	0.12	0.91
o-for-*au*	1	0.03%	0.88%	0.06	0.99
o-for-*ou*	1	0.03%	0.88%	0.06	0.99
o-for-*u*	1	0.03%	0.88%	0.06	0.99
oo-for-*e*	0	0.00%	0.00%	0.00	NA
oo-for-*o*	3	0.08%	2.65%	0.18	0.99
oo-for-*ou*	3	0.08%	2.65%	0.18	0.99
u-for-*a*	1	0.03%	0.88%	0.06	0.97
u-for-*e*	6	0.16%	5.31%	0.36	0.93
u-for-*i*	0	0.00%	0.00%	0.00	NA
u-for-*o*	0	0.00%	0.00%	0.00	NA
y-for-*ay*	0	0.00%	0.00%	0.00	NA
ya-for-*ea*	0	0.00%	0.00%	0.00	NA

Appendix G

Features Tables: Chinese Dialogue

For all features tables (Appendices D-G), the column labels denote the following: *N* is the raw number of occurrences, *% Global* is the percentage a feature or category contributes to all coded features, *% Local* is the percentage a feature or subcategory contributes to the higher order category in which it is embedded, *Freq.* is the normalized frequency of a feature or category (per 1000 words), and *DP* is the deviation of proportions (or dispersion measure).

Feature	N	% Global	% Local	Freq.	DP
FEATURES-TYPE					
TOTAL	**4366**			**547.74**	
lexical	529	12.12%		66.37	
morphosyntactic	1931	44.23%		242.25	
orthographic	40	0.92%		5.02	
phonological	1866	42.74%		234.10	
LEXICAL-TYPE					
TOTAL	**529**	**12.12%**		**66.37**	
general vocabulary	209	4.79%	39.51%	26.22	0.44
address	128	2.93%	24.20%	16.06	0.48
self address	76	1.74%	14.37%	9.53	0.48
inserts	38	0.87%	7.18%	4.77	0.71
wh- word	13	0.30%	2.46%	1.63	0.65
class shifting	5	0.11%	0.95%	0.63	0.65
lexical substitution	23	0.53%	4.35%	2.89	0.46
neologism	0	0.00%	0.00%	0.00	NA
reduplication	16	0.37%	3.02%	2.01	0.62
code mixing	21	0.48%	3.97%	2.63	0.72
MORPHOSYNTACTIC-TYPE					
TOTAL	**1931**	**44.23%**		**242.25**	
pronoun	192	4.40%	9.94%	24.09	
noun phrase	391	8.96%	20.25%	49.05	
verb phrase	796	18.23%	41.22%	99.86	
adjective/adverb	82	1.88%	4.25%	10.29	
negation	196	4.49%	10.15%	24.59	
complementation	4	0.09%	0.21%	0.50	
discourse organization	270	6.18%	13.98%	33.87	
PRONOUN-TYPE					
TOTAL	**192**	**4.40%**		**24.09**	
me subject	92	2.11%	47.92%	11.54	0.69
my subject	43	0.98%	22.40%	5.39	0.68
us subject	0	0.00%	0.00%	0.00	NA
I'se subject	0	0.00%	0.00%	0.00	NA
youse plural	0	0.00%	0.00%	0.00	NA
her subject	0	0.00%	0.00%	0.00	NA
him subject	26	0.60%	13.54%	3.26	0.65
them subject	1	0.02%	0.52%	0.13	0.88
me possessive	3	0.07%	1.56%	0.38	0.94

Feature	N	% Global	% Local	Freq.	DP
you possessive	4	0.09%	2.08%	0.50	0.74
he possessive	2	0.05%	1.04%	0.25	0.86
him possessive	3	0.07%	1.56%	0.38	0.92
him's possessive	0	0.00%	0.00%	0.00	NA
she possessive	1	0.02%	0.52%	0.13	0.99
them possessive	0	0.00%	0.00%	0.00	NA
my object	5	0.11%	2.60%	0.63	0.84
we object	1	0.02%	0.52%	0.13	1.00
he object	5	0.11%	2.60%	0.63	0.73
she object	0	0.00%	0.00%	0.00	NA
reflexive paradigm possessive	0	0.00%	0.00%	0.00	NA
reflexive paradigm objective	2	0.05%	1.04%	0.25	0.91
reflexive paradigm subjective	0	0.00%	0.00%	0.00	NA
object reflexive	1	0.02%	0.52%	0.13	0.98
demonstrative *him*	0	0.00%	0.00%	0.00	NA
demonstrative *them*	3	0.07%	1.56%	0.38	0.95
NOUN PHRASE-TYPE					
TOTAL	**391**	**8.96%**		**49.05**	
determiner	287	6.57%	73.40%	36.01	
plural marking	56	1.28%	14.32%	7.03	0.59
genitive marking	14	0.32%	3.58%	1.76	0.72
-man nominal suffix	34	0.78%	8.70%	4.27	0.59
DETERMINER-TYPE					
TOTAL	**287**	**6.57%**		**36.01**	
definite/indefinite generalization	0	0.00%	0.00%	0.00	NA
piece determiner	30	0.69%	10.45%	3.76	0.58
zero determiner	257	5.89%	89.55%	32.24	0.24
determiner insertion	0	0.00%	0.00%	0.00	NA
VERB PHRASE-TYPE					
TOTAL	**796**	**18.23%**		**99.86**	
time marking	149	3.41%	18.72%	18.69	
auxiliary/modals	173	3.96%	21.73%	21.70	
agreement/aspect-marking	401	9.18%	50.38%	50.31	
infinitive	73	1.67%	9.17%	9.16	
TIME-MARKING-TYPE					
TOTAL	**149**	**3.41%**		**18.69**	
invariant stem	137	3.14%	91.95%	17.19	0.33
regular past	1	0.02%	0.67%	0.13	0.98
participle past	3	0.07%	2.01%	0.38	0.93
-ing present/past	0	0.00%	0.00%	0.00	NA
unmarked present participle	8	0.18%	5.37%	1.00	0.62
AUXILIARY/MODAL-TYPE					
TOTAL	**173**	**3.96%**		**21.70**	
null modal	105	2.40%	60.69%	13.17	0.36
null *wh*-aux	28	0.64%	16.18%	3.51	0.40
auxiliary deletion	13	0.30%	7.51%	1.63	0.65

Feature	N	% Global	% Local	Freq.	DP
aux substitution	1	0.02%	0.58%	0.13	0.95
done completive	0	0.00%	0.00%	0.00	NA
go auxiliary	8	0.18%	4.62%	1.00	0.58
make causative	18	0.41%	10.40%	2.26	0.75
AGREEMENT/ASPECT-TYPE					
TOTAL	**401**	**9.18%**		**50.31**	
invariant present	106	2.43%	26.43%	13.30	0.30
zero copula	187	4.28%	46.63%	23.46	0.27
first-person *-s*	1	0.02%	0.25%	0.13	0.99
second-person *-s*	0	0.00%	0.00%	0.00	NA
plural *-s*	1	0.02%	0.25%	0.13	0.98
a- prefixing	0	0.00%	0.00%	0.00	NA
invariant *be*	2	0.05%	0.50%	0.25	0.89
was/were generalization	2	0.05%	0.50%	0.25	0.98
is/are generalization	1	0.02%	0.25%	0.13	0.98
am non-first-person	2	0.05%	0.50%	0.25	0.91
's/'se/is first-person-be	1	0.02%	0.25%	0.13	0.98
's/'se first-person-have	2	0.05%	0.50%	0.25	0.89
have/has generalization	15	0.34%	3.74%	1.88	0.59
lexical *have* deletion	4	0.09%	1.00%	0.50	0.87
belong copular	21	0.48%	5.24%	2.63	0.69
catch generalized	29	0.66%	7.23%	3.64	0.70
got possessive/existential	27	0.62%	6.73%	3.39	0.52
INFINITIVE-TYPE					
TOTAL	**73**	**1.67%**		**9.16**	
null particle	72	1.65%	98.63%	9.03	0.34
infinitive *-s*	1	0.02%	1.37%	0.13	0.98
ADJECTIVE/ADVERB-TYPE					
TOTAL	**82**	**1.88%**		**10.29**	
adjective/adverb leveling	1	0.02%	1.22%	0.13	0.95
comparatives	8	0.18%	9.76%	1.00	0.83
much/many generalization	4	0.09%	4.88%	0.50	0.85
heap intensifier	25	0.57%	30.49%	3.14	0.86
much intensifier	17	0.39%	20.73%	2.13	0.60
plenty intensifier	21	0.48%	25.61%	2.63	0.68
-side locative-suffix	3	0.07%	3.66%	0.38	0.85
-time temporal-suffix	3	0.07%	3.66%	0.38	0.85
NEGATION-TYPE					
TOTAL	**196**	**4.49%**		**24.59**	
multiple negation	6	0.14%	3.06%	0.75	0.56
no preverbal	168	3.85%	85.71%	21.08	0.29
no generalization	22	0.50%	11.22%	2.76	0.53
ain't as *be*	0	0.00%	0.00%	0.00	NA
ain't as *have*	0	0.00%	0.00%	0.00	NA
ain't as negator	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
COMPLEMENTATION-TYPE					
TOTAL	**4**	**0.09%**		**0.50**	
for complement	4	0.09%	100.00%	0.50	0.91
DISCOURSE ORGANIZATION-TYPE					
TOTAL	**270**	**6.18%**		**33.87**	
null subject	133	3.05%	49.26%	16.69	0.35
null object	40	0.92%	14.81%	5.02	0.52
null preposition	65	1.49%	24.07%	8.15	0.44
null coordinator	4	0.09%	1.48%	0.50	0.86
null subordinator	3	0.07%	1.11%	0.38	0.85
preposition insertion	2	0.05%	0.74%	0.25	0.83
word order	23	0.53%	8.52%	2.89	0.68
ORTHOGRAPHIC-TYPE					
TOTAL	**40**	**0.92%**		**5.02**	
eye dialect	11	0.25%	27.50%	1.38	0.79
ambiguous	29	0.66%	72.50%	3.64	0.65
PHONOLOGICAL-TYPE					
TOTAL	**1866**	**42.74%**		**234.10**	
consonant substitution	896	20.52%	48.02%	112.41	
consonant deletion	84	1.92%	4.50%	10.54	
insertion	689	15.78%	36.92%	86.44	
vowel substitution	87	1.99%	4.66%	10.91	
metathesis	1	0.02%	0.05%	0.13	0.96
syllable deletion	96	2.20%	5.14%	12.04	0.50
doubling	2	0.05%	0.11%	0.25	0.99
exaggerated	11	0.25%	0.59%	1.38	0.77
CONSONANT SUBSTITUTION-TYPE					
TOTAL	**896**	**20.52%**		**112.41**	
b-for-*p*	0	0.00%	0.00%	0.00	NA
b-for-*v*/*f*	36	0.82%	4.02%	4.52	0.55
c-for-*b*	0	0.00%	0.00%	0.00	NA
ch-for-*sh*	0	0.00%	0.00%	0.00	NA
ch-for-*t*	38	0.87%	4.24%	4.77	0.67
d-for-*r*	0	0.00%	0.00%	0.00	NA
d-for-*t*	0	0.00%	0.00%	0.00	NA
f-for-*p*	0	0.00%	0.00%	0.00	NA
f-for-*th*	11	0.25%	1.23%	1.38	0.78
f-for-*v*	3	0.07%	0.33%	0.38	0.99
j/*g*-for-*d*	0	0.00%	0.00%	0.00	NA
k-for-*t*	0	0.00%	0.00%	0.00	NA
l-for-*f*	1	0.02%	0.11%	0.13	0.95
l-for-*n*	0	0.00%	0.00%	0.00	NA
l-for-*r*	604	13.83%	67.41%	75.77	0.29
l-for-*t*/*d*	3	0.07%	0.33%	0.38	0.95
l-for-*th*	9	0.21%	1.00%	1.13	0.86
l-for-*w*	9	0.21%	1.00%	1.13	0.98
l-for-*y*	10	0.23%	1.12%	1.25	0.98

Feature	N	% Global	% Local	Freq.	DP
m-for-*n*	0	0.00%	0.00%	0.00	NA
n-for-*m*	0	0.00%	0.00%	0.00	NA
n-for-*ng*	4	0.09%	0.45%	0.50	0.76
p-for-*b*	0	0.00%	0.00%	0.00	NA
p-for-*f*	0	0.00%	0.00%	0.00	NA
r-for-*t*/*d*	0	0.00%	0.00%	0.00	NA
r-for-*f*	0	0.00%	0.00%	0.00	NA
r-for-*l*	3	0.07%	0.33%	0.38	0.94
s-for-*sh*/*ch*	7	0.16%	0.78%	0.88	0.74
s/*c*-for-*t*	1	0.02%	0.11%	0.13	0.99
s-for-*th*	1	0.02%	0.11%	0.13	0.96
sh-for-*ch*	0	0.00%	0.00%	0.00	NA
sh-for-*j*	0	0.00%	0.00%	0.00	NA
sh-for-*s*	0	0.00%	0.00%	0.00	NA
t-for-*d*	0	0.00%	0.00%	0.00	NA
t-for-*s*	2	0.05%	0.22%	0.25	0.88
t/*d*-for-*th*	140	3.21%	15.63%	17.56	0.80
th-for-*s*	0	0.00%	0.00%	0.00	NA
th-for-*t*	0	0.00%	0.00%	0.00	NA
v-for-*w*/*wh*	0	0.00%	0.00%	0.00	NA
w-for-*v*	14	0.32%	1.56%	1.76	0.95
y-for-*h*	0	0.00%	0.00%	0.00	NA
y-for-*j*	0	0.00%	0.00%	0.00	NA
CONSONANT DELETION-TYPE					
TOTAL	**84**	**1.92%**		**10.54**	
cluster reduction	40	0.92%	47.62%	5.02	0.60
word-initial deletion	19	0.44%	22.62%	2.38	0.68
word-final deletion	23	0.53%	27.38%	2.89	0.61
inter-vocalic deletion	2	0.05%	2.38%	0.25	0.85
INSERTION-TYPE					
TOTAL	**689**	**15.78%**		**86.44**	
initial insertion	4	0.09%	0.58%	0.50	
medial insertion	28	0.64%	4.06%	3.51	
final insertion	657	15.05%	95.36%	82.42	
INITIAL INSERTION-TYPE					
TOTAL	**4**	**0.09%**		**0.50**	
a- initial	2	0.05%	50.00%	0.25	0.95
ee-/*i*- initial	0	0.00%	0.00%	0.00	NA
h- initial	1	0.02%	25.00%	0.13	0.98
y- initial	1	0.02%	25.00%	0.13	0.86
MEDIAL INSERTION-TYPE					
TOTAL	**28**	**0.64%**		**3.51**	
-*e*- medial	0	0.00%	0.00%	0.00	NA
-*ee*-/-*y*- medial	4	0.09%	14.29%	0.50	0.86
-*es*- medial	0	0.00%	0.00%	0.00	NA
-*k*- medial	0	0.00%	0.00%	0.00	NA
-*l*- pre-vocalic	21	0.48%	75.00%	2.63	0.81

Feature	N	% Global	% Local	Freq.	DP
-m- medial	0	0.00%	0.00%	0.00	NA
-n- medial	0	0.00%	0.00%	0.00	NA
-r- pre-vocalic	0	0.00%	0.00%	0.00	NA
-r- post-vocalic	2	0.05%	7.14%	0.25	0.94
-ob- medial	0	0.00%	0.00%	0.00	NA
-u- medial	1	0.02%	3.57%	0.13	0.95
FINAL INSERTION-TYPE					
TOTAL	**657**	**15.05%**		**82.42**	
-a final	15	0.34%	2.28%	1.88	0.79
-e final	5	0.11%	0.76%	0.63	0.91
-ee/-y/-i final	616	14.11%	93.76%	77.28	0.32
-en final	0	0.00%	0.00%	0.00	NA
-k final	2	0.05%	0.30%	0.25	0.85
-o final	9	0.21%	1.37%	1.13	0.84
-r final	5	0.11%	0.76%	0.63	0.85
-s final	1	0.02%	0.15%	0.13	0.98
-ti final	0	0.00%	0.00%	0.00	NA
-um final	4	0.09%	0.61%	0.50	0.80
VOWEL SUBSTITUTION-TYPE					
TOTAL	**87**	**1.99%**		**10.91**	
a-for-*ai*	0	0.00%	0.00%	0.00	NA
a-for-*au*	2	0.05%	2.30%	0.25	0.94
a-for-*e*	2	0.05%	2.30%	0.25	0.89
a-for-*ea*	0	0.00%	0.00%	0.00	NA
a-for-*ei*	0	0.00%	0.00%	0.00	NA
a-for-*o*	9	0.21%	10.34%	1.13	0.75
a-for-*ou*	2	0.05%	2.30%	0.25	0.95
a-for-*u*	0	0.00%	0.00%	0.00	NA
a-for-*y*	1	0.02%	1.15%	0.13	0.94
ai-for-*i*	0	0.00%	0.00%	0.00	NA
au-for-*a*	0	0.00%	0.00%	0.00	NA
au-for-*o*	0	0.00%	0.00%	0.00	NA
ay-for-*ea*	0	0.00%	0.00%	0.00	NA
ay-for-*y*	0	0.00%	0.00%	0.00	NA
e-for-*a*	1	0.02%	1.15%	0.13	0.99
e-for-*ai*	1	0.02%	1.15%	0.13	0.99
e-for-*ay*	1	0.02%	1.15%	0.13	0.98
e-for-*i*	0	0.00%	0.00%	0.00	NA
e-for-*o*	6	0.14%	6.90%	0.75	0.80
e-for-*ou*	0	0.00%	0.00%	0.00	NA
e-for-*u*	2	0.05%	2.30%	0.25	0.98
e-for-*y*	0	0.00%	0.00%	0.00	NA
ee/y-for-*a*	1	0.02%	1.15%	0.13	1.00
ee-for-*e*	0	0.00%	0.00%	0.00	NA
ee-for-*ea*	0	0.00%	0.00%	0.00	NA
ee/y-for-*i*	1	0.02%	1.15%	0.13	0.98
ee-for-*ou*	0	0.00%	0.00%	0.00	NA

Feature	N	% Global	% Local	Freq.	DP
ee-for-*y*	13	0.30%	14.94%	1.63	0.80
i-for-*a*	10	0.23%	11.49%	1.25	0.93
i-for-*ai*	0	0.00%	0.00%	0.00	NA
i-for-*e*	7	0.16%	8.05%	0.88	0.71
i-for-*ea*	0	0.00%	0.00%	0.00	NA
i-for-*o*	1	0.02%	1.15%	0.13	0.98
i-for-*oi*/*oy*	0	0.00%	0.00%	0.00	NA
i-for-*u*	0	0.00%	0.00%	0.00	NA
i-for-*ui*	1	0.02%	1.15%	0.13	0.88
i-for-*y*	3	0.07%	3.45%	0.38	0.95
o-for-*a*	0	0.00%	0.00%	0.00	NA
o-for-*au*	3	0.07%	3.45%	0.38	0.86
o-for-*ou*	7	0.16%	8.05%	0.88	0.84
o-for-*u*	2	0.05%	2.30%	0.25	1.00
oo-for-*e*	1	0.02%	1.15%	0.13	0.99
oo-for-*o*	1	0.02%	1.15%	0.13	0.98
oo-for-*ou*	0	0.00%	0.00%	0.00	NA
u-for-*a*	0	0.00%	0.00%	0.00	NA
u-for-*e*	8	0.18%	9.20%	1.00	0.76
u-for-*i*	1	0.02%	1.15%	0.13	0.86
u-for-*o*	0	0.00%	0.00%	0.00	NA
y-for-*ay*	0	0.00%	0.00%	0.00	NA
ya-for-*ea*	0	0.00%	0.00%	0.00	NA