Provocative Metrics for Assessing Stylistic Quality in Fiction

Abstract

Despite the subjective nature of literary critique and analysis, I posit that that some characteristics of stylistically and aesthetically pleasing writing can be generalized and quantified for the purpose of provoking discussion and further inquiry into style. I focus on rhythmic qualities like pattern and recurrence of phonemes and words, and measures of diversity like unique term counts, repeated N-grams, and turnover in diction. I found that although some metrics demonstrated some potentially interesting trends, few showed consistent differences between well-written and poorly-written works of fiction.

Introduction

I enjoy reading, but I am very much an amateur reader. I read both genre and "literary" fiction, but though I have a nebulous sense of what I consider "good writing," I have always struggled to understand what precisely is meant by the phrase "well-written." There has been considerable work done in the area of machine assessment of writing quality of persuasive non-fiction, mostly for the purpose of grading or assisting the grading of essays on standardized tests[1]. However, very few attempts have been made to provide computer assistance to the assessment of fiction and literature. In one of the few papers to do so, Plaisant et al.[2] describe using text mining and classification interface to help humanities researchers consider the use of erotic language in the correspondence of Emily Dickinson. Borrowing terminology from Plaisant et al., here I attempt to develop provocative metrics for assessing stylistic writing quality in fiction, where "provocative metrics" are quantifiable properties of a passage of text intended to provoke discussion about the stylistic quality of that passage.

I focused strictly on non-semantic stylistic qualities in an attempt to parse the perception of writing quality away from the interpretation of the meaning of a text. To this end, I discussed the issue with two associates with backgrounds in literary analysis (one is pursuing an MFA in creative writing and the other is a production assistant of documentary television), each of which cited the highly subjective nature of stylistic quality, but also noted the importance of rhythm, word choice, and sentence length. These, then, formed the basis of my analysis, which I divided into two categories: rhythm and diversity.

---

[1] Hearst, M. (2000). The debate on automated essay grading. Intelligent Systems and Their Applications, IEEE [see also IEEE Intelligent Systems], 15(5), 22- 37.

[2] Plaisant, C., Rose, J., Yu, B., et al. (2006). Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 141-150.

Methods

Corpora

Examples of poorly written fiction included winning passages from the Bulwer-Lytton Fiction Contest, a competition in which contestants compete to write the most poorly written passage of text[3]. I used the fictional portions of the Brown Corpus in the search for cliches and repeated N-grams, and to provide baseline values for all metrics. Note that the Brown Corpus is diverse and contains multiple authors. Test corpora included excerpts of several novels I have read, usually composed of whole first chapters. These included J.R.R. Tolkein's *Fellowship of the Ring*[4], J.K. Rowling's *Harry Potter and the Prisoner of Azakaban*[5] and *Harry Potter and the Sorcerer's Stone*[6], Kazuo Ishiguro's *Never Let Me Go*[7], Susanna Clark's *Jonathan Strange and Mr Norrell*[8], and David Mitchell's *Black Swan Green*[9] and *Cloud Atlas*[10]. I also included the full text of Cory Doctorow's recent novel *Eastern Standard Tribe*[11], Jane Austen's *Pride and Prejudice*[12], Melville's *Moby Dick*[13], and Hawthorne's *Scarlet Letter*[14] from Project Gutenberg[15].

Rhythm

Alliteration and Consonance

Alliteration is the repetition of sounds at the beginning of words, and consonance is the repetition of consonant sounds anywhere in the words. In this study, I use "consonance" to mean the repetition of *any* phoneme within words. I determined both by first representing all words in a corpus as collections of phonemes using the Carnegie Mellon Pronouncing Dictionary[16]. For words not in the CMU Dictionary, I attempted to break the words into substrings that were. Substrings in the CMU Dictionary were ranked by

---

[3] http://www.bulwer-lytton.com/

[4] http://www.powells.com/biblio?show=MASS%20MARKET:SALE:0345339703:3.98&page=excerpt

[5] http://www.scholastic.com/harrypotter/books/prisoner/chapter.htm

[6] http://www.scholastic.com/harrypotter/books/stone/chapter.htm

[7] http://www.bookbrowse.com/excerpts/index.cfm?book_number=1556

[8] http://www.bookbrowse.com/excerpts/index.cfm?book_number=1463

[9] http://www.bookbrowse.com/excerpts/index.cfm?book_number=1777

[10] http://www.randomhouse.com/catalog/display.pperl?isbn=9780375507250&view=excerpt

[11] http://www.gutenberg.org/etext/17028

[12] http://www.gutenberg.org/etext/1342

[13] http://www.gutenberg.org/etext/15

[14] http://www.gutenberg.org/etext/33

[15] http://www.gutenberg.org

[16] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

length and then by frequency in the Brown corpus. I measured alliteration as the number of words alliterated for a given phoneme within a passage of text given some gap value (4 for all experiments). I measured consonance as the number of times a given phoneme occurred within words spaced at a gap value apart (2 for all experiments). For example, the string “my mawkish hawk” would have an alliteration score of 2 for the “m” in “my” and “mawkish,” and a consonance score of 4 for the “aw” in “mawkish” and “hawk” and the “k” in the same two words.

Rhythm Detection
I detected rhythm by first representing a list of tokens or phonemes as “beats” of “True” or “False” depending on whether the token or phoneme possessed the target rhythmic property (e.g. emphasis, presence of a particular phoneme, etc.). I then created a list of indices of the “True” beats, and traversed the list of beat indices as (PREV, CURR, NEXT) tuples. If the distance from PREV to CURR was equal to the distance from CURR to NEXT (or the the difference was within a given threshold), then all three beats were designated as in rhythm.

I measured rhythm of emphatic phonemes using the emphatic markers on the CMU Dictionary phonemes (phonemes ending with “1” have the primary stress in a word), and rhythm in the presence of individual phonemes themselves.

Diversity
Cliches and Recurring N-Grams
Cliche and redundant, recurring phrases could indicate poor writing quality, so I calculated frequency distributions of N-grams where N was between 4 and 7.

Term Diversity
For all term diversity metrics, I tokenized for words only, and filtered out stop words and punctuation, but not proper names. All tokens were stemmed with the Porter stemmer. I measured word diversity by calculating the mean number of unique terms given different subset lengths of a corpus (e.g. the mean number of unique terms in passages 40 terms in length). I also calculated unique word turnover between adjacent passages of a given length, where turnover was defined as the proportion of unique words in a passage that were not in the previous passage of the same length. Finally, I performed a simple rank-frequency analysis for all unique words in a corpus.

Sentence Length
I calculated sentence length by tokenizing without filtering punctuation or stopwords, and performing simple pattern-based sentence boundary detection.

Results

Rhythm
Alliteration and Consonance
Number of alliterated words increased linearly with the sample size, which was unsurprising (Figure 1). None of the test samples had mean alliteration counts outside one standard deviation of the mean count from the Brown Corpus.

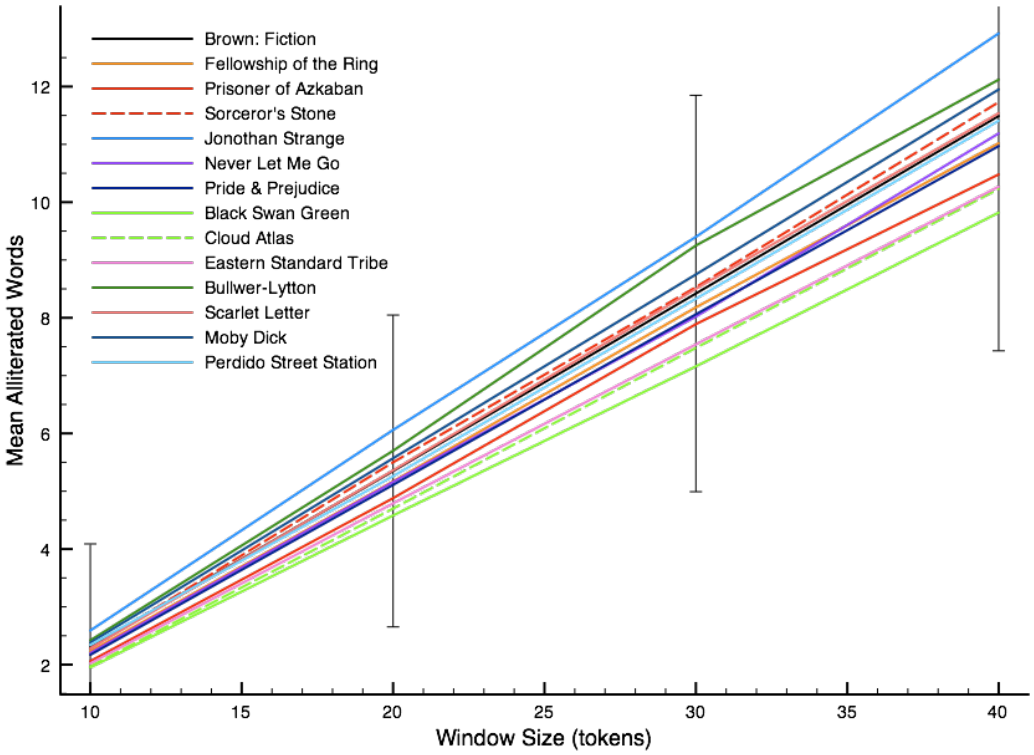

Figure 1. Alliteration. *Mean alliterated words for every phoneme in windows of 10, 20, 30, and 40 words. Error bars show one standard deviation for the Brown Corpus (black).*

Consonance across all phonemes showed a similar pattern to alliteration, with none of the mean values lying beyond one standard deviation from the Brown means (Figure 2).

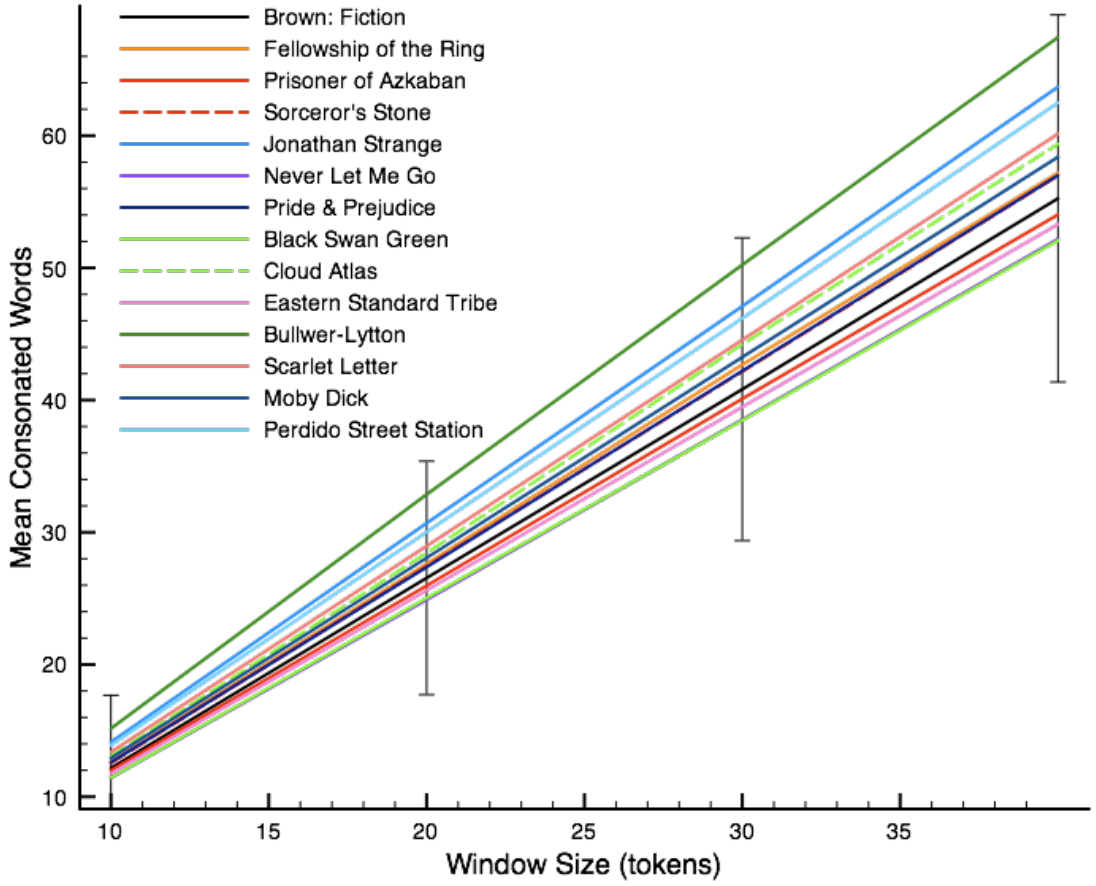

Figure 2. Consonance. *Mean instances of consonance for every phoneme in windows of 10, 20, 30, and 40 words. Error bars show one standard deviation for the Brown Corpus (black).*

Rhythm Detection

Rhythm did not vary greatly across corpora. Emphatic rhythm showed most corpora hewing closely to Brown, with *Black Swan Green* having the most rhythm and the Bulwer-Lytton sentences having the least (Figure 3).

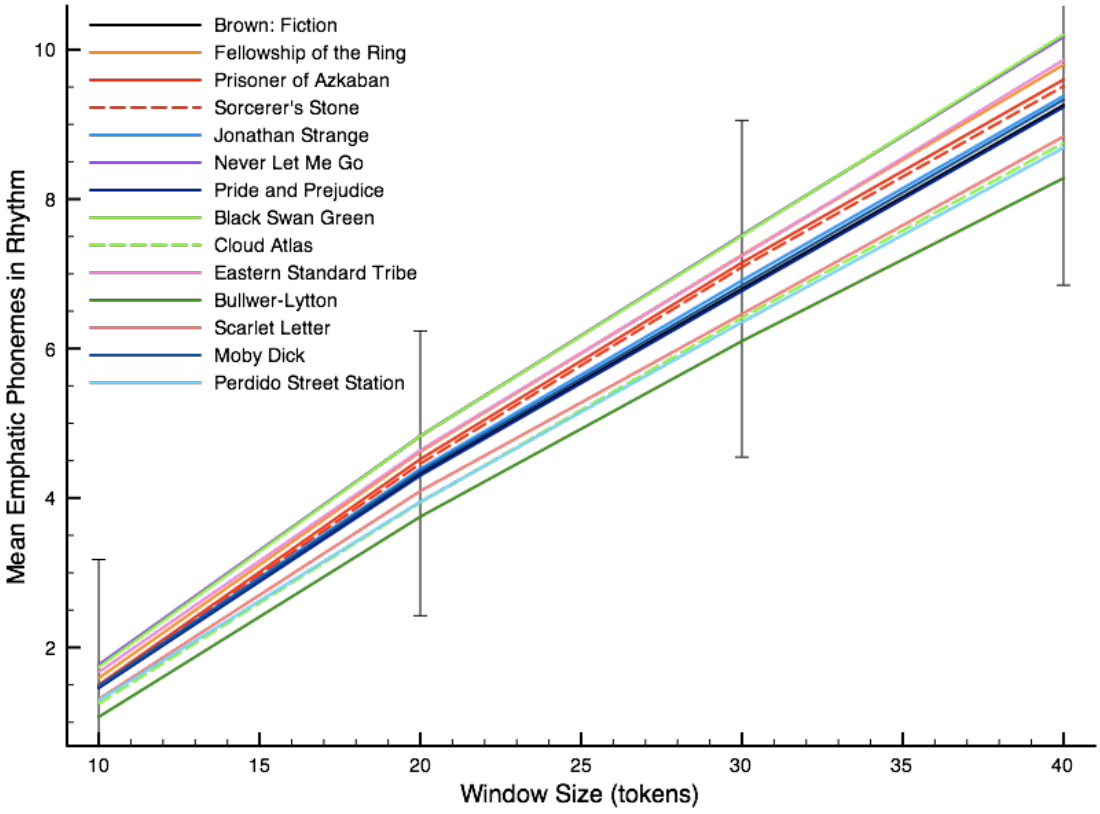

Figure 3. Emphatic Rhythm.

Differences in phonetic rhythm were even less distinctive, as the amount of variation in each corpus dwarfed the mean values (Figure 4). *Never Let Me Go* tended to have lower means than other works.

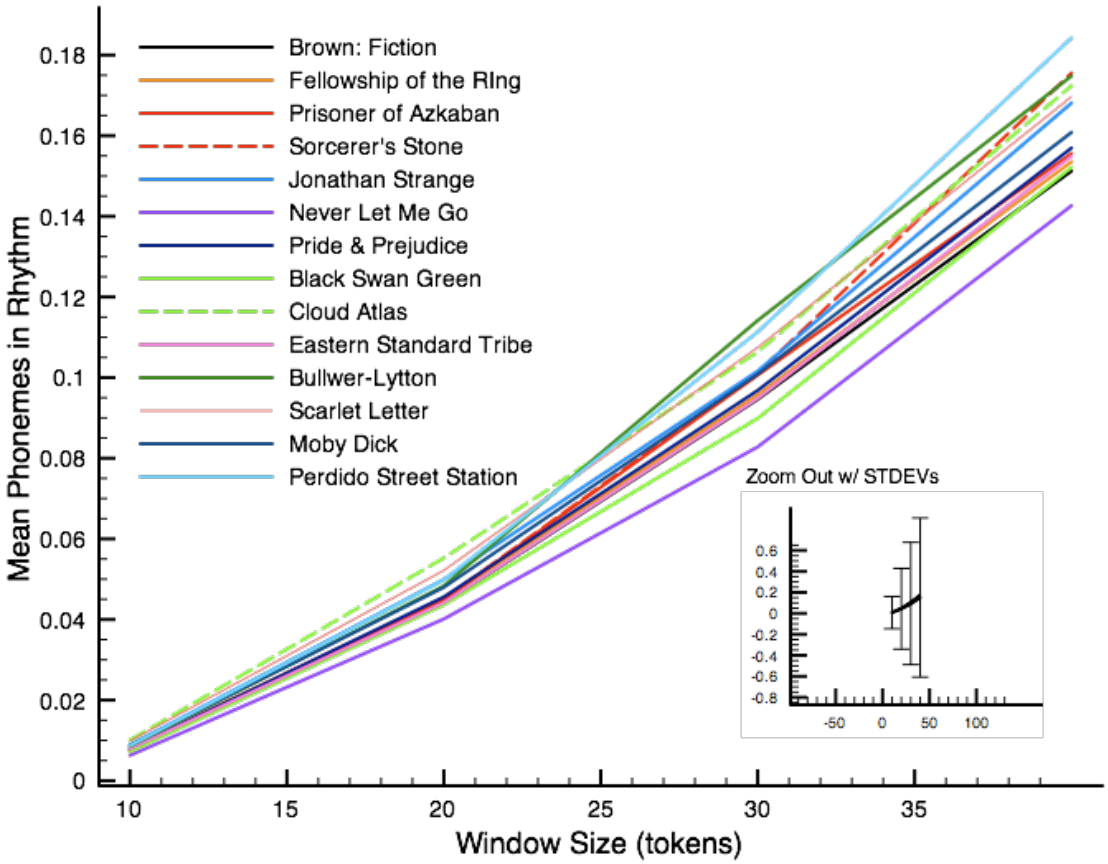

Figure 4. Phonetic Rhythm. *STDEVs in inset are for Brown.*

Diversity

Cliches and Recurring N-Grams

Searching for long N-grams did not reveal any obvious cliches, though it may have shown particular phrasing and turns of speech that could be unique to an author. For the most part, counts of long N-grams were very low, providing little basis for analysis (Table 1).

Table 1. A Sample of Recurring N-Grams. *Proper nouns and numbers omitted, count in parentheses.*

	Fellowship of the Ring	***Pride & Prejudice***	***Black Swan Green***
Top 4-Grams	and the old man (2) can say what you (2) find somewhere where i (2)	i do not know (19) at the same time (16) the rest of the (15)	house in the woods (5) the house in the (4) british bulldogs one two (3)
Top 5-Grams	can say what you like (2) find somewhere where i can (2) half of you half as (2)	as soon as they were (9) in the course of the (7) i am not afraid of (5)	the house in the woods (4) british bulldogs one two three (3) got to go home now (3)
Top 6-Grams	half of you half as well (2) of you half as well as (2) you can say what you like (2)	it was not to be supposed (4) was not to be supposed that (4) herself as well as she could (3)	i've got to go home now (3) a drag on his cigarette that (2) about not going into my office (2)
Top 7-Grams	half of you half as well as (2)	it was not to be supposed that (4) after the health of her family she (2) and i wish with all my heart (2)	a drag on his cigarette that lasted (2) but the person on the other end (2) drag on his cigarette that lasted an (2)

Term Diversity

Term diversity generally followed a Zipf distribution as expected, though this metric was of little use for the shorter, single chapter corpora (Figure 5). Interestingly, *Eastern Standard Tribe* had a curve very similar to the Brown curve, while *Pride and Prejudice* and *Moby Dick* had curves noticeably above and below the Brown curve, respectively.

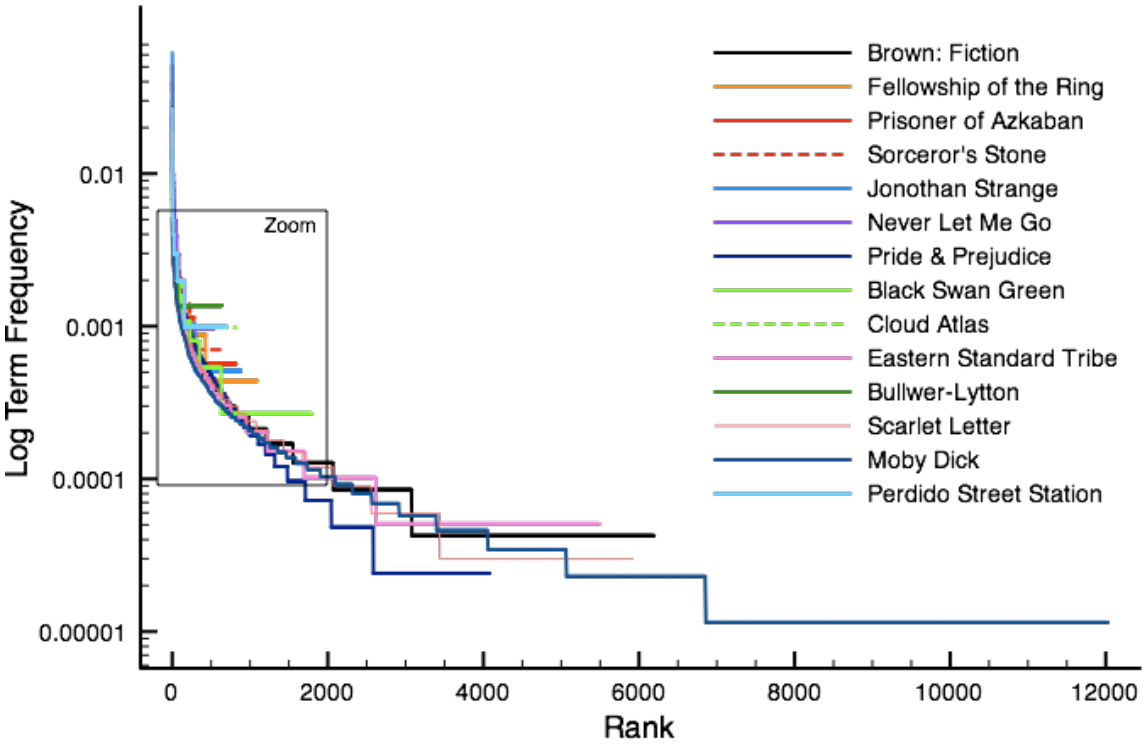

Figure 5. Rank / Frequency distribution of unique terms. *See Figure 6 for a closer look at the middle of the curves.*

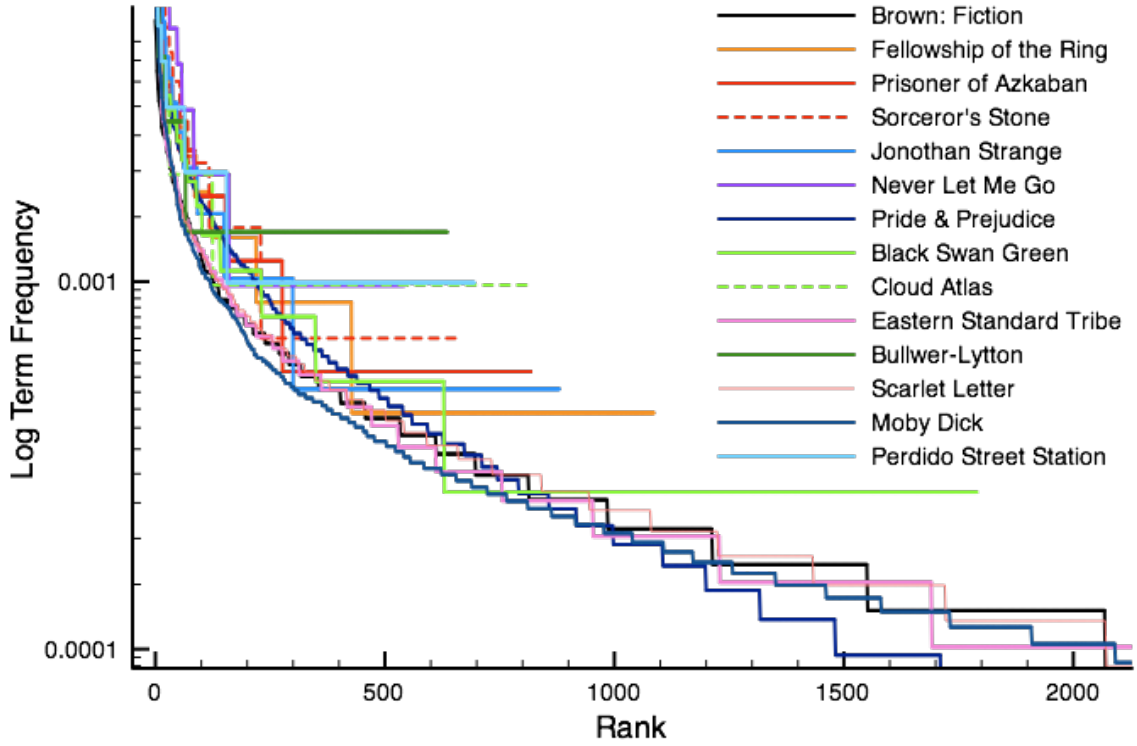

Figure 6. Middle of Rank / Frequency Distributions.

Cloud Atlas had a significantly higher number of unique terms in larger window sizes, and Never Let Me Go and Harry Potter tended to have the lowest (Figure 7).

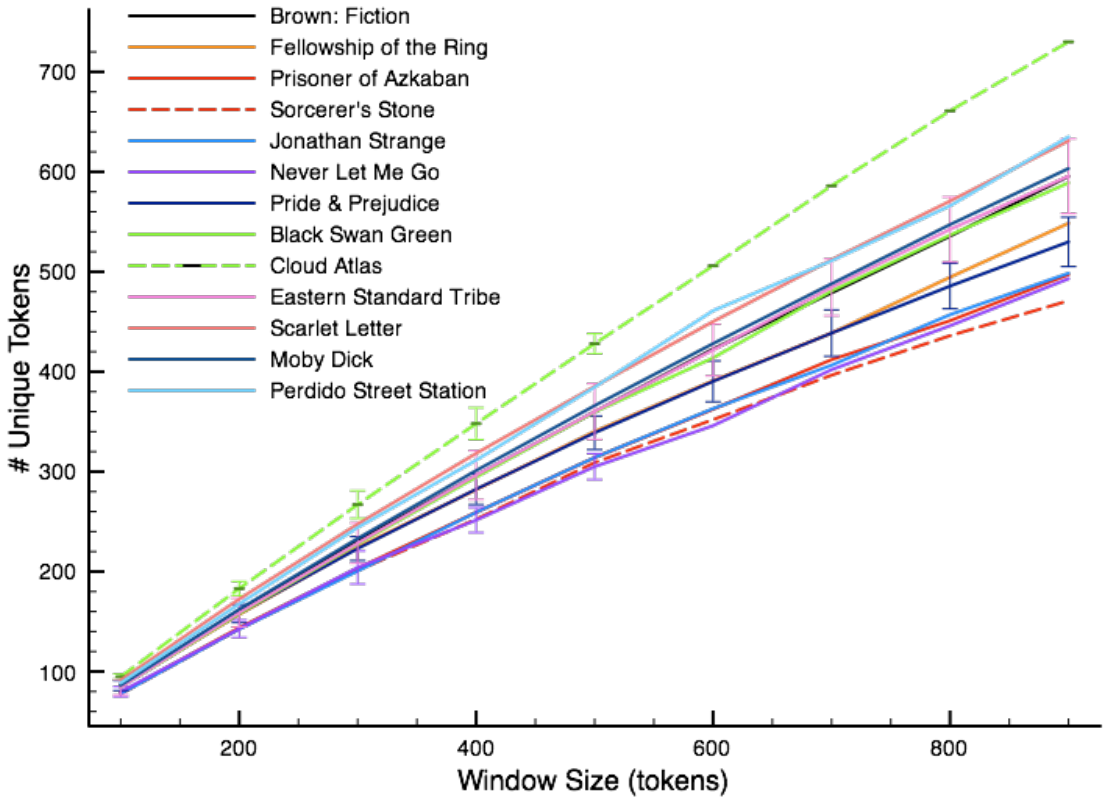

Figure 7. Unique Terms. *Error bars are standard deviations for* Cloud Atlas*,* Eastern Standard Tribe*,* Pride & Prejudice*, and* Never Let Me Go*. STDEVs that appear to be zero artifacts of a bug in the code I was unable to resolve at the time of writing.*

Turnovers between windows were highly variable, but mean turnover at different window sizes tended to decrease at different rates (Figure 8). Brown, *Moby Dick*, and *Eastern Standard Tribe* all experienced very little change in turnover as window size increased.

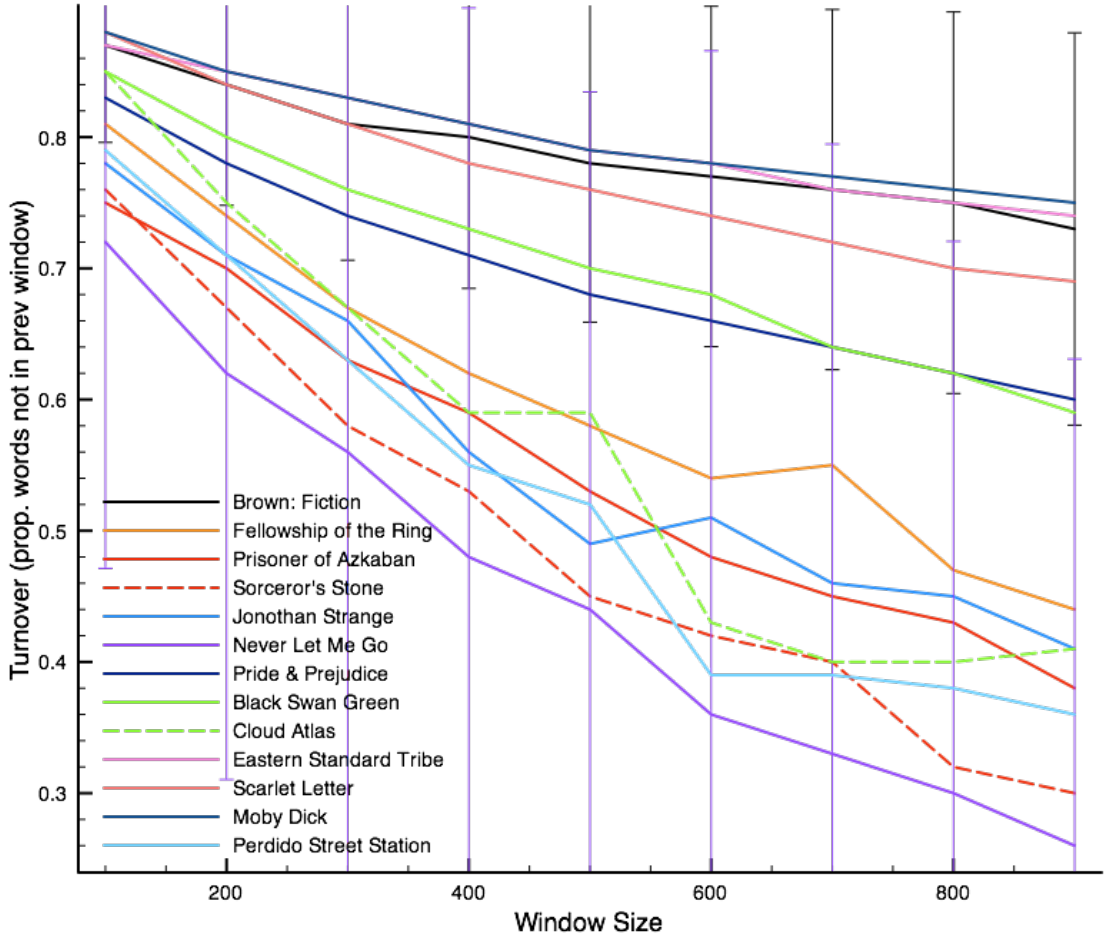

Figure 8. Turnover of Unique Terms. *Error bars are standard deviation for Brown and* Never Let Me Go*.*

Sentence Length

Mean sentence lengths did not vary greatly across corpora, even between full text and single chapter corpora, with the exception of the Bulwer-Lytton sentences, which were longer and more variable in length than any other source (Figure 9).

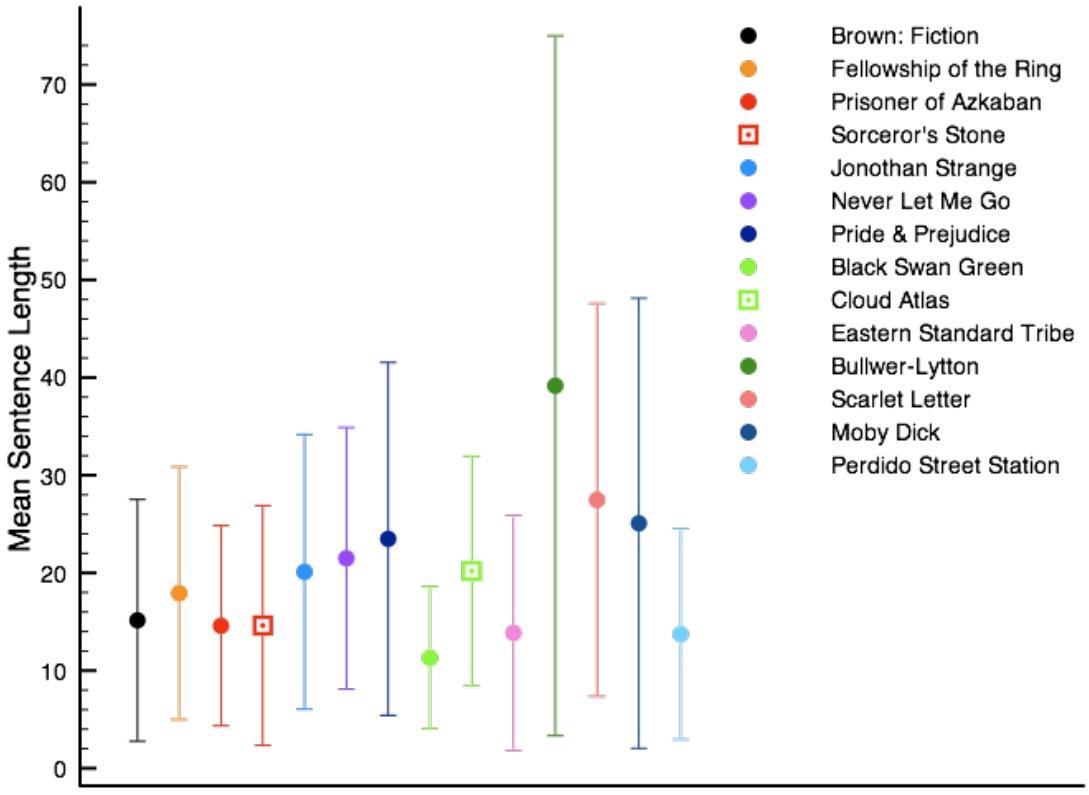

Figure 9. Sentence Length.

Discussion

Fiction is highly variable in almost all respects, and the results shown here demonstrate this. None of the metrics show any consistent pattern across works, or serve to split the works into "well-written" and "poorly written." It was my hope that works that were intentionally bad (Bulwer-Lytton) would show similar values to works that I consider bad (*Perdido Street Station*) or works I consider mediocre (*Eastern Standard Tribe*), but no such patterns seem to exist.

Rhythm

Trends in the rhythmic analysis displayed some potentially interesting results. Bulwer-Lytton sentences, which were intentionally written to be bad, scored highly in both alliteration and consonance, and works that I consider to be very well-written, like *Black Swan Green*, scored low (Figures 1 and 2). However, *Eastern Standard Tribe* and *Perdido Street Station* did not seem extreme or remarkable in either experiment. The high scores for Bulwer-Lytton may result from the idea that poor writing is intentionally florid, with overuse of alliteration and consonance. Overplayed poeticism may be considered trite.

Emphatic rhythm was largely homogenous, but if we consider *Black Swan Green* and Bulwer-Lytton to be significantly different from the rest, we could say that the poor writing of Bulwer-Lytton lacks the rhythm found in excellent works like *Black Swan Green* (Figure 3). However, *Eastern Standard Tribe*'s high scores would seem to negate this. The phonetic rhythm numbers were so noisy that it is doubtful any differences between the mean values are meaningful. This is probably because these numbers are lumped for all phonemes. Perhaps an analysis that looks at each phoneme separately would show clearer patterns, with less variance.

Overall, rhythm metrics failed to show very decisive patterns in the data, and the differences between corpora did not seem great enough to even provoke discussion. However, this analysis was certainly not comprehensive. Perhaps reducing the threshold of the rhythm detection algorithm might have produced results with less variance, or limiting the gap size in the alliteration and consonance algorithms. I also did not consider punctuation and its effect in the rhythm of a sentence, which could be important.

Diversity

The diversity metrics yielded slightly more interesting results. The rank / frequency curves were not very relevant for the short, single-chapter corpora, but they showed some interesting differences between the full-text works. *Pride and Prejudice*'s curve had a higher center, perhaps indicating that Austen had a larger set of words that she used frequently than the other authors (Figure 6). Though it seems unlikely that this difference has much bearing on the quality of the writing (can it be said that Austen was a better writer than Melville?), it is interesting that Austen, a British 19th century author was higher than the contemporary American author Cory Doctorow, who was in turn higher than Melville, an American 19th century author.

Simply plotting number of unique terms actually showed the greatest differences between works, with David Mitchell's *Cloud Atlas* scoring significantly higher than the others (Figure 7). This result may be biased, however, as this sample chapter from *Cloud Atlas* may have a higher number of proper names than other corpora, as it contains a rather lengthy description of the Maori invasion of the Chatham Islands and their subjugation of the native people. Thus, filtering out proper names might have changed this graph considerably. *Harry Potter* and *Never Let Me Go* both had fairly low levels of unique words, and it is tempting to deduce that this results from the former being written for children and young adults and the latter being written in the voice of a young adult recalling her childhood. Again, this says little about the quality of style (some might argue that Ishiguro's style is simple while Rowling's is simplistic), but the theory might be further explored by calculating the metric for more young adult novels and "adult" novels written about children and young adults.

The turnover results seemed provocative at first, but the variance is so great that it is doubtful they are very meaningful. The turnovers in the shorter, single-chapter corpora all fall off at the larger window sizes as window size approaches the size of the corpus itself (Figure 8). Even so, as in the unique words analysis, *Never Let Me Go* and *Harry Potter* are at the bottom with the least amount of turnover, possibly suggesting that not

only do they have less unique words overall, but that they have more consistent word usage. I was at first intrigued by *Eastern Standard Tribe*'s high turnover values and its similarity to the Brown Corpus, because the Brown Corpus is multi-author and thus turnover should be very high. This result might be due to the fact that *Eastern Standard Tribe* often uses colloquial speech, sometimes even text from computer chat rooms, that use language that differs significantly from the rest of the prose. However, *Moby Dick* and *The Scarlet Letter* both show similar curves, so perhaps high turnover is a property of American fiction. More examples might tell. Again, this metric does little to distinguish works I consider well-written from others, but it could be useful for provoking general discussion.

As with the other metrics, sentence length was highly variable, though the degree of variance may say something about quality. The Bulwer-Lytton sentences showed the highest variability by far, and the highest mean, which could indicate the fact that many of them are lengthy run-on sentences with multiple subordinate clauses. Also interesting is that many of the works written in a modern style, like *Black Swan Green*, *Eastern Standard Tribe*, and the Harry Potter books, all show relatively low sentence lengths, while older works like *Pride and Prejudice* and *Moby Dick* are relatively high, perhaps indicating a shift in style over the past 200 years. If this metric is to say anything about stylistic quality, it may be necessary to examine works of the same time period.

Conclusion

Overall, most of these metrics failed to show any conclusive differences between "good" and "bad" writing, although one might argue that some of them could spark discussions about quality. There are many avenues for future work, if only to more conclusively prove that stylistic quality is hard or impossible to pin down with numbers. I could extend the recurring N-grams experiment to include a rank / frequency diagram, and the search for recurring N-grams could include N-grams with gaps (e.g. "she * to him" matching "she said to him") or parts of speech. Rhythm analysis could include other types of rhythm other than a steady beat, like syncopation, and, as previously mentioned, should probably consider the pacing of punctuation in addition to phonemes. All diversity metrics could no doubt be improved with better stemming and proper noun filtering. I had also wanted to try some shallow or full parsing techniques, to measure features like number of clauses, noun phrase length, etc. I suspect the Bulwer-Lytton sentences would appear over-structured with many clauses.