Computer Analysis of Word Usage in *Emma*

DAVID ANDREW GRAVES
David Andrew Graves is a software architect in California's Silicon Valley. He and his wife Carol Ann are avid readers of Jane Austen and life members of JASNA, and enjoy Regency period dancing.

"Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty-one years in the world with very little to distress or vex her" (5). From the very first sentence in *Emma* the reader can immediately recognize the language and style as that of Jane Austen. Words like "handsome," "clever," "comfortable," and "blessings" are characteristic of Jane Austen's positive view, and even words like "distress" and "vex" appear when describing the absence of negative feelings, or as a fleeting state. Her command of language has made her works some of the brightest stars in all of English literature. While Austen's style, wit, and humor are very much her own, I assert that even the vocabulary she used is recognizable as hers alone.

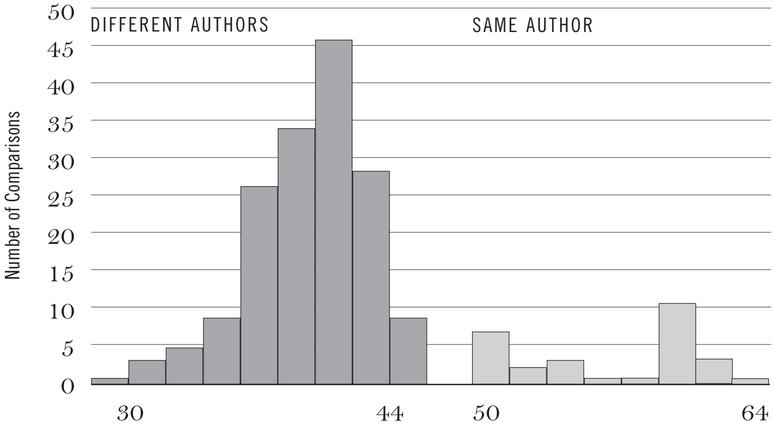
In this paper I discuss the use of computer software to analyze the vocabulary of *Emma*, compared to Austen's other five major novels and texts by several other authors. For the last two years I have been using software as a tool for analyzing texts for patterns in word sequence and word frequency. I have found that works by a given author show similarities in word usage, which can be measured. Using simple arithmetic, a numerical "index of similarity" for the word frequency in any two texts can be calculated. I have found that works by the same author tend to have a higher index of similarity, while works by different authors have a lower index of similarity.

The core methodology is quite simple: the occurrences of each word in the text are counted, and word frequencies are calculated as percentages of the total number of words. For each word that appears in both texts, the two word frequencies are reduced to an index of similarity for that word, then all these values are summed, producing a single numeric value. A text compared to itself using this method would yield an index of similarity of 100. Comparing two texts with no words in common would yield an index of similarity of zero. In my experiments with fifteen texts by five authors, I found that the index of similarity varies from about 50 to 65 when comparing works written by the same author, and about 30 to 45 for works written by different authors.

My initial set of computer-readable texts consists of *Emma*, *Mansfield Park*, *Northanger Abbey*, *Persuasion*, *Pride and Prejudice*, and *Sense and Sensibility*; *Dr. Johnson and Fanny Burney* by Fanny Burney; *David Copperfield*, *Oliver Twist*, and *A Tale of Two Cities* by Charles Dickens; *The Scarlet Letter* and *The House of the Seven Gables* by Nathaniel Hawthorne; and *The Antiquary*, *The Heart of Mid-Lothian*, and *Rob Roy* by Sir Walter Scott. Most well-read human readers can easily distinguish, for example, Austen from Hawthorne. To distinguish works whose authorship might not be obvious to a human reader, ideally experiments should include works by a number of authors from each time period, genre, and subject matter. Thus far, my work has been limited by the availability of computer-readable texts.

Regardless of the size of a writer's vocabulary, it seems that there are certain words that appear in many of their works. Of these words, some tend to appear more frequently, and some consistently appear less frequently. I like to think of this as the "vocabulary fingerprint" for a writer. Different writers have different vocabulary fingerprints, of course. The very common words such

Index of Similarity Distribution



as "the" and "a" cause a certain amount of dilution of each author's distinct vocabulary fingerprint, but I have found that by ignoring these common words, along with all proper names, the fingerprint becomes more distinct, thus increasing the accuracy of the comparison. I constructed my current list of about four hundred ignored words by identifying all the words that are common to all the texts in my sample set of fifteen texts written by different authors in different periods.

Creating a tally of every word used in a novel would be a very tedious task, but a computer makes short work of it. In only a few seconds, a desktop computer can read a novel and create the word list and frequency table that defines the vocabulary fingerprint for that novel. Whenever we read a fine novel, we can easily recognize certain authors without having to compile lists of the words they use. However, my research has been in applying computer technology to measuring similarities between texts. Attributes such as style, irony, and wit are not taken into account, but rather this method takes advantage of the ease with which a computer can be programmed to count occurrences of words, recognize sequences of words, and so on.

I have been asked how I ever thought of such a project. It came to me, almost all at once, while attending a talk at the 1997 JASNA Annual General Meeting in San Francisco. I was listening to Inger Brodey's presentation on Austen's Multiculturalism.

One statement in particular set my mind into a spin. Brodey said, "In fact, words like 'estrangement,' 'imprisonment,' 'alienations,' 'removals,' and 'alone' have an unusual prominence in *Persuasion*" (137). I pondered that this researcher had read the text with an eye for noting recurring instances of a word or class of words. I wondered what it would be like for a computer program to scan a text looking for each and every instance of recurring words. A profile of word usage could be generated, and profiles of different texts could be compared. Perhaps this could give new insight into a particular author's craft.

Soon after, I learned of efforts to collect and preserve rare texts, some of which are of unknown authorship. Some of these texts are from the 17th and 18th centuries, and are in danger of crumbling away. Efforts are in progress to preserve these works by various methods, such as typing the texts into a word processor. Once the texts have been preserved, experts analyze the anonymous works in an attempt to identify the author. I immediately recognized a practical application for my hypothetical computer program. I imagined that works of unknown authorship could be compared to works of known authorship by computer, as an aid to identifying the author of the unknown works. The software could provide an additional tool and additional data for the human experts. I wrote a prototype of my computer program and found that it could, in fact, correctly identify texts that were written by the same author for my sample containing fifteen texts by five authors. Since then I have been refining my method to improve the accuracy of its measurements.

Others have used computers to do analysis of literary texts. At Claremont University, researchers Elliott and Valenza have analyzed the works of Shakespeare using various techniques such as measuring the frequency of hyphenated compound words and relative clauses, grade-level of writing (measured by word-length and sentence-length), and percentage of open-ended and feminine-ended lines. One method used at Claremont was "modal testing" that divides a text into blocks, counts for 52 keywords in each block (middling common words such as "about," "again," and "ways") and measures and ranks the frequency of those words. In contrast, my method tallies the entire vocabulary of the text, with the *exception* of the common words. For my method, I have found that ignoring the commonly used words increases the accuracy of the vocabulary fingerprint, since the common words only dilute the index of similarity. The Claremont approach to modal testing is certainly valid; it's just a different approach. There is more than one way to do textual analysis by computer, and my work is certainly not the first.

So, knowing now how we got here, let's look at the "vocabulary fingerprint" of *Emma* and compare it to Austen's other novels. Keep in mind that we are applying the software tools in a manner for which they were not designed. The tools were originally designed to indicate whether the same author wrote any two texts. While the measurements indicate that the same person wrote the six novels, this is surely not news to anyone. Still, looking at word frequency and repeated word sequences, we can make some interesting comparisons between *Emma* and the other novels.

Let's start by comparing *Emma* to the other novels on the basis of word frequency. The word frequencies in all the novels are similar, but of the five, *Emma* is most similar to *Persuasion*, *Mansfield Park*, and *Pride and Prejudice*. Not as similar is *Sense and Sensibility*, with *Northanger Abbey* holding the distinction of being the most dissimilar, with respect to word frequency. It is easy to guess why—*Northanger Abbey* was Austen's spoof on Gothic horror, which is reflected in the vocabulary of the novel, relating all the anxieties of Catherine Morland, both real and imagined, with words like "agitation," "distress," "fear," and "torment."

Table 1

Word Frequency of *Emma* versus Austen's other Novels

Novels compared	Index of Similarity
Emma, *Persuasion*	62.4
Emma, *Pride and Prejudice*	62.0
Emma, *Mansfield Park*	61.9
Emma, *Sense and Sensibility*	59.7
Emma, *Northanger Abbey*	57.0

At first glance it may seem that these values are very close together, but remember that the range of values for works by the same author is about 50 to 65, a spread of only fifteen points.

Simply knowing that the vocabulary of *Emma* is more similar to *Persuasion*, *Mansfield Park*, and *Pride and Prejudice* is not very satisfying. One immediately wonders *why* these works are more similar. Reading the computer-generated list of most frequently used words reveals a high incidence of words about feelings (both positive and negative), cognition, judgment, discourse, and relationships. These words appear frequently in all of Austen's novels, and make a significant portion of her vocabulary fingerprint. Frequently occurring words were grouped into semantic categories for analysis. Words with meanings spanning multiple semantic categories were omitted, to avoid creating "blurred" results, since the words are simply counted by software that cannot take the context into account.

Each semantic category was rather large and included multiple forms of the same root word. The "positive feelings" semantic category contains words such as "happy," "love," "pleasure," and "affection." The "negative feelings" category includes words such as "anxious," "afraid," "angry," and "sad." The cognition category contains "think," "know," and "understand," while the judgment category contains words such as "good," "better," "opinion," and "judgment." Using these word categories, the frequency of words by semantic category can be measured, revealing some interesting differences between *Emma* and the other novels. For each category, the number of occurrences per 50,000 words was normalized as a percentage of the maximum value for that category. Thus, the top value in each category is listed as 100, and the lower values appear as lesser percentages, as if we were grading exam papers on a curve.

Table 2

Relative Frequency Ranking of Semantic Word Categories

	Negative	Positive	Discourse	Relation	Cognition	Judgment
Emma	82	98	72	93	100	100
Mansfield Park	87	68	76	86	70	99
Northanger Abbey	100	91	72	76	87	64
Persuasion	83	67	81	87	89	90
Pride & Prejudice	89	91	100	100	78	66
Sense & Sensibility	90	100	69	99	70	61

Of the six novels, *Emma* has the lowest incidence of words describing negative feelings, while *Northanger Abbey* has the highest incidence. Thus, the guess made earlier about dissimilarity of these two works is supported by measurement of these word frequencies by category. *Emma* scores in the middle range for the discourse and relationships word category, while *Pride and Prejudice* tops the list for both of these categories.

Emma has the highest incidence of words on judgment and cognition, to be expected in a novel featuring our scheming heroine. *Emma* also has a high incidence of words describing positive feelings, second only to *Sense and Sensibility*. Incidentally, *Sense and Sensibility* seems to be more about sensibilities than sense. While it ranked highest in words describing positive feelings, it ranked lowest for words describing cognition and judgment. It seems that the text devoted to Elinor's thoughtfulness is overshadowed by that covering Marianne's gush of emotions.

Identifying recurring sequences of words is another way to compare any two texts by computer. To do this by hand would be so tedious as to make the task impossible. In this case, the computer program compares every sequence of words in a text to every other sequence, scanning for the recurrence of word sequences of a specified length. I have found that any two texts written by the same author do tend to have similar patterns of recurring words, and texts by different authors tend to have fewer word sequences in common. Thus, word sequence analysis is another form of "fingerprinting" which may be helpful in identifying the authorship of anonymous works.

Examples of some of Austen's most frequently used three-word sequences are "in the world," "she could not," and "a great deal." I found that her repeated three-word sequences were evenly distributed throughout the six novels. From the viewpoint of three-word sequences, *Emma* is just like the other novels.

I tried comparing Jane Austen's twenty most frequently used three-word sequences to the top-twenty for three other authors. Austen and Burney shared "in the world," "would have been," and "not to be." Burney and Scott used "as well as" more frequently than any other sequence, whereas for Austen it ranks tenth. All four authors shared the sequence "would have been." The greatest number of shared three-word sequences (out of their top twenty) was seven, shared by Dickens and Scott.

After much experimentation, I have found that searching for shorter sequences of words gives a stronger indication of whether two works are by the same author. However, it can be interesting to look for longer sequences as well. Austen seldom made use of repeated sequences of eight or more words, but notably, when these repetitions do occur, they are often associated with tedious or undesirable suitors, such as Mr. Elton, Mr. Collins, John Thorpe, and Mr. Rushworth. In each case, the character was either repeating himself, or echoing back what he had just heard from another character. In *Emma*, for example, Mr. Elton echoes Emma's observation that there are "'No husbands and wives in the case at present'" (46). He also repeats himself in some shorter sequences, such as "'an old married man'" and "'my dancing days are over'" (327). In *Pride and Prejudice*, Mr. Collins is paraphrased three times (in narrative) regarding Charlotte naming the day that was to make him "the happiest of men" (122, 128, 139). In *Northanger Abbey*, John Thorpe repeats that he "'did not come to Bath to drive my sisters about'" (48, 99). And in *Mansfield Park*, Mr. Rushworth goes on about his costume for the play, how he will hardly know himself in "'a blue dress and a pink satin cloak'" (138, 139). It is no wonder that Edmund Bertram says to himself, "'If this man had not twelve thousand a year, he would be a very stupid fellow'" (40).

The single most repetitious character in *Emma* is, of course, the incessantly speaking Miss Bates, whose spoken text includes seven percent of the repeated sequences in the novel, while occupying less than four percent of the total volume of text. Close behind her is the most disagreeable Mrs. Elton, whose speech includes five percent of the repeated word sequences in the novel, while she speaks about the same volume of words as Miss Bates. Maple Grove is mentioned thirty-one times in *Emma*, always by Mrs. Elton. Mr. Elton says "'Exactly so'" (42, 44, 46, 48, 144, 370) exactly six times in the novel, and Emma thinks it once: "it will be an 'Exactly so,' as he says himself" (49).

The next steps in my research include further tuning of the software to increase its ability to distinguish between authors with similar vocabulary fingerprints, trying new methods such as identifying words that are *unique* to a particular text rather than those words that are in common, further comparison of spoken text by character, and validation of the methods with a larger number of texts and authors. While there are repositories of computer-readable texts freely downloadable from organizations such as Project Gutenburg, I have had trouble finding texts by particular authors, or from a particular period. For example, I would like to compare Austen's six novels to the juvenilia and texts by her contemporaries, but I have not yet been able to obtain them. The web-based repositories are advancing rapidly, so I'm hopeful that I can soon acquire computer-readable copies of the texts I seek. Ultimately, I hope that my software can be applied to identifying the authorship of anonymous texts, working in conjunction with a preservation project.

I hope no one has been offended by the use of a computer to reduce Austen's writings into sorted lists of words and numbers. We love her novels not *just* because of the vocabulary, but because of her plots and characters, her irony and her wit. My intent has been to use software as a tool to give a new viewpoint on *Emma* and the other novels and thereby perhaps provide some enrichment to your reading experience. I find that now I occasionally savor an individual *word.* From the viewpoint of word frequency by semantic category, *Emma* stands as Jane Austen's lightest and brightest novel, strongly positive, and with the lowest incidence of negative feelings, just as she promised us from the very first sentence.

WORKS CITED

Austen, Jane. *The Novels of Jane Austen.* Ed. R. W. Chapman. 3rd ed. Oxford: OUP, 1933.

Brodey, Inger Sigrun. "Resorting and Consorting with Strangers: Jane Austen's Multiculturalism." *Persuasions* 19 (December 1997): 130-143.

Elliott, Ward and Valenza, Robert. "A Touchstone for the Bard," *Computers and the Humanities* 25 (1991): 199-209.