

Language in Place: A Computational Analysis of Dickens' "American Notes"

Diana C. Archibald

As Michaela Mahlberg points out in *Corpus Stylistics and Dickens' Fiction*, until recently scholarly investigations of Dickens' language have tended to "focus on the identification of stylistic features in the form of overviews or inventories illustrated with isolated examples rather than on the linguistic analysis of fictional worlds" (2). Mahlberg (2005-2018), Tomoji Tabata (1995, 2002), Garrett Stewart (2001), Masashiro Hori (2004), Patricia Ingham (2011), and others have been working to fill that gap within Charles Dickens studies by exploring what corpus linguistics can reveal about Dickens' fiction. In recent years, scholars have renewed interest in his nonfiction, especially *American Notes*, but the book has not attracted much attention from corpus linguists, who sometimes include it in fiction corpora rather than treating it as a categorically different text. This project applies basic computational and literary analysis methods to the study of *American Notes* (hereafter abbreviated *AN*) as part of Dickens' nonfiction corpus, offering preliminary findings. Corpus stylistics uses computer programs systematically to uncover hidden patterns and elements that may point to significant factors contributing to a text's meaning. By integrating literary and linguistic analysis, researchers can better determine when statistical significance is actually meaningful. Likewise, the aims of literary criticism guide the investigation in which corpus linguistic tools and methods are employed. Without the application of computational analysis tools, some of the clues to textual meaning in a

large body of work remain obscured. Without the use of literary analysis, statistical findings may not appear meaningful for literary criticism. Both approaches intertwined in one corpus stylistics study can yield interesting results. Each digital humanities project is unique, built through a series of choices that are tailored to the unique demands of texts, research questions, and critical landscapes.

Renewed scholarly debate about *AN* centers on a few questions.¹ Did Dickens actually dislike America as much as scholars have claimed? What was the true nature of Dickens' distaste for the young republic? Does Dickens treat New England favorably in *AN* only because it reminded him of England? And does this minor work deserve all of the attention it has garnered recently? This project contends that computational analysis can help open up new ways of seeing the text – both by itself and in relation to other works – and can produce important new evidence to deepen critical arguments. In particular, this project seeks to answer the following research question:

What stylistic differences can computational analysis reveal 1) between the way different regions in *AN* are portrayed within the book, and 2) between *AN* and Dickens' other nonfiction, and how do these differences illuminate Dickens' attitudes toward America?

I. METHODOLOGY

In brief, this project employs a methodology similar to the “corpus-driven approach” Mahlberg describes in her book, *Corpus Stylistics and Dickens' Fiction* (13), albeit applied to a much smaller corpus and using fewer linguistic analysis tools. The research question has guided the way this study was initially constructed, and scholarly debate provides a variety of answers to those questions, as a critical context for the study as a whole. However, the project was not undertaken in order to prove a hypothesis. Rather, computational analysis was performed to uncover what the data show, leading to a recursive process of emerging questions and additional data analysis designed to pursue leads. Theoretical assumptions certainly

underpin the modest conclusions of this small study, but every attempt was made at open exploration of the data without preconceived notions of what those data should reveal.

A. Text Selection and Preparation

In corpus linguistics, corpus design and compilation lay the foundation for a viable project. To construct an appropriate corpus, scholars must carefully consider a) what texts must be included in order to make the corpus representative; b) how many pieces to use in order to yield statistically significant results; and c) how the corpus should be organized into separate units (files) to facilitate the analysis. A project's research questions as well as practical considerations influence these decisions. The corpus for this project is comprised of Dickens' 1842 travel book *AN* (without prefaces or postscript) and three of his other nonfiction works for comparison (reference corpus). While this combined corpus is by no means exhaustive, it does include an extensive collection of representative samples of travel writing and literary sketches from the total number of available Dickens nonfiction texts, with 425,938 total words divided into 115 files. In corpus linguistics, this is a small study; most scholars analyze corpora with several million words. Since this is an exploratory effort, however, a comparatively small representative sample is sufficient to produce preliminary findings that can be further investigated in a more comprehensive study in the future.

The text most similar to *AN* is Dickens' other travel book, *Pictures from Italy* (1846) (hereafter abbreviated *PI*). This travelogue was published only a few years after *AN* and was written under similar circumstances – to earn money while traveling during a novel-writing hiatus. Dickens' 1842 trip to North America was five months long, and he brought his wife Catherine with him, though none of their children accompanied them. In 1844, Dickens and his whole family spent almost a year in Europe, mostly in Italy – from which experience he composed *PI*. *AN* is the longer of the two works, with *PI* at only 71% the length of the first book, despite covering more than twice the amount of time. The two

works have a similar style overall, with attention to the logistics of physical travel, description of local color, and accounts of sightseeing visits. Both also have their anomalies: *AN* with its slavery chapter, comprised of excerpts from newspaper accounts of runaway slaves; *PI* with its strange dream sequence chapter.

The reference corpus includes the short nonfiction pieces found in the first three sections of *Sketches by Boz* (1839). Originally published in several newspapers and periodicals from 1833-36, these pieces capture Dickens' early nonfiction style and share some elements in common with the travel books: keen observation of people and places, occasional inclusion of passionate social commentary, and engaging narrative. As Patricia Ingham asserts in her introduction to *AN*, "Certainly in respect of the generic nature of the characters and fragmented anecdotal narrative, the *Sketches by Boz* are broadly a model for *American Notes*" (xviii). The twelve short stories from the "Tales" section of *Sketches* have not been included since these are fictional works. That leaves a total of forty-four pieces from "Our Parish," "Scenes," and "Characters" that are incorporated into the reference corpus. Clearly, Dickens applies fiction techniques to his nonfiction writing, and contemporary theory problematizes the very notion of nonfiction; nevertheless, the distinction is an important one to investigate, especially since Dickens' nonfiction has received less scholarly attention than his fiction. All of the sketches and reminiscences from the complete version of *The Uncommercial Traveller* (1860, 1864-65, 1868-69) were also included in the reference corpus, providing later examples of Dickens' comparable nonfiction writings. Those thirty-seven articles form 34% of the corpus, with *Sketches by Boz* accounting for another 25%, making a total of 59% taken from Dickens' writings for periodicals. With *Pictures from Italy* providing another 17%, that brings the total for the reference corpus to 76% of the project and *AN* at 24%.

Several other works were not included in the corpus. The travelogue *The Lazy Tour of Two Idle Apprentices* (1857) was co-written with Wilkie Collins, and the use of the fictional personae, the exaggerated qualities of

these characters, and the interpolation of fictional tales create a very different text than Dickens' relatively straightforward travel books. *The Life of Our Lord* (written 1846-49) and *A Child's History of England* (1851-53) are distinctly different genres from *AN*. Articles from *Household Words* were also omitted, although they may be added to the corpus for future study. Although Dickens Journals Online has uploaded images of the journals, these are not available in downloadable text files needed for computational analysis and would be too labor intensive to create as text files for the present study. In the future, more comparable texts could be added to create a fully comprehensive corpus, but for a preliminary study, the reference corpus is reasonably extensive and representative of Dickens' nonfiction sketches.

All of the texts come from Project Gutenberg, where plain text files of public domain books are available for free download and use. Both *AN* and *PI* are transcribed from the 1913 Chapman and Hall combined edition and were originally uploaded to the Project Gutenberg site in 1996, with revised files released in 2013. The text for *Sketches by Boz* is taken from the 1903 Chapman and Hall edition, and *The Uncommercial Traveller* comes from the 1905 Chapman and Hall edition. Although these e-texts are not based on scholarly editions, for the purposes of corpus linguistics analysis, minor errors in the texts do not impede analysis since scholars always cross-check findings with a scholarly edition to ensure accuracy of key claims.²

AN was divided three different ways: by the original 18 chapters and – given the importance of geography to this study – by 21 location-modified enhanced or fractured chapter divisions and by 6 regional clusters. For instance, the very end of the second chapter includes a page-long description of Boston (the subject of the following chapter), so the end of chapter 2 was added to the start of chapter 3, creating an enhanced chapter file. Likewise, sometimes material within a chapter was excised, thus creating a fractured chapter, as in the 590-word description of Halifax, Canada appearing in the Atlantic Crossing story of chapter 2. While creating a file that small presents some challenges for linguistic analysis,

given the importance of geography to the research question, these adjustments were necessary for the analysis.

B. Corpus Linguistics Tools and Method of Data Generation

Three open-source corpus linguistics tools were used in this project: Voyant Tools, AntConc, and CLiC Dickens, as well as two other tools: R programming language and Microsoft Excel. Voyant Tools is “a web-based text reading and analysis environment.”³ Researchers upload their own texts into the environment and use an array of tools to process information, generating everything from word clouds and scatter plots to corpus summaries of high-level statistics about the uploaded texts and their relationship to one another. Voyant was the application used most frequently for this project; after completing analyses using other tools, Voyant helped uncover further information about those data. Although stop word lists (common words the application will leave out of calculations such as frequency counts) can be edited, users have no control over the underlying settings that generate results (for instance, setting P-values). Still, Voyant is indispensable, since users can upload multiple versions of texts to analyze and compare findings. Results are also easy to interpret in a wide range of user-friendly presentation tools.

Unlike Voyant Tools, AntConc, “a freeware corpus analysis toolkit for concordancing and text analysis,”⁴ allows the user considerable control over the settings, but the interface is much simpler and utilitarian. After downloading the free program, users can run their own text files through the software to extract data. For this project AntConc was most useful in determining the keyness of *AN* keywords in relation to the reference corpus. In corpus linguistics, keyness is measured in order to determine which words that surface in frequency counts are truly significant either in terms of overuse or underuse compared to the reference corpus and expected norms. To establish keyness, linguistic patterns in English must be taken into account, in addition to the usual measures such as effect size and P-value. AntConc employs a rigorous

algorithm to determine keyness. While this tool may be more challenging to use for those new to corpus linguistics, this program offers a high degree of reliability and flexibility.

For this project, AntConc generated positive and negative keyness lists that were then used to adjust the stop list for Voyant and analyze words in the list that AntConc confirmed had a high degree of keyness. The keyness words were also cross-checked against the Dickens fiction corpus and nineteenth-century non-Dickens reference corpus in CLiC Dickens, a web application "designed specifically for the analysis of literary texts." CLiC is comprised of pre-loaded corpora available to "demonstrate through corpus stylistics how computer-assisted methods can be used to study literary texts and lead to new insights into how readers perceive fictional characters."⁵ Dickens' novels can be analyzed as individual texts or as a group. For this study, the most significant *AN* terms with keyness were checked within Dickens' novels and, for good measure, in CLiC's default reference corpus of other nineteenth-century fiction to determine how unusual these key words are within Dickens' nonfiction, Dickens' fiction, and other writers' fiction. While the settings in CLiC have some degree of flexibility, this app does not replicate AntConc's algorithms. Cross-checking results across applications is acceptable for a preliminary study such as this, but for further analysis it would be better to use the same program to analyze keyness.

Data generated from these corpus linguistics tools was also manipulated and analyzed using R and Excel to determine significance. For instance, initial computational analysis of text files was conducted using Voyant. Preliminary findings of word counts, words per sentence, vocabulary density, and word frequency were downloaded into a CSV (comma-separated values) file and uploaded into Excel. Text files were processed using R programming to extract frequency counts for punctuation (not possible with Voyant or AntConc). Results were uploaded into the spreadsheet where calculations determined statistical significance of the punctuation marks by calculating $1.645 \times$ the standard

deviation from the mean, generating a 90% statistical significance level.⁶ Within the spreadsheet, different results were compared and alternative groupings of files were tested and results compared, in a recursive process of analysis and interpretation.

II. PRELIMINARY FINDINGS

Within the field of corpus linguistics, this small study is best seen as a pilot project designed to uncover preliminary findings that must be further investigated to determine their true significance and to contextualize these findings within specific critical debates. The section below provides an overview of what appears to be significant, given both the methods of corpus stylistics and the broad critical context of *AN*. The first section presents findings within *AN*, with particular attention to apparent anomalies. In the second section, *AN* is compared to the reference corpus of selected, representative nonfiction texts written by Dickens, with occasional comparisons to the CLiC reference corpora of Dickens novels or other nineteenth-century fiction. Subsections present findings on punctuation use, sentence length, and the keyness of keywords in comparison across corpora. For this section, *AN* was compared as a whole to the other corpora, but then findings were cross-checked by chapter to pinpoint the exact location of the phenomenon uncovered by computational analysis.

A. Findings within *American Notes* Alone

1. Length, Distance, and Time

AN was published in two volumes, organized loosely by chronological order of the locations Dickens visited (see Table 1, pages 160 and 161 below). Volume I contains eight chapters, from the boarding of the *Britannia* in Liverpool through the description of Washington D.C.; volume II includes ten chapters and begins with the journey from the U.S. capital to Virginia and ends with the voyage home followed by a long chapter on slavery and a final chapter of commentary. A simple numerical analysis of original chapters points to a few anomalies. The most

significant is the "Boston" chapter, which takes up 15% of the whole book and is over twice the length of the next longest chapter (2.7) and over four-and-a-half times longer than the smallest chapter (2.5). Omitting the borrowed material from the Perkins School for the Blind that forms a large portion of the Boston chapter, the remainder is 11% of the book, still the longest chapter. Of course, Dickens deliberately included the Perkins material, and given the length of the excerpts that he includes, clearly this experience was important to him. The story of Laura Bridgman was bound to be appealing to the creator of Little Nell, and, in fact, Dickens' readers were deeply interested in the deaf-blind student at the famous Boston school. As I have written elsewhere, Dickens was impressed with Massachusetts and its progressive institutions, and computational analysis adds further evidence of the extent of his interest.⁷ He was eager to write about the innovations that he observed and devoted a disproportionate amount of the book to the Commonwealth.

Parsing content according to region, unsurprisingly, New England still takes the most space, 23.1% of the book (see Table 2, page 162 below). Yet Dickens only spent 12% of his time on the continent in that region and only 6.1% of the distance traveled within North America. Put another way, the ratio of amount written about New England to the percentage of distance traveled is 3.8 times higher than a one-to-one correlation. The South shares a similarly high ratio of 3.0, and Dickens also spent a relatively small amount of time there, compared to his entire time on the continent, yet he devoted almost 20% of the book to discussing it. The grouping of material for this region includes the "Slavery" chapter (2.9), most of which came from a borrowed source, the abolitionist Theodore Dwight Weld's pamphlet, *American Slavery As It Is*.⁸ Notably, the two regions with the highest number of words-to-distance ratio are also the only two regions to include a significant chunk of borrowed text: Dickens is highlighting institutions in both – enlightened education in New England and chattel slavery in the South.

The Mid-Atlantic region's words-to-distance ratio is closer to 1/1 (0.8), showing more of a balance between amount written and distance

traveled. The Midwest is underrepresented in the text, considering the distance traveled and time spent. In the case of the Midwest, in particular, Dickens writes remarkably little. Notably, he spends one chapter – albeit, the shortest in the whole book – on an overnight trip to the “Looking-glass Prairie,” but vast tracts of land are glossed over in the other Midwestern chapters. The only other time Dickens devotes a full chapter to a short outing is “An American Railroad. Lowell and its Factory System.” These two chapters share a few numerical features that seem notable, despite their contrasting subjects (see Table 1, pages 160 and 161 below). Both are just over 3% of the book length, with similar average words per sentence rates and a nearly identical ratio of the average words per sentence to chapter length, as well as length of chapter title. These unusual similarities point to a possible pattern and prompt consideration of the function of these chapters in the book. Both begin with a description of the mode of travel (railroad, horseback) and then the destination (factory town, wild land), offering carefully constructed thematic representations of an idea, emblematic of their regions: New England progress and Midwestern wildness. The themes of these two exemplars carry on throughout the rest of each region’s material. Like the Midwest, Canada is also underrepresented in the text, given distance and time, at only 5.7% of the total length of the book with a words-to-distance ratio similar to the Midwest. Readers were expecting a book about a republic not a colony, and that is what they got; nevertheless, Dickens made sure to include short but intriguing descriptions of the Canadian destinations that bookended his trip.

Examining patterns of sentence length, two more pairings emerge. Average words per sentence is highest in the two chapters that discuss the ships used to cross the Atlantic Ocean (1.1 and 2.8). Although both of these chapters are short, the length of sentences does not appear to be a function of total words per chapter since other short chapters have far fewer words per sentence averages. Instead, there seems to be something about the subject that lends itself to more complex sentences. On the other end of the spectrum, average words per sentence is lowest in the “Slavery”

and "New York" chapters. These chapters' average sentence length is about half the Atlantic travel chapters. While the sentence length in "Slavery" is undoubtedly impacted by Dickens' use of the runaway slave ads from periodicals that he borrowed from Weld, it is very interesting that the "New York" chapter employs the same significantly shorter sentence style. Are slavery and New York City connected in Dickens' text? An analysis of distinctive words by region may shed some light.

2. Distinctive Words by Region

Voyant generates "distinctive words" lists by comparing each file to the rest of the corpus in order to determine the words that stand out as being associated with a particular file. Analysis of the six region files reveals the geographically distinctive words in *AN* (see Table 3, page 163 below). Some regions are fairly predictable. The Atlantic Ocean is characterized by the seafaring words one would expect, as well as abstract words that reveal attitudes toward life on board a ship. The Canadian material is dominated by images of water and the military, again, not surprising since this section of the book contains the visits to Niagara Falls and Canadian forts, but the list also includes rainbows, angels, brave, and strength – positive terms. New England, too, generates words with mostly positive connotations, many related to education and to industry. Notably, this region generates the most positive abstract terms, including words such as beloved, share, radiant, kissed, enlightened, and intellectual. The Midwest distinctive words are relatively neutral, concrete, and also largely expected, for instance: coach, straw, prairie, Indian(s), log, horses, inn, temperance, corn, and chicken. Only two clearly negative words occur: slimy and desolate. Although Dickens spent more time in the Midwest than other regions, he wrote little about it, and from this list we can imagine why. He must have found the travel tedious and taxing, with little to break up the overwhelming landscape except periodic halts at provincial inns.

The most interesting distinctive regional words are associated with the South and the Mid-Atlantic. The South includes not only expected

words related to slavery, but also many words used to describe the nation's capital: opinion, honourable, public, liberty, and freedom. Although this list may seem similar to New England, the context for these words demonstrates the intense sarcasm associated with their use in context, further highlighting the profound contrast between the two regions. In the Mid-Atlantic, the list emphasizes loss of freedom: cell(s), confinement, prison, and tombs, with abstract negatives such as solitude, wretchedness, and dreadful. However, as gloomy as these words are, the list also includes several neutral words: tambourine, salmon, puddings, and pickles, for instance. Dickens visited many prisons and jails throughout his trip and chose to focus on some of those he saw in New York and Pennsylvania, but Dickens also carries over the language of incarceration to other scenes such as the Shaker village in Lebanon Springs, New York, as well as the description of New York City in chapter 1.6. The preponderance of confinement-related language in the Mid-Atlantic chapters is amplified in the "Slavery" chapter near the end of the book. In fact, due to the order of chapters, Mid-Atlantic always precede Southern chapters. "New York" and "Philadelphia and its Solitary Prison" are followed by "Washington. The Legislature and the President's House." Even "Slavery" is preceded by Dickens' disapproving description of the Shaker community. In the transition from one region to the other, what is bad becomes worse; cells become slaves, solitary confinement becomes shots and scars, and the wretchedness of the inmates becomes the lash of the runaway slave.

B. Findings Across Corpora

A comparative analysis of *AN* and other corpora reinforces the significance of some of the findings outlined above and yields additional insight into what makes the book special within the context of Dickens' other nonfiction, as well as compared to his fiction and nineteenth-century fiction by other authors. Vocabulary density measures yielded no significant results. Density appears to be a function of file size, as expected, with no particular sections deviating significantly from the

norm: larger files have low density and smaller files have high density. Analysis of punctuation frequencies and keyness of key words, on the other hand, yielded the most promising data.

1. Punctuation

An examination of punctuation usage in *AN* compared to the nonfiction reference corpus shows one strongly significant finding and a few instances of moderate potential significance. The most important discovery is the overuse of question marks in the "New York" chapter of *AN* (see Table 4, page 164 below). Adjusted for file size, this chapter has one of the highest uses of question marks out of any of the 115 files in the combined corpora, second only to chapter 8 of *The Uncommercial Traveller* (*UT*). Use of question marks in "New York" is almost two and a half times the statistical deviation, a very strong result. By way of contrast, use of question marks in the "Boston" chapter is only half of one standard deviation below the mean, not a statistically significant low number but definitely on the lower end of question mark usage overall. Compared to "New York," however, this is a marked contrast.

Examining how question marks are used in this chapter, only 19 out of 53 mark direct questions within quoted dialogue; the remainder are a combination of paraphrased questions from verbal exchanges and questions formulated within the narrator's mind, reflecting his opinions. One example will suffice to illustrate. It begins with what appears to be a simple question Dickens asks, followed by the answer his guide to the city provides. Dickens takes those answers, however, and stylizes them in such a way that they are clearly not what the guide actually said but rather a reflection of Dickens' own attitudes:

Are there no amusements? Yes. There is a lecture-room across the way, from which that glare of light proceeds [...]. For the young gentlemen, there is the counting-house, the store, the bar-room: the latter, as you may see through these windows, pretty full. Hark! to the clicking

sound of hammers breaking lumps of ice, and to the cool gurgling of the pounded bits, as, in the process of mixing, they are poured from glass to glass! No amusements? What are these suckers of cigars and swallowers of strong drinks, whose hats and legs we see in every possible variety of twist, doing, but amusing themselves? What are the fifty newspapers, which those precocious urchins are bawling down the street, and which are kept filed within, what are they but amusements? (*AN*, 1.6)

The persistent use of question marks throughout the chapter is remarkable. As many scholars have discussed, New York City is highly questionable for Dickens, but punctuation also appears to add to the effect. Computational analysis of the whole corpus suggests that overuse of question marks is a tool Dickens uses to express disapproval of perceived injustice.

The highest overuse of question marks in Dickens' nonfiction is in "The Great Tasmania's Cargo," in which Dickens visits a Liverpool workhouse where army soldiers who had been unfairly discharged are being cared for in the hospital ward. The treatment and great suffering of these men who had fought for the British in the Sepoy Mutiny in India infuriated Dickens, and a battery of questions assaults the reader. In this piece, however, almost all of the question marks appear within direct quotations, though occasionally questions are reported without being directly quoted, and there is one rhetorical question with no quotation marks in the third paragraph:

Who doubts that if we all did our duty as faithfully as the soldier does his, this world would be a better place?

Dickens was clearly outraged by the injustice, but the use of question marks here does not compare to the satirical questions in "New York." Likewise, in "Poor Mercantile Jack," Dickens again questions inhumane treatment. This piece begins with a battery of internal questions that Dickens identifies as "this was my thought," but then throughout the

rest of the piece, the questions appear within direct quotations, again part of a conversation in which he is questioning people in an institution. In fact, among all of the other strongly significant cases of question mark usage, this punctuation is used as a part of direct dialogue or as simple, direct questions without quotation marks, with few exceptions. Three of the eight pieces also demonstrate statistical significance in use of quotation marks, as well, but comparatively, "New York" has a relatively low usage of quotation marks, 21% below the mean, in fact. While almost all of the cases of overused question marks share thematic similarities, the particular way this punctuation is used is notably different in "New York," indicating that chapter is unique within Dickens' nonfiction.

Analysis of other punctuation marks shows some potentially interesting results, though a more in-depth and systematic approach would need to be taken to determine true significance. In terms of exclamation points, *AN* is not statistically significant overall, but *Pictures from Italy* (*PI*) has three files that demonstrate strong significance. Of the chapters in *AN*, the highest use of exclamation marks is the description of the *Britannia* in the first chapter, with the "New York" chapter the next highest – though neither is out of the realm of normal for Dickens' nonfiction. Chapter 2.3 contains no exclamation points – apparently, the Midwest was too unremarkable to warrant such ejaculations, though that chapter, as well as 2.4 and 2.8, did use a remarkable number of colons. Unlike *Sketches by Boz* (*SB*), which includes many chapters with a statistically significant use of the dash, *AN* is not outside the norm. *PI* leads in usage of commas, quite notably, but again, *AN* is well within Dickens' nonfiction punctuation norms.

2. Keyness

For this project, there were 73 positive keyness keywords and 23 negative keyness keywords for *AN* compared to the reference corpus (see Tables 5 and 6, pages 165-68 below). Findings in this section help demonstrate the ways in which *AN* is both unique and typical.

a. Proper Nouns

Generally, it is expected that important proper nouns in a text will demonstrate notable keyness, since unique names and places are bound to appear more often in a particular text than its reference corpus. In this small study, 20 proper nouns had positive keyness. Unsurprisingly, over one quarter refer to *America* or *American*, and while these words do appear elsewhere in Dickens' nonfiction, they appear much less frequently. The word *states*, in every instance in *AN*, is used to refer either to the United States or to a specific state or grouping of states, while elsewhere the word only refers to America in 3 out of 9 cases. By contrast, *America* appears 57 times in 11 of Dickens' novels, with the obvious instances in *Martin Chuzzlewit* (30), plus 5 *Bleak House*, 5 *Barnaby Rudge* (referencing "where the war is"), 3 *David Copperfield* (2 of which are about characters who escape to America), 1 *Great Expectations* reference to buying a rifle and going to America, 1 *Hard Times* reference to a fugitive, 1 *Little Dorrit* comparison with the snows of America, 5 *Nicholas Nickleby* about starting over or travel, 2 *Our Mutual Friend* comments about America by a pretentious character, 3 in *Tale of Two Cities*, and 1 *Pickwick Papers* reference to emigration (plus additional instances of alternative spellings).

Most of the other proper nouns do not appear at all in the reference corpus. Only two, *Boston* and *Philadelphia*, can be found elsewhere (2 and 1 times, respectively – all in *UT*); still, their appearance at all outside of *AN* seems to signal the importance to Dickens of these particular places. By contrast, Washington was a major city, yet it is never mentioned in the reference corpus. The proper nouns *York* and *England* are unusual in that when preceded by the word *New*, they refer to an American locale, but when used alone, they refer to Britain. *York* in both corpora overwhelmingly refers to New York (38 out of 39 cases). What is more surprising is that only 7 of the 48 instances of *England* appearing in *AN* refer to New England, leaving 41 references to the old country. In these cases, Dickens is sometimes introducing an emigrant from England, referring to an event in England, or making a direct comparison of

America to England. By contrast, *England* appears only 10 times in *PI*, once in *SB*, and 30 times in *UT*. Adjusting for relative frequency per text length, *England* is mentioned over three times more frequently in *AN* than in the reference corpus.

While some argue that Dickens' attitude toward the different regions of the United States was influenced by how much they reminded him of England, keyness analysis does not support this argument. No clear pattern of usage emerges upon close examination; there are negative, neutral, and positive usages of *England* throughout *AN* and within the reference corpus. Nevertheless, Dickens is clearly interested in the interaction between, and comparison of, his home country and the young republic in *AN* in a way that he is not in his other travel book. There, his references to England are mostly about his intermittent travel back home: "bound for a good many places (England among them)" (ch. 6) or mild comparisons such as "what are called Christmas candles in England" (ch. 11). In *AN*, on the other hand, *England* is much more present but in unpredictable ways.

b. Function Words

Function words such as determiners, pronouns, prepositions, conjunctions, and linking verbs are usually included in corpus linguistics stoplists.⁹ Since these words are so common that they tend to skew the results of frequency counts, they are omitted from certain types of analyses. For keyness, however, AntConc searches every word in uploaded files. In this project 12 function words had positive keyness, and 5 had negative keyness.

The most significant findings in both lists are pronouns. In *AN*, the third-person and first-person feminine pronouns *her* and *she* are ranked 9 and 10 for positive keyness due to a large spike in use of these words in the "Boston" chapter description of Laura Bridgman, which creates an uncharacteristic emphasis on female pronouns not found even in Dickens' novels, where only the masculine pronoun *his* has keyness (ranked #23). Undeniably, Bridgman is one of the focal points of *AN*, and although Dickens did not write that material, he did curate and place it deliberately

within a chapter highlighting progressive New England institutions. The travel books do seem to favor third-person point of view, as both *AN* and *PI* have positive keyness for *its* and negative keyness for *you*, *your*, and *my*. Interestingly, the travel books also both have negative keyness for the courtesy titles *Mr.* and *Mrs.*, presumably a function of genre.

The style of *AN* is illuminated by a few other findings. In *AN* *this*, *these*, and *those* have positive keyness, and *the* has negative keyness, showing a tendency toward identifying nouns through demonstratives rather than definitive articles. *The* can be replaced with a demonstrative for many reasons: signaling physical proximity or distance, pointing to that which is difficult to identify, indicating emotional distance, or highlighting something particularly important. The preposition *upon* is ranked #5 in keyness in *AN*, which actually is even more significant because *upon* also has high keyness (#4) in Dickens' fiction compared to the nineteenth-century reference corpus. In other words, Dickens' tendency to overuse *upon* was even stronger in *AN*, but there is no correlation with travel; chapters with the most travel description do not spike in usage of *upon*. The subordinating conjunction *though* has keyness as well. Of the 98 cases, 28 are combined with *as* to mean "as if," and the rest are used alone to mean "in spite of something" – a negation or reversal, perhaps significant in light of scholarly debates.

c. Substantive Words

The largest category of keyness words might be called "substantives," that is, nouns, adjectives, and verbs that are more specific or colorful than common function words. *AN* is a travel book, so it is no surprise that 12 of the words with positive keyness are related to travel, especially by water. While some of these words appear elsewhere in Dickens' nonfiction, their relative frequency in *AN* is significant. In 1842, travelers within North America often utilized canal and river system conveyance since the young country's roads were not well-developed. Water travel was the fastest, most comfortable, and most convenient mode of traveling long distances. The sheer expanse of the

continent and the vast rivers he encountered necessarily pushed rivers to the foreground of his descriptions. After all, he traveled around 4,500 miles, crisscrossing half of the continent from Halifax to Richmond and from the Atlantic coast to St. Louis, Missouri. Rivers were a dominant feature of much of the undeveloped land through which he traveled. Even in the Eastern cities, however, he remarks on the rivers, such as the "light-headed, thoughtless, and brisk [...] young" Merrimack River in the mill town of Lowell, Massachusetts (1.4), or the "wide and rapid" Potomac River (2.1). In the wilder areas, he describes everything from the "filthy" Mississippi River, which after dark "seemed to be alive with monsters" (2.6) to the "beauty of this noble stream," the St. Lawrence River (2.7). In *PI*, most rivers are described merely in passing, and the presence or absence of bridges is noted or picturesque waterfalls. In *SB* the word *river* usually refers to the unnamed Thames, with very few references to *river* overall. The most significant appearance of *river* in the reference corpus is in *UT* in "Chatham Dockyard."

Perhaps the most surprising keyness travel word in *AN* is *railroad*. With a reference corpus spanning Dickens' entire career, *railroad* might be expected to appear more commonly throughout his nonfiction. In fact, there are only half the number of appearances of *railroad* in the reference corpus as in *AN*, even though the reference corpus is three times larger. Although "An American Railroad" uses the word in its title and describes the ride from Boston to Massachusetts, that chapter only accounts for 5 of the word's appearances. About half of the chapters in *AN* include *railroad*, but in the reference corpus it appears in only 7 cases. On the other hand, the word *train* is much more consistently used across all of Dickens' nonfiction and has no keyness in *AN*, either positive or negative. Why is *railroad* significant? Perhaps Dickens' comparatively frequent usage of the term may be explained by the relative infrequency of being able to travel via such means in a continent dominated by water travel and rough roads. By using the term *rail-road*, he may be emphasizing that these are roads. Discussing railroads is also more like describing schools and prisons, with the

focus on an institution or system. The historical context might also shed some light, since he was writing *AN* in the wake of the massive American railroad loan defaults that proved so devastating to British banks and pensioners. In an early draft of the book, Dickens originally planned to include an allusion to this catastrophe in his epigraph, but it was omitted before publication.¹⁰ Railroads were quite topical at the time, so it is natural that the author would report on the state of American railroads.

In addition to travel terms, a few of the important themes of *AN* show up in the keyness word list. One expects that in a book known for its castigation of slavery, the words *negro*, *slaves*, and *slavery* show keyness. Related to this topic, *scar* emerges as significant, with numerous instances of slaves with scars appearing in the “Slavery” chapter. This finding reveals one of the distinctive characteristics of the way that Dickens presents slavery. As noted above, Dickens lifted the excerpts of newspaper ads for runaway slaves directly from Weld’s abolitionist pamphlet without attribution. Dickens’ principle of selection appears somewhat random; however, the keyness of the word *scar* in *AN*, compared to his other nonfiction, may offer a clue. The word *scar* is rare in the reference corpus, only appearing 3 times there. In “The Hospital Patient” from *Sketches by Boz*, the battered woman covers her “scarred and bruised face” with a pillow. In *The Uncommercial Traveller (UT)* there is a reference to a wish to have a villain’s back “scarified often and deep” (ch. 30), and *scar* is mentioned when trying to identify the body of a drowned man in “Shipwreck” (ch. 2). With cases of runaway slaves, as with corpses, scars prove identity and point to crimes. By focusing so heavily on the scars that runaway slaves bore, Dickens chooses to highlight the embodiment of oppression, scars as a sort of text bearing witness to their mistreatment. The “Slavery” chapter in *AN* is by far the most notable place where *scar* appears. Further computational analysis of the “Slavery” chapter in comparison with Weld’s entire work might help determine the hidden principles upon which Dickens determined which excerpts to include in his own chapter.

Also related to the slavery theme is the word *shot*, most frequently appearing in the "Slavery" chapter but also elsewhere in *AN*, although only one of these instances uses the word in a way that refers to firing a gun: an accidental shooting when "guns were discharged [...] after a candidate's successful election," and "the coachman of one of them was actually shot in the body, though not dangerously wounded" (2.7). Of the 7 times *shot* appears in the reference corpus, only one case relates to a gun: an "accomplished and amiable countryman [...] who had shot himself deaf in whole years of competition" (*UT* ch. 7). Other appearances of *shot* in both corpora always mean to jolt forward. In Dickens' nonfiction, by and large, being shot by a gun is an American experience, and in most cases, it is a master shooting a slave. Notably, other words related to slavery such as *whip*, *chain*, *irons*, or *manacle* appear in *AN* but do not have keyness. With *shot*, as with *scar*, keyness results reveal another principle of selection for the Weld quotations. In order to foreground the brutality of slavery, Dickens chose selections that highlighted the brutal violence of guns turned toward human flesh.

The keyness of the word *shot* also points to another theme: American wildness and violence in general. Critics have noted Dickens' critique of American gun violence in *Martin Chuzzlewit*, with its fictional gun-slinging villain, Hannibal Chollop, and other characters who use threats of gun violence to gain social and economic advantage over those whom they intimidate. Yet, in *AN*, computational analysis reveals the extent to which Dickens' "Slavery" chapter makes this case about American violence, as well. The word *gun*, while it does not have keyness, does appear in both *AN* and the reference corpus. The most interesting appearance of the word *gun* in the reference corpus is in "Medicine Men of Civilization" when he describes "the honourable and savage member who has a loaded gun and has grown impatient of debate, plunges out of doors, fires in the air, and returns calm and silent to the Palaver" (*UT*, ch. 28). This "savage" uses a gun to release tension, but he does so by shooting in the air, not at his fellow legislators. Thus, even a so-called savage is less savage than the Americans. In Dickens' novels, the word *gun* appears 49 times, *pistol* 31

times, and *musket* 8 times. In most cases, *guns* are associated with an individual's crime and *muskets* with a comic character, but the many references to *pistol* are associated with the brutal Sikes in *Oliver Twist* (1837-39). While *gun* often denotes an armament with military associations (cannon salute, prison wall sentries, pilot's signal), the word *pistol* is much more frequently associated with personal brutality, and, indeed, 10 out of the 12 times *pistol* appears in *AN* are in the "Slavery" chapter.

Several additional findings may be important. As noted above in the distinctive terms analysis, words related to confinement are significant in certain parts of *AN*. Keyness analysis shows this also to be true of *AN* in relation to the reference corpus. Words such as *cell*, *jail*, and *prison* cluster in certain parts of Dickens' other nonfiction and fiction. As scholars have noted, incarceration is a significant theme throughout much of his oeuvre, but this project shows that *AN* exhibits a statistically significant use of these terms. The thematic term, *public*, shows keyness in *AN* and seems mostly to be a combination of positive uses of the word in New England and sarcastic uses elsewhere. While the connotations are split, the effect is to emphasize the importance to Dickens of an examination of American public life. Finally, two words with keyness seem to signal an emphatic style in *AN*: *very* and *yes*. Closer analysis reveals that *yes* is used almost exclusively within dialogue as an affirmative answer rather than as an interjection, so this finding seems perhaps a bit less significant. The adjective or adverb *very*, however, is used solely to emphasize that something is of a high degree or potency. America is *very*. The "Slavery" chapter is the one anomaly, with *very* appearing far less frequently there. Other means were used in that chapter to convey emphasis.

Negative keyness results are also interesting. At first glance, it might appear significant (though predictable) that *dollars* shows positive keyness. But the word *money* shows negative keyness in *AN*, showing it has been replaced by the word *dollars*. By underutilizing *money* in favor of *dollars*, Dickens highlights a special type of American perspective on

money. One of the most intriguing findings is the negative keyness of the word *family* in *AN*. The book includes only 11 instances, and half of these are in the "Boston" and "Lowell" chapters (1.3, 1.4) with the rest in the Midwest and the journey home – none in the Mid-Atlantic or South and far fewer overall than in the reference corpus. Finally, a predictable pairing emerges of *new* showing positive keyness and *old* showing negative keyness in *AN*. While one would expect this correlation with the Old and New Worlds, the actual uses of the words do not always refer to geography, such as the many cases of old women rather than old cities or buildings.

CONCLUSION

The application of corpus stylistics methods to Dickens' nonfiction can yield important insights into the meaning of texts, Dickens' style, and the place of his nonfiction both within his complete corpus and in relation to other authors of the period. Preliminary findings outlined above indicate there are several viable research questions to pursue in future. What are the stylistic similarities between *AN* and Dickens' fiction, and what do those similarities reveal? How does the presentation of America in *AN* compare to other nineteenth-century nonfiction travel books – is Dickens in keeping with the norm or adding a unique twist? What linguistic commonalities do the portraits of different regions of North America in *AN* share with Dickens' other writing – are there pairings between nonfiction and fiction that we may not recognize without the aid of computational analysis? And my next project: how, on a linguistic level, does Dickens characterize a perceived American propensity to gun violence, materialism, and jingoism? Further, this study applied only some of the basic computational analysis techniques, so there are several methods of corpus linguistics analysis still to be tried, such as an examination of clusters, lexical bundles, collocates, functions, and more. Given the promising nature of this study's preliminary findings, it seems likely that such methods will also yield interesting and significant results.

University of Massachusetts Lowell

Notes

¹ See, for instance, Archibald and Brattin, eds., *Dickens and Massachusetts*; Claybaugh, "Toward a New Transatlanticism"; John, *Dickens and Mass Culture*; Keirstead, "Tourism, Mass Culture"; McBratney, "The Failure of Dickens' Transatlantic Dream"; and Meckier, "Dickens and Tocqueville" and his important earlier work, *Innocent Abroad: Charles Dickens' American Engagements*.

² Mahlberg also uses Project Gutenberg texts for "practical" reasons, recognizing that "the texts will not be perfect, but are sufficient" for corpus linguistics analysis (44). She notes some of the problems with using Project Gutenberg editions, including e-texts being comprised of work done by multiple volunteers (43-44). In the case of this study, however, all the e-editions used were transcribed by the same individual.

³ <http://docs.voyant-tools.org/start/>

⁴ <http://www.laurenceanthony.net/software/antconc/>

⁵ [https://www.birmingham.ac.uk/schools/edacs/departments/english language/research/projects/clic/index.aspx](https://www.birmingham.ac.uk/schools/edacs/departments/english/language/research/projects/clic/index.aspx)

⁶ Statistical significance of word frequencies is automatically checked in AntConc and CLiC, which employ algorithms with appropriate recognition of power distributions, Zipf's law, and other factors affecting the calculation of word frequencies in English.

⁷ See Archibald, *Dickens and Massachusetts*.

⁸ For a discussion of Dickens' plagiarism of Theodore Dwight Weld's abolitionist pamphlet in *AN*, see Johnson, "The Source of the Chapter on Slavery," and Brattin, "Slavery in Dickens' Manuscript."

⁹ See, for instance, the stoplist provided by MALLET (Machine Learning for Language Toolkit), <http://mallet.cs.umass.edu/index.php>.

¹⁰ See Archibald, "Introduction" to *American Notes* (xxi) on Michael Slater's and Sydney Moss's discussion of the title and epigraph.

Works Cited

- Anthony, Laurence. AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University, 2018. Available from <http://www.laurenceanthony.net/software>.
- Archibald, Diana C. "Introduction." *American Notes for General Circulation*. By Charles Dickens. Ed. Archibald. Universitas P, 2018.
- Archibald, Diana C. and Joel J. Brattin, eds. *Dickens and Massachusetts: The Lasting Legacy of the Commonwealth Visits*. Amherst: University of Massachusetts Press, 2015.
- Brattin, Joel J. "Slavery in Dickens' Manuscript of *American Notes for General Circulation*." In Archibald and Brattin, 147-62.
- Claybaugh, Amanda. "Toward a New Transatlanticism: Dickens in the United States." *Victorian Studies*, Vol. 48, No. 3 (Spring 2006): 439-60.
- Dickens, Charles. *American Notes for General Circulation*. Ed. Patricia Ingham. London and New York: Penguin, 2000.
- . *American Notes for General Circulation*. Illustrated by Marcus Stone. The Project Gutenberg eBook. Release date: 18 February 2013 [eBook #675, file first posted on July 18, 1996]; transcribed from the 1913 Chapman & Hall, Ltd. edition by David Price. <http://www.gutenberg.org/files/675/675-0.txt>.
- . *Pictures from Italy*. Illustrated by Marcus Stone. The Project Gutenberg eBook. Release date: 17 February 2013 [eBook #650, file first posted on 17 September 1996]; transcribed from the 1913 Chapman & Hall, Ltd. edition by David Price. <http://www.gutenberg.org/files/650/650-0.txt>.
- . *Sketches by Boz*. The Project Gutenberg eBook. Release date: 6 December 2009 [eBook #882]; transcribed from the 1903 Chapman & Hall edition by David Price. <http://www.gutenberg.org/files/882/882-0.txt>.
- . *The Uncommercial Traveller*. Illustrated by Harry Furniss. The Project Gutenberg eBook, Release date: 4 January 2015 [eBook #914, file first posted on 20 May 1997]; transcribed from the 1905 Chapman & Hall edition by David Price. <http://www.gutenberg.org/files/914/914-0.txt>.

- Hori, Masashiro. *Investigating Dickens' Style: A Collocational Analysis*. Basingstoke: Palgrave Macmillan, 2004.
- Ingham, Patricia. "Introduction." *American Notes for General Circulation*. By Charles Dickens. Ed. Patricia Ingham. London and New York: Penguin, 2000.
- . "The Language of Dickens." *A Companion to Charles Dickens*. Ed. David Paroissien. Oxford: Wiley-Blackwell, 2011. 126-41.
- John, Juliet. *Dickens and Mass Culture*. Oxford: Oxford University Press, 2010.
- . *Dickens' Villains: Melodrama, Character, Popular Culture*. Oxford: Oxford University Press, 2001.
- Johnson, Louise H. "The Source of the Chapter on Slavery in Dickens' *American Notes*." *American Literature*, Vol. 14, No. 4 (January 1943): 427-30.
- Keirstead, Christopher M. "'In Search of the Great Human Family': Tourism, Mass Culture, and the Knowable Community of Dickens' *American Notes*." *Nineteenth-Century Prose*, Vol. 33, No. 1 (Spring 2006), 117-32.
- Mahlberg, Michaela. *Corpus Stylistics and Dickens' Fiction*. New York: Routledge, 2013.
- , et al. "CLiC Dickens: Novel Uses of Concordances for the Integration of Corpus Stylistics and Cognitive Poetics." *Corpora*, Vol. 11, No. 3 (2016), 433-463.
- McBratney, John. "The Failure of Dickens' Transatlantic Dream in *American Notes*." *Nineteenth-Century British Travelers in the New World*. Ed. Christine Devine. New York: Routledge, 2013. 69-86.
- Meckier, Jerome. "Dickens and Tocqueville: Chapter 7 of *American Notes*." *Dickens Studies Annual: Essays on Victorian Fiction*, Vol. 45, No. 1 (2014), 113-23.
- . *Innocent Abroad: Charles Dickens' American Engagements*. Lexington: University of Kentucky Press, 1990.
- Sinclair, Stéfan, Geoffrey Rockwell and the Voyant Tools Team. *Voyant Tools* (web application, 2012). Available from <https://voyant-tools.org/>.

Diana C. Archibald, "A Computational Analysis of Dickens' 'American Notes'"

- Stewart, Garrett. "Dickens and Language." *The Cambridge Companion to Charles Dickens*. Ed. John O. Jordan. Cambridge: Cambridge University Press, 2001. 136-151.
- Tabata, Tomoji. "Investing Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution." *English Corpus Linguistics in Japan*. Ed. Toshio Saito, Junsaku Nakamura, and Shunji Yamazaki. Amsterdam: Rodopi, 2002. 165-182.
- . "Narrative Style and the Frequencies of Very Common Words: A Corpus-based Approach to Dickens' First Person and Third Person Narratives." *English Corpus Studies* 2 (1995), 91-109.
- Tyler, Daniel, ed. *Dickens' Style*. Cambridge: Cambridge University Press, 2013.
- [Weld, Theodore Dwight.] *American Slavery As it Is*. Facsimile edition of the 1839 pamphlet, with a new preface by William Lorenz Katz. New York: Arno Press and *The New York Times*, 1969.



Table 1
Comparative Data on Chapter Length, Words Per Sentence Average Length,
and Title Length in Original Chapters
Part 1 of 2

Volume and Chapter #	Chapter Title	chapter length as % of total book length	chapter length as % of book's avg chapter length
1.1	Going Away	3.2%	58%
1.2	The Passage Out	5.9%	107%
1.3	Boston	15.0%	269%
1.4	An American Railroad. Lowell and its Factory System	3.5%	64%
1.5	Worcester. The Connecticut River. Hartford. New Haven. to New York	3.4%	61%
1.6	New York	6.9%	125%
1.7	Philadelphia, and its Solitary Prison	6.3%	114%
1.8	Washington. The Legislature, and the President's House	6.5%	116%
2.1	A Night Steamer on the Potomac River. A Virginia Road, and a Black Driver. Richmond. Baltimore. The Harrisburg Mail, and a Glimpse of the City. A Canal Boat	6.7%	121%
2.2	Some further Account of the Canal-Boat, its Domestic Economy, and its Passengers. Journey to Pittsburgh across the Alleghany Mountains. Pittsburgh	4.3%	77%
2.3	From Pittsburgh to Cincinnati in a Western Steam-boat. Cincinnati	3.6%	65%
2.4	From Cincinnati to Louisville in another Western Steam-boat; and from Louisville to St. Louis in another. St. Louis	4.6%	82%
2.5	A Jaunt to the Looking-glass Prairie and back	3.2%	58%
2.6	Return to Cincinnati. A Stage-coach Ride from that City to Columbus, and thence to Sandusky. So, by Lake Erie, to the Falls of Niagara	6.8%	122%
2.7	In Canada; Toronto; Kingston; Montreal; Quebec; St. John's. In the United States again; Lebanon; the Shaker Village; and West Point	7.3%	131%
2.8	The Passage Home	3.3%	60%
2.9	Slavery	6.0%	108%
2.10	Concluding Remarks	3.5%	62%

Table 1
Comparative Data on Chapter Length, Words Per Sentence Average Length,
and Title Length in Original Chapters
Part 2 of 2

Volume & Chapter #	Avg # of words per sentence	Chapter's avg # wds per sentence as % of book's avg words per sentence	Ratio of title length to chapter length	Ratio of avg # of words per sentence to chapter length
1.1	43.3	137%	0.35	0.43
1.2	31.1	98%	0.29	1.09
1.3	33.2	105%	0.04	2.56
1.4	33.2	105%	1.28	0.61
1.5	22.1	70%	1.69	0.87
1.6	22.0	70%	0.16	1.80
1.7	26.9	85%	0.45	1.33
1.8	36.3	115%	0.62	1.01
2.1	27.2	86%	2.36	1.41
2.2	32.5	103%	2.77	0.75
2.3	31.3	99%	1.58	0.65
2.4	35.8	113%	2.12	0.72
2.5	29.1	92%	1.60	0.62
2.6	28.5	90%	2/10	1.35
2.7	38.8	123%	1.48	1.07
2.8	40.6	128%	0.51	0.47
2.9	21.5	68%	0.09	1.58
2.10	35.0	111%	0.33	0.56

Table 2
Words to Distance Ratio by Regional Grouping

Region	Length as % of total book length	Region's % of total N. American travel distance	Ratio
Atlantic Ocean	12.0	n/a	n/a
New England	23.1	6.1	3.8
Mid-Atlantic	21.4	25.7	0.8
South	19.9	6.6	3.0
Midwest	18.1	43.9	0.4
Canada	5.7	17.6	0.3

Table 3
Distinctive Words by Region Compared to the Rest of *American Notes*
(Top 50, with place names removed)

Atlantic Crossing: voyage (8), vessel's (6), stewardess (5), pilot (8), whist (4), stewards (4), sickness (4), January (4), captain's (4), ship (32), captain (14), sailor (5), stowed (3), slab (3), portmanteaus (3), included (3), esquire (3), dolphin (3), ding (3), card (3), b'ye (3), astern (3), anchor (3), sea (26), waves (4), sofa (4), sanguine (4), ocean (4), nook (4), gallant (4), cloud (4), bags (4), aboard (4), lazy (9), whirled (2), wave (2), unexpected (2), unconcerned (2), unbelievers (2), tumbled (2), sunrise (2), stumbling (2), squally (2), special (2), skylights (2), sailing (2), rubber (2), plucked (2).
Canada: soldiers (6), rapids (3), raft (3), fortress (3), torrent (2), streaming (2), stirred (2), soldier (2), rides (2), rainbows (2), pipe (2), fort (2), eddied (2), citadel (2), cataracts (2), brave (2), border (2), angels (2), recruit (3), parallel (3), monument (3), leap (3), Catholic (3), Canadian (3), preserved (4), mouth (4), worthless (2), wharfs (2), uncommon (2), strength (2), speech (2), shaking (2), realities (2), quays (2), proof (2), Parliament (2), navy (2), lasting (2), impulse (2), headlong (2), gangway (2), framed (2).
Mid-Atlantic: cell (29), Shaker (15), cells (11), Shakers (6), confinement (6), prison (25), 'em (5), I'm (7), canal (11), thinks (4), they're (4), theatre (4), released (4), mountains (4), Johnny (4), elders (4), solitude (6), dance (6), tables (8), memorable (5), dreadful (5), criminals (5), wretchedness (3), tombs (3), thief (3), termination (3), tambourine (3), starts (3), shad (3), salmon (3), resigned (3), refuge (3), Quaker (3), purchases (3), puddings (3), presided (3), pioneer (3), pickles (3), liver (3), hood (3), grate (3), forester (3), esteemed (3), Easters (3).
Midwest: crocus (17), coach (31), straw (17), brown (15), halted (5), prairie (7), Indians (7), Dr (7), baby (7), log (11), chair (11), varying (4), rivers (4), Boz (4), tavern (6), Indian (6), ha (6), woman (16), inn (9), sir (34), doctor (15), hat (33), horses (14), temperance (5), roots (5), portrait (5), mound (5), corn (5), added (5), uncle (3), tribe (3), team (3), slimy (3), seriously (3), pressure (3), negotiations (3), mustered (3), jaunt (3), desolate (3), chicken (3), banners (3).
New England: Laura (16), blind (20), teacher (10), mills (8), alphabet (7), key (7), fingers (17), letters (16), imitate (6), examined (6), perception (5), manual (5), haven (5), factory (5), evident (5), commenced (5), beloved (5), station (12), weeks (7), pulpit (7), asylum (7), taught (4), spells (4), sightless (4), showed (4), share (4), reflect (4), radiant (4), ma'am (4), manufacturing (4), labels (4), kissed (4), enlightened (4), earnings (4), becomes (4), waiter (6), patient (6), motions (6), mad (6), knowledge (6), intellectual (6), imitation (6), factories (6), evil (6), eagerly (6), classes (6), school (9).
South: slaves (24), slave (18), scar (13), Vinyard (11), slavery (16), Jiddy (9), negro (37), reward (8), James (8), major (7), capitol (7), Arndt (7), driver (17), pistol (6), pill (6), opinion (25), cries (9), named (14), horses (23), sa (5), hi (5), brutal (5), Arpin (5), loose (8), chair (12), public (44), coach (19), legislative (7), hall (7), county (7), the (7), shot (18), servants (4), runaway (4), M'kane (4), loo (4), lee (4), lash (4), honourable (4), Gillespie (4), gaily (4), ally (4) liberty (10), freedom (10), south (6), owner (6).

Table 4
Statistically Significant Use of Question Marks in Dickens' Nonfiction
(n=115, $\bar{x}$ =0.1834, sd=0.1808, strong significance=2 x sd, moderate significance=1.65 x sd)

Text	Chapter	Title	Raw Freq	Words/File	Significance
UT	8	The Great Tasmanian's Cargo	28	3656	0.7659
AN	1.6	New York	53	7144	0.7419
UT	5	Poor Mercantile Jack	34	4803	0.7079
UT	2.4	An Old Stage-coaching House	21	3326	0.6314
SB	Characters 5	The Parlour Order	12	1926	0.6231
UT	21	The Short-timers	24	4656	0.5155
SB	Scenes 23	The Pawnbroker's Shop	15	3001	0.4998
SB	Scenes 16	Ominibuses	8	1651	0.4846

Table 5
Positive Keyness Results
Page 1 of 3

Rank	Freq	Keyness	Effect	Keyword
1	67	148.23	0.0013	America
2	41	116.69	0.0008	Washington
3	39	102.20	0.0008	steamboat
4	46	100.96	0.0009	negro
5	384	93.89	0.0074	upon
6	38	93.37	0.0007	Boston
7	752	83.32	0.0142	which
8	40	75.00	0.0008	cabin
9	419	70.12	0.0080	her
10	275	64.57	0.0053	she
11	62	61.90	0.0012	passengers
12	42	60.03	0.0008	American
13	1062	59.98	0.0199	is
14	32	55.45	0.0006	York
15	34	52.78	0.0007	cell
16	69	52.26	0.0013	river
17	18	51.23	0.0003	crocus
18	17	48.38	0.0003	Laura
19	47	47.35	0.0009	deck
20	16	45.53	0.0003	Cincinnati
21	570	44.74	0.0108	this
22	28	43.60	0.0005	states
23	15	42.69	0.0003	Shaker
24	24	41.59	0.0005	slaves
25	410	38.34	0.0078	very

Continued on next page

Table 5
Positive Keynes Results
Page 2 of 3

Rank	Freq	Keyness	Effect	Keyword
26	13	37.00	0.0003	dollars
27	13	37.00	0.0003	log
28	13	37.00	0.0003	Lowell
29	256	36.79	0.0049	these
30	23	36.72	0.0004	shot
31	93	36.69	0.0018	kind
32	47	34.79	0.0009	England
33	12	34.15	0.0002	Montreal
34	59	33.66	0.0011	boat
35	17	33.13	0.0003	slavery
36	14	33.05	0.0003	cabins
37	24	32.62	0.0005	jail
38	68	32.12	0.0013	board
39	49	31.47	0.0009	yes
40	11	31.30	0.0002	Canada
41	11	31.30	0.0002	fix
42	11	31.30	0.0002	Niagara
43	11	31.30	0.0002	Vinyard
44	23	30.49	0.0004	vessel
45	344	30.43	0.0066	its
46	13	30.34	0.0003	scar
47	16	27.72	0.0003	named
48	572	27.61	0.0109	are
49	22	26.67	0.0004	railroad
50	137	26.42	0.0026	those

Continued on next page

Table 5
Positive Keyness Results
Page 3 of 3

Rank	Freq	Keyness	Effect	Keyword
51	19	25.67	0.0004	labour
52	9	25.61	0.0002	Jiddy
53	9	25.61	0.0002	Pittsburgh
54	98	24.53	0.0019	public
55	31	23.98	0.0006	free
56	12	23.77	0.0002	steamboats
57	27	23.76	0.0005	amount
58	51	23.76	0.0010	pretty
59	44	23.59	0.0009	prison
60	36	23.35	0.0007	ran
61	54	23.06	0.0010	means
62	102	23.02	0.0020	new
63	8	22.77	0.0002	Kingston
64	8	22.77	0.0002	Lebanon
65	8	22.77	0.0002	Sandusky
66	10	22.31	0.0002	Philadelphia
67	19	22.22	0.0004	directly
68	18	21.82	0.0003	canal
69	23	21.48	0.0004	institutions
70	16	21.13	0.0003	growth
71	98	21.08	0.0019	though
72	204	20.97	0.0039	every
73	22	20.92	0.0004	reached

Table 6: Negative Keynes Results

Rank	Freq	Keyness	Effect	Keyword
1	159	86.85	0.0030	you
2	44	80.38	0.0008	Mr
3	1	70.34	0.0000	Mrs
4	281	66.97	0.0054	s
5	121	56.83	0.0023	old
6	6414	56.07	0.0972	the
7	4	41.26	0.0001	shop
8	26	39.23	0.0005	your
9	11	37.09	0.0002	family
10	9	35.73	0.0002	London
11	1	35.67	0.0000	parish
12	16	34.37	0.0003	church
13	60	33.27	0.0012	young
14	163	28.66	0.0031	little
15	8	26.18	0.0002	money
16	262	25.68	0.0050	my
17	14	25.42	0.0003	yard
18	152	24.21	0.0029	man
19	4	22.89	0.0001	jack
20	2	22.58	0.0000	bill
21	7	22.13	0.0001	wine
22	3	21.65	0.0001	d
23	44	21.2	0.0008	street