5. Analysis

The linguistic analysis of the results was not directly influenced by stylostatistical research because of its more literary emphasis and the difference in granularity. I simply compiled the significant differences that had been identified by the statistical tests, then examined the samples to find out what qualitative measures of style were correlated with the quantitative differences that had been detected.

To find out whether the differences between POST and the Brill tagger were large enough to lead to different results in the stylistic tests, I carried out the comparisons of tag frequencies on the separate data from each tagger. Since I analyzed the POST results before analyzing the Brill results, in each section I first discuss the POST results and then discuss the Brill results only where these differed from those of POST.

5.1 Results: Matching pairs

I first examined part ones and twos that were either written by the same person, or where the part two was written as a conclusion to someone else's part one (condition C)—that is, those cases where I expected stylistic homogeneity across parts. For each of the tests, there were 20 comparisons: 18 part ones compared with their respective matching parts and two comparisons from condition C.

5.1.1 Two- and three-letter words

In the first chi-square test, I compared texts for the ratio of two- and three-letter words to words of other lengths.

Rankings Nine of the part ones and part twos that matched were ranked in the top five, and fourteen were ranked in the top ten, indicating fairly accurate matching. Neither of the condition C comparisons were in the top ten.

Figure 1 shows a histogram of the results. Each box represents a part two, and its position on the x-axis denotes the level at which it ranked as a match to its part one. For example, a box at 3 represents a part two that was only the third-best match to its true part one. In a perfect match, all boxes would have ranked first. White boxes represent texts written under condition A, that is, written all in one session; dark grey boxes represent those from condition B, that is, written in two sessions a week apart; and light grey boxes represent those from condition C, that is, written after reading a different subject's part one.

Significance levels None of the statistically significant differences in use of two- and three-letter words involved the comparison of a part one and a part two written by the same person. There were also no significant differences between a part one and a part two written by another writer after having read that part one. These findings suggest that the ratio of two- and three-letter words is relatively stable within writers.

5.1.2 Distribution of word length

In the next test, I compared texts for their word-length distribution.

Rankings The rankings made according to word-length distribution were the most accurate. Four part ones were correctly matched to their part two. Altogether, fifteen matches were ranked in the top five. Neither of the samples from condition C (in which the subjects read the first half, then wrote the conclusion) were ranked in the top five. See figure 2.[24]

[24] All of the histograms are located at the end of the thesis.

Significance levels Only one of the statistically significant differences in word-length distribution involved the comparison of a part one and a part two written by the same person. There were no significant differences between a part one and a part two written by another writer after having read that part one.

The matching comparison was just slightly over the significance level. The difference between the parts was in the high use of two-letter words in the first part and in the high use of four-letter words in the second part. The large number of two-letter words in the part one was mainly due to the use of the present perfect (e.g., *…he is put in isolation…*), whereas the simple present was used in the second part (e.g., *They agree*…). The two-letter versus four-letter word difference was also heightened by the high pronominal usage and the different subjects in the two parts: in part one, most of the action revolved around one man (*he*), whereas in part two, most of the action involved groups of people (*they*, *them*).

5.1.3 Tags overall

In the next two tests, I compared texts for the frequency distribution of part-of-speech tags[25] over the complete text. I did this separately for the POST tagging and the Brill tagging.

Rankings The second parts were correctly ranked highest in only two cases. For half of the texts, however, the correct part two was ranked in the top four, and only four matches were not ranked in the top ten. See figures 3 and 4 for the results for the POST tagging and Brill tagging respectively. These histograms show the clustering of the correct part twos near the top of the rankings.

Significance levels Interestingly, although the rankings were the best, the significance levels for the tags overall varied more than for any of the other comparisons. Ten matching comparisons (half of them) showed significant differences (five of these at the .001 level) with the POST tagger. One of these comparisons was from condition C—it was written by a different author after reading part one. Examination of these ten comparisons[26] revealed that only six tag categories were responsible for the differences observed. In eight of the comparisons, the category *noun* (which included nouns and cardinal numbers) showed wide variations; in six, the use of proper nouns diverged; in three, the category *determiner* (determiners and predeterminers) varied; in three, there was a large difference in the use of past-tense verbs; in two, differences in the use of present-tense third-person singular verbs were large; and one showed a discrepancy in adjective use (including comparatives and superlatives).

The most common cause of the differences between part one and part two was that the subjects obviously did not recall the names of all of the main characters while writing the first half, but did during the second half of the experiment. Therefore, characters were generally referred to using a noun phrase in the first half (e.g., *the man*, *the director*), but by name in the second half (e.g., *Charles*, *Mr. Cox*).[27] In two cases, the lack of knowledge was reflected in the usage of determiners and nouns. Four more samples showed large differences in the frequency of proper nouns and nouns for this reason. In the condition C sample, the same problem resulted in variation in the use of proper-noun, noun, and determiner categories. One sample showed a large change in the use of proper nouns only, and one showed a change in the use of nouns only.

---

[25] See Appendix B for a list of the tag categories that were used.

[26] The numbers add up to more than ten because some examples showed more than one type of inconsistency.

[27] A few of the sentences were so ridiculous that this lack of knowledge was obvious (e.g. *Ben recognized one of the children as the young version of the father.*). Since the subjects had been encouraged to take notes during the viewing, and the experimenter had been available for questioning during the writing, I had not anticipated such a problem.

The second most common cause of stylistic inconsistency was change in verb-tense usage. Three samples showed discrepancies in the use of the past-tense verb, and one of these also varied in the use of the present-tense third person singular verb. All of these writers used tense inconsistently, even at times within the same section, a not uncommon problem for writers. Only one of the part twos was written the following week; the other two were written during the same session.

Finally, one sample, which was written in one sitting, had many adjectives in the first half, but very few in the second half. When the sample was checked, one difference was that part one was longer than part two by 29 words (that is, part two was only about nine-tenths of the length of part one). Perhaps the subject was in a hurry to finish the second part, so that he could leave, and therefore was less descriptive. Upon reading the samples, I noticed that part two was more action-oriented, and therefore the modifiers tended to be adverbs, rather than adjectives. This difference might be a reflection of the amount of action in the first half compared to the second half of the video.

Differences between taggers Nine of the ten comparisons discussed above were independent of the tagger used. Both taggers found ten significantly different matching comparisons, but one showed differences only in the Brill-tagged text, and one only with POST. Both of these tagger-independent differences were significant at the .001 level.

The first of these was affected by an error in the preprocessing for the Brill tagger that should, in principle, not have affected statistical significance: several sentences were deleted in the part two. When this incomplete part two was compared with its matching part one by the Brill tagger, there were large differences in several categories, but when the full texts were compared by the POST tagger, there were no significant differences. Two categories accounted for most of the divergence between the Brill and the POST data: plural nouns (NNS) and base-form verbs (VB). In the sentences that were deleted, the POST tagger tagged eight plural nouns, one of which was actually a third-person singular verb (*Charles* <u>*leaves*</u>...). There were more plural nouns in this part of the writing because of what was happening: *the residents* (occurred three times) were trying to sneak down *the stairs* (occurred twice). The POST tagger tagged four base-form verbs in this part, although two of them were actually third-person plural verbs (e.g., *They both* <u>*go*</u> *outside*...). Thus, tagger error on the part of POST also contributed to the taggers' dissimilar results.

The second comparison that showed differences was a result of the accumulation of errors in opposite directions by the two taggers. In the POST comparison, the part one had a lot of common nouns and few proper nouns, whereas the part two had a lot of proper nouns and relatively few common nouns. The counts in the Brill-tagged text were slightly higher for the proper nouns in the part one, and slightly lower for the common nouns in the part two, thus reducing the difference.

5.1.4 Sentence-initial tags

Next, I compared texts for the frequency distribution of the different types of tags in sentence-initial position. I had five categories in the analysis of sentence-initial parts of speech—conjunctions (both coordinate and subordinate), determiners, modifiers (adjectives and adverbs), nominals (nouns, proper nouns, and cardinal numbers), and pronouns (both personal and possessive). Again, I did this separately for the POST and the Brill tagging.

Rankings The rankings of the initial tags were similar to the rankings of the tags overall, although only two-thirds were ranked in the top ten. Aside from one impossible comparison of initial tags, the three lowest ranked tags were the same for each. See figures 5 and 6.

Significance levels Two comparisons of matching parts for both taggers were not calculated, since the chi-square test is not reliable when the observed frequency is zero. The lost comparisons were due to the absence of modifiers in both cases and the absence of conjunctions in one case.

At .001 significance, none of the matching parts, including those part twos that completed a part one written by a different author, were significantly different. At the .05 level, two correct matches showed statistically significant differences from one another. The main differences were found to be in the use of determiners and nouns in both texts. The discrepancy was again primarily due to the fact that the subjects did not know all of the main characters' names while writing the first half, but did while writing the second half. One of these samples was also found to have significant differences in the usage of nouns and determiners in the whole text (see above).

Differences between taggers The only differences between the taggers involved impossible comparisons. Both taggers had two impossible comparisons, one of which was independent of the tagger used.

Due to an error by the Brill tagger, the sole modifier in sentence-initial position in a part one was not tagged as a modifier, and its matching part two had no modifiers. Therefore, a comparison that was possible, but not significant, for POST could not be made.

In the Brill data, one difference in the tagging of a part two allowed a comparison that was not possible by POST. The Brill tagger tagged *next* as an adjective, whereas POST tagged it as a preposition (*<u>Next</u> we see Charles waking up Ben…*)[28], thus allowing the calculation of the modifier category. The comparison did not show significant differences between the parts.

5.1.5 Sentence-final tags

Next, I compared texts for the frequency distribution of the different types of tags in sentence-final position. There were four categories in the analysis of sentence-final parts of speech—modifiers (adjectives and adverbs), nominals (nouns, proper nouns, and cardinal numbers), pronouns (both personal and possessive), and verbs and particles.

Rankings The rankings were more spread out for the sentence-final tags than for any of the other rankings (see figures 7 and 8). Also, the texts that were ranked low were different from the texts that were ranked low in the initial-tag and tags-overall data. This finding suggests that the use of sentence-final tags is distinctive from the use of sentence-initial tags or all tags. However, further investigation is required before any firm conclusions can be drawn.

Significance levels There was one matching pair for which no comparison could be made because the expected value of two tags was zero: neither part had occurrences of modifiers or pronouns in sentence-final position. As mentioned above, an impossible comparison suggests homogeneity rather than heterogeneity. There were no matching comparisons that showed significant differences between part ones and part twos at the .001 level or at the .05 level. The lack of variability among the samples might have contributed to the data spread in the rankings: since the differences in the values of the chi-squares between the first and last ranked part twos were not large, the rankings might not be meaningful.

There were no differences between the taggers.

5.1.6 Comparison of means: Sentence length

One matching pair showed significant differences. The average sentence lengths of the two parts were fairly different, but when the standard deviations were taken into account, the difference was much greater. The first part had more adjectival and adverbial clauses, and many more compound and complex sentences than the second part. One reason for the difference may have been that the

[28] I would tag it as an adverb.

writer included more description in the first half to set the scene of the story. The second part was shorter than the first part by more than 200 words, so another likely reason for the difference was that the writer was in a hurry to finish the second part so that he could leave.

5.1.7 Differences between conditions

Since there were only two writing samples in condition C, it is impossible to draw even tentative conclusions about whether reading another person's document, and then adding to it, influences stylistic choices. Comparison of the rankings and chi-square values for conditions A and B did not reveal any pattern of differences (see histograms). Overall, second halves that were written a week later did not reveal any more inconsistencies than did second halves written immediately after the first half. Perhaps a longer intervening period of time would affect a writer's style. However, the absence of a time effect is predicted by stylometric theory, which holds that writing style is stable in mature writers.

5.2 Results: Non-matching pairs

I next examined paired part ones and part twos that were not written by the same person—those cases in which I expected to see stylistic differences, or heterogeneity, between parts. For each of the tests, there were 340 possible comparisons (16 part ones compared with each of 19 non-matching part twos, and the two part ones that were used in condition C compared with each of 18 non-matching part twos).

5.2.1 Two- and three-letter words

Fifty-three (out of 340) of the chi-squares indicated significant differences at the .05 level; only three were significantly different at the .001 level.

Significant differences in the ratios of two- and three-letter words to words of other lengths were caused by a variety of factors in each case. The use of certain parts of speech (coordinate conjunctions, pronouns, and prepositions) was associated with high percentages of two- and three-letter words, whereas adjectives occurred more often in samples with lower two- and three-letter word ratios. Perfect tenses (e.g., *is going* rather than *goes*; *was talking* rather than *talked*) were also associated with a high percentage of two- and three-letter words. Overall, samples with a high ratio of two- and three-letter words tended to have short sentences, many prepositional phrases, and vocabulary that was simple (e.g., *sad* rather than *depressed*; *old* rather than *elderly*) and colloquial (e.g., *kid* rather than *child* or *youth*; *bad guy* rather than some other descriptor for the protagonist's ornery best friend). In general, then, this test appears to differentiate between a simple and a more descriptive style.

5.2.2 Distribution of word length

Out of 340 possible comparisons, there were 102 that were significantly different at the .05 level; 33 were different at the .001 level.[29]

One-letter words Only one sample had a high percentage of one-letter words, but the one-letter words made a difference in nine comparisons. The high one-letter word count was due to the use of the indefinite article *a*. Not surprisingly, the sample with many one-letter words was a part one since writers had to introduce the characters, etc. in their first parts, and therefore tended to use more indefinite articles than in their second parts. In the part twos, most characters, places and objects had already been mentioned, so the indefinite article was not often needed. The sample with many indefinite articles had a higher percentage of them than other part ones for two reasons. First, it contained a great deal of description, and introduced many people and objects that were

[29] The number of significant comparisons listed below add up to more than 102 because some examples showed more than one type of inconsistency.

not mentioned in other samples. Secondly, the writers of the other part ones tended to use previously-mentioned characters and objects to introduce new characters and objects. Compare, for example, the first mention of the main character and one of the minor characters, after only the setting has been described: *A smartly-dressed old man…proudly tells a nurse that his son is coming for him.* versus *The house nurse notices one of the residents on the staircase: it's Charles….* The high use of the indefinite article, then, is not entirely due to the situation of writing an introduction. To some extent, how a writer introduces new information is a stylistic choice.

Two-letter words Twenty-six comparisons showed differences caused by two-letter words. The samples with a high number of two-letter words had many two-letter pronouns (*he, it, we, me*) and had many instances of the third-person singular form of *to be* in both the simple present (e.g., *He is old.*), and the present progressive (e.g., *is trying*), whereas the samples with few two-letter words did not.

Three-letter words The most common word-length difference was in three-letter words—there were 67 significantly different comparisons. In all cases, there were a number of factors that contributed to high and to low three-letter word counts, although not all of these factors contributed to all differences. Overall, samples with a high percentage of three-letter words had many pronouns, certain verb tenses and few adjectives. The pronoun category contributed to the three-letter word counts because of the high number of three-letter pronouns used (primarily *his* and *him*). The verb tenses that were common in the samples with many three-letter words were the perfect (e.g., *has found*), the past perfect (e.g., *had realized*), the past progressive, (e.g., *had been doing*) and the use of the third-person singular verb *to be* in the simple past (*was*). There were few adjectives in almost all the samples with many three-letter words and the ones that were used tended to be simple (e.g., *old, bad, sad*).

Coordinate conjunctions and prepositions did not contribute a great deal to either two- or three-letter word distributions, although they did to the combined two- and three-letter words test. Since there are both two- and three-letter coordinate conjunctions and prepositions, the reason for the difference seems to be that those categories only make a notable contribution when the two- and three-letter word counts are combined.

Four-letter words Variations in the usage of four-letter words contributed to 59 significant differences. Most of the samples with few four-letter words were part ones, whereas most of the samples with many four-letter words were part twos. Most of the samples involved in the significant four-letter word comparisons had a relatively high number of pronouns, and therefore the difference resulted mainly from writers using *they/them* instead of *he/him*. Just as in the matched comparison that showed significant differences, discussed in section 5.1.2, the divergence was chiefly due to the fact that the action revolved primarily around the main character in the first half, but focused on a group in the second half. The part one that had many four-letter words also had many *they*'s and *them*'s because the writer used these pronouns in a generic way (e.g., *Some people do stay young—do they have a secret?*), as well as to refer to the groups of people who appeared in the first part. In two of the three part twos that had few four-letter words, there were not many *they*'s and *them*'s. In one case, the writer often used *their* (e.g., *Their escape…*; *In their plan…*) and in both samples noun phrases and names were preferred to pronouns. In the third case, *they* and *them* were often used, but several four-letter words that were common in most other samples were not: the word *home* was never used (to refer to the old-age home where the action took place), and neither were *life* or *live*, which are theme-related words.

Five-letter words One sample with many five-letter words contributed to eleven significantly different comparisons and another such sample contributed to an additional three. The five-letter words in these samples did not fall into particular categories. Both samples had more than one of:

their, *young*, and *other*, but these were common words in many other samples. The two samples also had many four-letter verbs in the simple present (e.g., *tells*), but so did one of the samples with few five-letter words. The first sample also had more than one instance of: *again*, *would*, *being*, and *while*; the second had more than one occurrence of: *there*, *about*, *after*, and *magic*. Other than the fact that each writer used a lot of five-letter words, there does not seem to be any systematic reason for the high number, nor is there any obvious difference between samples with many and samples with few five-letter words.

Six-letter words There were only nine comparisons that showed significant differences due to six-letter words. Six of these involved a part one that had many six-letter words. Upon examining this sample I found that two of the main characters' names were never used by the writer, but rather one was referred to throughout as *friend*, and the other was continually referred to as *doctor*. Although most writers used the term *friend* to refer to the main character's best friend, there were several different terms that were used to refer to the man who ran the institution (e.g., *director*, *psychiatrist*). Also, the writer of the sample with many six-letter words used both terms many times, whereas most writers who used them only did so once or twice. The other three comparisons involved a part two that had few six-letter words. There was no apparent reason for the low number. One of the three part ones with many six-letter words used the term *father* throughout the sample to refer to the main character, but there were no six-letter words that were used more than a few times in the other two part ones.

Seven-letter words There were 28 comparisons in which seven-letter words made a difference. The differences were primarily a result of the fact that the main character's first name, nickname, and last name all had seven letters (*Charles* (*Charlie*) *Whitely*). In every sample that had many seven-letter words, the main character was frequently referred to by name (in fact, all of these samples had high proper-name counts), whereas in the samples that had few seven-letter words, his name was never used in any form.

Eight-letter words Eight-letter words were involved in differences in nineteen comparisons. High numbers of eight-letter words were associated with certain vocabulary related to the plot that was used frequently by the writers of these samples, but not much by other writers. In most of the latter cases, other terms that were not eight-letters long were used to refer to the same people or concepts.

The most common eight-letter word in these samples was *children*. Children appeared as characters and were central to the story thematically, so they were referred to both in descriptions of the action, and in discussions of the motivation for the plot. Writers who did not use *children* often, used various other terms to refer to them, such as: *youth*, *kids*, or *boys*.

In the show, the main character talked about the *Fountain of Youth*. Not all writers used this term in reference to the theme of rejuvenation, but those who did often used it several times, thus contributing to the eight-letter word-count.

Both *roommate* and *director* were often used in these samples to refer to two of the main characters. Writers who knew the names of these characters rarely used these terms, and other writers used different terms to refer to the same characters (e.g., *friend* and *doctor*).

Finally, since the program was a *Twilight Zone* episode, references to the show increased the eight-letter word count; not all writers referred to it and most did so only once.

Nine-letter words Differences in the use of nine-letter words occurred in nine instances, involving two part ones with many nine-letter words four times each. In both of these writing samples, certain nine-letter words were used several times: in one, *residence* was used four times (to refer to the old-age home), and *geriatric* was used twice, and in the other, *questions* was used three times and *residents* was used twice. Since nine-letter words are not very common, the use of these words in addition to a number of other nine-letter words was enough to make a difference in some comparisons.

Words with more than nine-letters None of the significantly different comparisons had a notable amount of their difference caused by ten-letter, eleven-letter, or twelve-or-more letter words. Since there were not many words of these lengths, it is not surprising that they did not contribute much to significantly different comparisons. In other types of writing (e.g., technical writing), longer words might be more common and would therefore be more likely to contribute to the differences.

Conclusion Overall, significant differences caused by short words (one to four letters) indicated differences in the use of certain grammatical categories (e.g., rate of indefinite articles), whereas differences caused by longer words (six to nine letters) indicated differences in vocabulary preferences (e.g., *friend* versus *roommate* to refer to one of the main characters). Examination of samples with many five-letter words suggest that high numbers of five-letter words result from both grammatical and vocabulary preferences, which is perhaps why I was not able to fully account for the significant differences caused by them.

The differences caused by long words (primarily eight- and nine-letter words in these samples) were also associated with the level of vocabulary used. Samples with many long words tended to have a more elevated vocabulary (e.g., *geriatric*, *residence*, *children*), whereas those that did not had less formal terms (e.g., *old*, *home*, *kids*). Differences that result from words of more than nine-letters would probably be similar to differences that result from nine-letter words: they would be due to word choice and vocabulary level. When samples with a significantly high number of short words were compared to samples with a significantly high number of long words, differences in vocabulary level were not apparent to the reader unless the samples with many short words also had very few long words. In other words, vocabulary level was associated with whether long words were present or absent, and the rate of short words did not seem to be an independent contributing factor.

5.2.3 Tags overall

One comparison could not be made due to an observed frequency of 0 for the category *other punctuation*. Of the 339 possible comparisons, 277 pairs were significantly different at the .05 level; 198 were significantly different at the .001 level. According to stylometric theory, writing samples by different people will often be significantly different, so these high numbers are not surprising.

Due to the extremely high number of comparisons that showed significant results, a subset was selected from the entire set to be examined for general trends in the data.

Analysis revealed certain clusters of differences between the pairs, some of which were observed in the analyses of matching pairs (section 5.1.3). The main patterns of differences were found in nominal preference, verb-tense preference, adverb and adjective use, rate of personal pronouns, number of coordinate conjunctions, use of prepositions, and the frequency of sentence-final punctuation (i.e., periods, question marks, and exclamation marks), commas and other punctuation.

The frequencies of certain tags (possessive endings, past participles, *wh*-words, modals, particles, and subordinate conjunctions) were relatively stable and their usage never varied greatly from sample to sample. Other tags (possessive pronouns, present participles/gerunds, infinitive markers, and non-third-person singular present verbs) occasionally varied a moderate amount from one sample to the next, but were never large-enough contributing factors to use to distinguish between writing samples. Perhaps these parts of speech are generally stable in this type of factual retelling (particularly modals and *wh*-words, which probably vary more in persuasive texts, for example), or they may be generally stable in most texts. Another possible reason for the lack of distinctiveness of certain tags may simply be that they did not occur often enough relative to other tags to make an appreciable difference.

Nominals As discussed in section 5.1.3, the differences in nominal preference arose mainly from the avoidance of character's names by writers who had forgotten them. Of the twenty part twos, fourteen of them included all three of the main characters' names, and all of them had at least one of those names, whereas, of the eighteen part ones, only one included all three names, and four of them had none of them. It is therefore not surprising that contrasts in nominal preference were common.

Verb tense Inconsistency in verb tense has also been discussed in section 5.1.3. Most writers did use tense consistently, but since tense is to some extent a matter of choice, the preferred tense wasn't the same in each sample, nor would it be in most collections of writing samples. Therefore, there was a higher percentage of verb-tense inconsistencies in the non-matching pairs than in the matching pairs.

Adjectives and adverbs Samples with high adjective counts usually had a ratio of adjectives to nouns that was close to 1:2, whereas those with low counts had ratios close to 1:4. In the former samples, important objects and people in the story were almost always modified the first time they were mentioned, and often were modified again and again. Samples with low adjective counts tended to have adjectives that were related to the theme of the narrative, (e.g., the *old/young* dichotomy was prevalent in most samples), but did not have many other adjectives. The other adjectives in the "low" samples were mainly everyday, and often over-used, words such as: *good*, *bad*, *new*, *large*, *sad*. In the "high" samples, more unusual adjectives were prevalent: *grand*, *sleek*, *ornate*, *fatalistic*, *curmudgeonly*.

Samples with high percentages of adverbs provided information about time, location, manner, intensity, doubt, and cause. Many adverbs of negation were also used. In fact, in these samples a wide variety of adverbs in their various roles were used. Samples with low adverb counts, however, usually did not show adverbs in many different roles. Rather, the adverbs in such samples tended to only be of two or three types, although not the same ones for all writers. For example, some writers provided a lot of information about time, whereas others focused on location or manner.

Samples with many adjectives and few adverbs did not contrast strongly with those that had many adverbs and few adjectives. In most cases, samples with higher adverb counts were part twos, while samples with higher adjective counts were part ones. The writing task itself contributed to this phenomenon: in the first half, writers needed to "set the scene" by describing the characters and the setting, as happened in the video (and indeed in most narrative) which resulted in more adjectives; the second part of the video, on the other hand, was more action-oriented, which resulted in more adverbs.

There were a few samples that had low percentages of both kinds of modifiers and a few that had high percentages of each. These samples stood out from the rest. The high or low use of modifiers seemed to be due to the individual writers' preferences, rather than something that was influenced by the writing task. Perhaps a combined count of both modifiers would provide more useful stylistic information than separate adverb and adjective counts, since the type of modifier used appears to be more heavily influenced by what is being described, than is the rate of modifier use.

Personal pronouns Various writing characteristics and strategies influenced the rate of personal pronouns[30] in the samples.

One such strategy was the use of proper nouns and noun phrases to avoid ambiguity. Since nearly all of the characters in the episode were men, there were many cases where the third-person masculine pronoun could have referred to more than one person (e.g., *There, his son told him that they were only going to talk about it. He was very sad.*) In many of the samples with few pronouns, their writers had been alert to this problem, and often avoided ambiguity by using the character's name or a noun phrase in cases where a pronoun would normally have been used (e.g., *After the sprinkler incident, Charles is ordered under strict observation, an action that his roommate fears will kill Charles.*).

One characteristic of many of the samples with few pronouns was ellipsis of the pronoun (e.g., *The psychologist describes his own distaste for the onset of senility, and ∅ promises to keep an eye on Whitely.*). Elision of the pronoun is more often possible in compound sentences, so it tended to occur more often in samples with many compound sentences, whereas in samples with many short, simple sentences, ellipsis was often impossible (e.g., *But Ben recognizes them. He sees the youth and realizes they are his old friends.* versus[31] *But Ben recognizes the youth and realizes that they are his old friends.*). Not all samples with few instances of pronoun ellipsis, however, had predominately simple sentences; some writers simply did not often employ ellipsis (e.g., *Then out of the trees came a man and he introduced it as a Twilight Zone story.*).

Samples with few pronouns tended to have more description of the setting than did those that did not, which may account for the fact that more part ones had few pronouns while more part twos had many. Often, the description was several sentences long, a relatively large proportion of most samples, and did not include any pronouns (e.g., *In a typical 1950's television version of an old age home, one resident attempts to resist the gradual lethargy that is overcoming his fellow residents. As the story opens, the other residents are seen in varying states of passivity and decrepitude. Our protagonist is more alert than the other residents.*).

Finally, samples with few pronouns often included commentary on the theme or symbolic nature of events. Such commentary rarely included personal pronouns (except, sometimes, a first-person pronoun), and often ran for several sentences (e.g., *Thematically, we are presented with a narrative dealing with the phenomenological experience of an aging man and his attempts to consciously break the norms and values attached to old age. That the aged are indeed subject to be passive recipients of authority (e.g., the son, the director) is presented as one aspect of a cruel and unfeeling society that views its elderly as frail in body and spirit. The attempt to break this stereotype is manifested and symbolized by a child's game, one that requires a certain degree of physical activity.*).

In general, high pronominalization was associated with simple sentences and a narrative writing style, whereas low pronominalization was associated with avoidance of ambiguity, pronoun ellipsis, compound sentences, and the inclusion of some interpretive and/or descriptive sections.

---

[30] The tagger category *personal pronouns* (PP) includes reflexives, but not possessives, which are tagged separately as PP$. I will use the term pronoun in the rest of this section to refer to personal pronouns tagged as PP.

[31] This sentence is my rewriting of the previous sentence.

Coordinate conjunctions Three samples were mainly responsible for the differences in number of coordinate conjunctions. The writers of these samples overused the connector *and*. Most of the sentences in these samples consisted of simple clauses or simple coordinate clauses. One of the samples, with an average of almost two *and*'s per sentence, had many run-on sentences (e.g., *Charlie responds that Ben is just afraid of new ideas, of looking silly and of making mistakes and refuses to go along with Ben's view that they are "old men" and need rest and cannot act impulsively and childishly any more*). Stylistically, the overuse of *and*, run-on sentences, and, to a lesser extent, little variation in sentence structure are not considered to be effective in getting one's message across or holding the reader's attention. Information about overuse of coordinate connectors might help writers improve such faults in their writing without the need for a human editor or a parser.

Prepositions The majority of samples with a high percentage of prepositions were part ones, which is due in part to the preposition's function of showing a relationship between two words. In introducing the scene and the characters, more relationships were described (e.g., *The story opens at an idyllic old-age home.*; *Inside the home we are introduced to some of its residents sitting quietly, not communicating with each other.*).

Although significant differences were found between samples with high versus low proportions of prepositions, the dissimilarity did not stand out when the samples were read. The most likely reason for this is that there were many errors made in the tagging of prepositions—both in tagging prepositions as some other part of speech, such as particle (e.g., *...the others sneak down* [RP] *the stairs...*), and in tagging other parts of speech as prepositions (e.g., *As he walked outside* [IN] *the other tenants stared at him...*). The high error rate is due to the difficulty involved in correctly identifying prepositions. Although prepositions belong to a closed class, having limited membership, many of the members of the preposition category belong to more than one part-of-speech category: many words that are prepositions also function as adverbs (e.g., *outside*), many function as particles (e.g., *up*), some function as conjunctions (e.g., *since*), some as adjectives (e.g., *opposite*) and the word *to* functions as an infinitive marker. The preposition was not the only part of speech that was subject to many tagging errors, but most of the others that were (e.g., subordinate conjunction, particle) didn't occur often enough to contribute very much to significant differences. Adjectives and adverbs also had a fairly high error rate, but since the majority of the words in both these classes are relatively unambiguous (e.g., in the former case, words ending with *-ic*, *-al*, *-ous*, *-ful*, etc. and in the latter case, words ending in *-ly*), the errors made in tagging adjectives and adverbs did not affect the results to the same extent as the errors made in tagging prepositions.

It is also possible that readers can tolerate a higher threshold of variation in certain parts of speech than in other parts of speech. Perhaps people are more attuned to the use of open class words, such as nouns, but don't notice large differences in closed class, or function, words, such as prepositions. Indeed, the motivation for the examination of function words by stylometrists is that function words are used less consciously than content words, and are therefore a better measure of the true author's identity. However, for my purposes, differences in preposition usage do not appear to be useful.

Sentence-final punctuation The differences in frequency of periods were (not surprisingly) directly related to the average sentence length: the differences were significant between samples from opposite ends of the spectrum. Most existing grammar checkers perform such computations, so access to this information is not new. If the different types of sentence-final punctuation were distinguished,[32] and each type occurred in each sample, this information might be more useful to the user than data about sentence length variations.

Commas The differences in frequency of commas were due to a variety of factors in each case. Samples that had a high comma frequency tended to have longer sentences. These samples usually had many parentheticals and adverbials, which were often placed in the middle of a sentence, thus requiring two commas rather than only one (e.g., *Mr. Whitely, the old man, will die here.* rather than *Ben… went to warn the director, Mr. Cox*). Conversely, samples with few commas tended to have short sentences, and few conjunctions, parentheticals and adverbials. Writers who used many commas were also more likely to have misused them and to have used them in optional places (e.g., before a coordinating conjunction when the comma was not required for disambiguation). These last two are copyediting issues, but flagging differences in comma usage may help writers eliminate unnecessary commas and ensure that they are consistent in their comma placement.

Other punctuation Although I had to collapse anything besides periods and commas into an *other punctuation* category, quotation marks were the punctuation marks primarily associated with high usage of other punctuation. Of the significant differences, most were a result of comparisons with two samples that did not have any punctuation besides periods and commas. Samples with a high percentage of other punctuation were all characterized by dialogue, quotations, quotation marks around names (e.g., *"Kick-the-can"*), and the use of quotation marks to show irony (e.g., *his "friend"*). A comparison of individual punctuation marks would be more helpful to writers who are trying to discover inconsistencies, since different punctuation marks imply different text characteristics. Also, some punctuation marks may be almost interchangeable in certain situations (e.g., commas rather than parentheses for parentheticals), and flagging their usage may increase consistency of usage. In larger text samples, it is likely that comparisons of each type of punctuation mark would be possible.

One problem affecting the analysis of other punctuation is that certain punctuation marks occur in pairs (e.g., parentheses), others do not (e.g., colons), and still others sometimes do and sometimes don't (e.g., dashes). I did not make any changes to the other punctuation counts in my data collection, but in retrospect, it would have made sense to have divided by two the counts of punctuation marks that consistently occur in pairs.

Differences between taggers The Brill- and POST-tagged texts had the same number of significantly different comparisons at the .05 level: 277, of which 271 were the same. At the .001 level, POST had 193 and Brill had 198 significantly different comparisons.

Brill had one more impossible comparison than did the POST-tagged text due to a typographical error in the data entry (a spurious semicolon) that was removed during the preprocessing of the text for the Brill tagger but was not removed from the text used for the POST tagger. Thus, in the Brill comparison, this sample could not be compared to another one, which did not have any punctuation besides sentence-final punctuation and commas. The POST comparisons found significant differences between these samples.

There were five other comparisons that were not significantly different for the Brill-tagged text, but were for the POST-tagged text.

---

[32] Neither tagger tags periods, question marks, and exclamation marks with distinguishing tags, but this would not be difficult to do.

In one case, there was the same accumulation of noun-tagging errors in the opposite direction, involving the same part two that was mentioned in the matching comparisons (section 5.1.3). The POST-tagged part one had many common nouns and few proper nouns, whereas the part two had many proper nouns and relatively few common nouns. Due to a few errors by both taggers and one preprocessing error, the counts in the Brill-tagged text were slightly higher for the proper nouns in the part one, and slightly lower for the common nouns in the part two, thus reducing the difference below significance.

In the second case, there were differences in how the taggers tagged base-form and past-tense verbs: the POST tagger erroneously tagged a couple of past-tense verbs as past participles (e.g., *Charles wants to know what changed Ben from the innocent young boy who once believed in magic.)* whereas the Brill tagger erroneously tagged a couple of non-third-person singular present-tense verbs as base forms (e.g., *Mr. Cox and Ben emerge from the manor.*). The POST tagger also tagged a few nouns as base-form verbs, and the Brill tagger tagged a few verbs as nouns, which resulted in an accumulation of errors in opposite directions.

In the third case, there were many modifiers in one sample in the POST comparison, compared to very few in another, which made the comparison significant, but because the particles were tagged as modifiers by the Brill tagger and the sample with few modifiers had a relatively large number of particles, the difference was lessened.

In the fourth case, the divergence in the tagging was negligible, and the difference in the chi-square value was very small. No single change contributed to the opposite results. Rather, changes of one or two counts in several categories were enough to alter the results.

In the last case, the divergence was primarily due to one word that was used frequently in the part two: *outside* occurred seven times. In five of these instances (e.g., *we hear the voices of the residents playing outside.*), the POST tagger incorrectly assigned the tag JJ (adjective), whereas the Brill tagger correctly assigned the tag RB (adverb). Therefore, both of these categories showed relatively large differences.

There were six comparisons that were significantly different in the Brill-tagged comparisons, but not in the POST-tagged comparisons. In three cases the comparisons involved a part one that had many particle tags (seven) in the POST tagging. All but one of these were tagged as adverbs by the Brill tagger, which, along with a few other words that the POST tagger did not tag as adverbs (e.g., *outside* —see above), resulted in a sizeable difference in the adverb count. In the other three cases, the comparisons involved one part two which had differences in the gerund/present participle category (VBG) because of errors by both taggers. The Brill tagger tagged some present participles as nouns (e.g., *Ben quickly realizes what is happening*), and the POST tagger assigned the tag VBG to some nouns (e.g., *After a brief back and forth about aging*). There were also divergences in the adverb category due to the failure of the Brill tagger to assign particles. In one of these cases, the part one also showed a variation that contributed to the change: there were more preposition tags in the Brill-tagged text, because the sample had many particles which had been tagged as prepositions (e.g., *Charles picks up the can*).

5.2.4 Sentence-initial tags

The five categories analyzed were conjunctions, determiners, modifiers, nominals, and pronouns. Eighteen non-matching comparisons were not possible because of expected frequencies of zero. The majority of these were due to the absence of modifiers in both parts, and some were due to the absence of conjunctions.

Of the 322 possible comparisons of pairs, 82 were significantly different at the .05 level; only twelve comparisons had chi-square values that indicated a lack of homogeneity between the samples at .001 significance.

Out of the 82 comparisons that showed lack of homogeneity, the main contrast in 36 of them was in the use of nominals and determiners; 28 were due to divergences in nominal usage; 13 were due to variation in usage of determiners; three were a result of both nominal and pronoun usage differences; one was due to pronoun variations only; and one resulted from distinctive uses of modifiers. Conjunctions did not contribute appreciably to any dissimilarities. Thus, the differences were caused almost exclusively by nominal and determiner usage.

Mismatches in nominal and determiner usage were primarily associated with the problem discussed earlier: the writer of one part knew the name of some or all of the characters, whereas the writer of the other part did not.

In the four cases in which pronoun use was distinctive, the three writers who used few pronouns sentence-initially knew the three main male characters' names and seemed to be trying to avoid the ambiguity of the masculine pronoun by using the characters' names as often as possible. The writer of the sample that had many pronouns (this sample was involved in three of the four comparisons) knew the main character's name, but not the names of the others; to circumvent this lack of knowledge, he used possessive forms to refer to the other characters (e.g., *<u>His</u> son; <u>His</u> roommate*), thus increasing his pronoun use.

Modifiers contributed to only one significant difference in sentence-initial tags. In this case, the part one contained many adverbial elements in sentence-initial position, whereas most of the sentences in the part two began with a nominal or a determiner. Although it appears that modifiers did not play a large role in these differences, however, the placement of adverbial elements, such as adverbs, prepositional phrases and adverbial clauses, contributed to more than just that one difference. Many adverbial elements begin with a preposition, rather than an adverb, so these variations would likely have been observed in preposition counts if prepositions had been included in the analysis.[33] Since they weren't, the contrasts primarily showed up in the preferred sentence-initial tag of the samples that did not have many adverbial elements when compared with those that did. For example, some of the differences in noun and determiner usage were due to comparisons involving a part one which had a high number of adverbials sentence-initially and which used all the characters' names; there were therefore few nouns or determiners in sentence-initial position.

Adverbial placement is mainly a stylistic factor, since adverbial elements are allowed more movement within a sentence than most other elements. For example, the adverb *reluctantly* can be placed sentence-initially (e.g., *Reluctantly, Charlie went back.*), medially (e.g., *Charlie reluctantly went back.*), or finally (e.g., *Charlie went back reluctantly.*). Although the placement of moveable elements often depends on the emphasis which the writer intends, and not all positions are possible for all adverbials (e.g., adverbial clauses can rarely be placed sentence-medially), many writers show a definite predisposition as to where they put moveable elements. Information about collaborative writers' preferred adverbial placement may be helpful to them when they are trying to make their documents more consistent.

Two writing styles were associated with adverbial placement in sentence-initial position: a reporting style and a connective style. The first of these appears to be due to a selected style of writing rather than preferred (and likely less conscious) adverbial placement. This reporting style emphasized time and location. Writing characterized by such a style had more adverbials throughout the

[33] In longer samples, constructions that a writer uses infrequently are more likely to occur.

writing, but especially sentence-initially since this is a salient position in the sentence (e.g., *Next day …*; *Upon returning to the residence…*). In the connective style, the adverbial elements that were placed sentence-initially were used mainly to establish conjunctive relationships, rather than providing temporal or locative information. Some of these modifiers were adverbs of cause (e.g., *Consequently,...; Therefore…*) or functioned as such (e.g., *More concretely,...*), connecting the new sentence to the previous one. The use of such connectives generally produced more cohesive documents.

Upon further examination, I found more evidence that certain relatively common sentence-initial tags (e.g., verbal elements and prepositions) that weren't used in my analysis skewed results to some extent, since the proportions of sentence-initial tags were altered. For example, a sample with a high proportion of prepositions sentence-initially was not significantly different from one with no prepositions sentence-initially, using the above five categories. However, when a comparison was made including the prepositions as a sixth category, there were statistically significant differences. I believe that they would have occurred often enough to be included in the analysis if the sample sizes had been larger, because it was primarily the shorter samples that did not have these initial tags. In larger sample sizes all relatively common initial tags should occur at least once, allowing them to be included in the analysis.

Differences between taggers[34] The Brill data had 25 impossible comparisons, whereas POST had only 18. Seventeen of these impossible comparisons were independent of the tagger used.

One comparison that was impossible for the POST tagger was possible for the Brill tagger. This comparison involved the same part two mentioned in the matching comparisons (section 5.1.4), which had no modifiers in sentence-initial position in the POST tagging, but had one in the Brill tagging. This time, the difference was significant. The determiner and noun categories accounted for the difference between the non-matching parts for the previously-mentioned reason that the writer of the part one did not use the characters' names, whereas the writer of the part two did.

Eight comparisons that were possible for the POST tagger were impossible for the Brill tagger. Three of these comparisons showed significant differences in the POST comparisons. One part one that had only a single modifier in the POST tagging had none in the Brill tagging due to a tagging error, thus preventing comparisons with four part twos that had no modifiers. Another part one also had only one modifier in sentence-initial position, and this word was tagged incorrectly by the Brill tagger, thus eliminating another three comparisons with the part twos that had no modifiers.[35] Finally, the part two that had some data accidentally removed during the preprocessing of the Brill data had only one occurrence of a conjunction sentence-initially, which was in the part that was deleted. A comparison was made impossible because there was a part one that had no conjunctions in sentence-initial position.

There were 82 pairs that showed significant differences in the POST comparisons, but only 74 in the Brill comparisons; 67 of these were independent of the tagger used.

There were six comparisons that were significantly different for Brill, but not for POST.

In the first case, the increase in the sentence count due to an error by POST was enough to move the comparison below significance: POST interpreted an abbreviation (*i.e.*) as the end of a sentence, which added one count to the conjunction category. The difference between the parts, how-

---

[34] Note that the taggers do not necessarily have the same number of sentence-final or sentence-initial tags in any comparison, since there are differences in how they tagged words and not all tags were included in the comparisons.

[35] The fourth part two with no modifiers was its matching part, mentioned in section 5.1.4.

ever, was due to the noun category: once again, the writer of the first part did not know the name of the protagonist, whereas the writer of the second part did. In the second case, several errors on the part of each tagger resulted in opposite outcomes. The difference between the parts was due to the noun and modifier counts. In the part one, there were many adverbials in sentence-initial position, whereas most of the sentences in the part two began with nouns. In the third case, the way I separated the dialogue into sentences when I preprocessed the samples for the Brill tagger was different from the way that dialogue was analyzed by POST. I treated fragments such as: *his roommate exclaims.* and *his son explains.* as part of the spoken sentence, whereas POST analyzed such fragments as full sentences if they occurred after the dialogue (i.e., after sentence-final punctuation and followed by sentence-final punctuation), but not when they occurred before the dialogue (e.g., *His son says, …*). The result was that in the Brill-tagged text there were fewer sentences, and thus fewer sentence-initial tags in the part two, which had a lot of dialogue. The sources of the significant difference were the determiner and noun categories. In the part one there was more description of the setting (*The time period*; *The residents*; *The home*; *A man*; *A nurse*) and the writer appeared to know only the protagonist's name; in the part two, there was more action, little description and the writer knew every character's name. In all three of these cases, the POST comparisons were very close to significance.

In the next three cases, only two part ones and two part twos were involved. The part one that was involved in two comparisons had no modifiers due to an error by the Brill tagger (mentioned above). The part two that was involved in two comparisons had a few minor differences due to tagging errors and how sentences in dialogue were analyzed differently. The other two parts were not tagged differently by the two taggers. In all three cases the source of the dissimilarity was the determiner category. Both part ones had many sentences that began with determiners (partly due to the obvious avoidance of characters' names), whereas both part twos had a wider variety of sentence-initial tags.

There were 12 comparisons that were significantly different for POST, but not for Brill.

Five of these comparisons involved the same part one, which had many modifiers and determiners in sentence-initial position; the five part twos had many proper nouns and pronouns in this position. There was a count difference in the noun and modifier categories of the part one due to an error by the Brill tagger that alone lessened the distinctiveness in two comparisons. In two other cases, minor changes in the determiner count for one and the noun count for the other contributed to the difference. In the last case, the comparison involved the part two which had some sentences deleted during the preprocessing. Three of the five sentences deleted during preprocessing began with a noun, thus skewing the results.

Another part one was involved in three of the twelve comparisons. In two cases, the part twos had the same tag counts, and the third case involved the part two with the deleted sentences. In the Brill tagging, the high preposition count of the part one was lowered because dialogue was handled differently.

A third part one was involved in two comparisons, one of which involved the part two with missing text. The part one had few nouns in sentence-initial position, apparently because the writer did not know the protagonist's name. In the Brill tagging, *one* was tagged as a cardinal number, whereas POST incorrectly tagged it as a pronoun (*<u>One</u> of the boys*). Since the scarcity of nouns contributed most of the difference in these comparisons, this change was enough to make these two comparisons non-significant in the Brill data.

In one case, the part one counts were the same, but the change in the part two due to missing sentences was enough to lessen the difference between it and a part one that had few nouns sentence-initially. In this part one, the writer knew the protagonist's name, but not the other character's names. Since the other characters were more often the actors in the sentences than the protagonist was, there were many pronouns and noun phrases in sentence-initial position.

In the final case, several errors by both taggers in both parts made a difference as to whether the parts were found to be significantly different or not. In the part one, there were many adverbials in sentence-initial position, whereas most of the sentences in the part two began with nouns.

5.2.5 Sentence-final tags

The four categories in the analysis of sentence-final parts of speech were modifiers, nominals, pronouns, and verbs and particles. There were ten non-matching pairs for which no comparisons could be made because the expected value of some tag was zero.

Of the 330 possible comparisons, only 33 of them showed significant differences at the .05 level. There were no comparisons that showed significant differences at the .001 level. The frequency distribution of sentence-final tags showed the least variability of all the tests over the samples.

In seven cases, the contrast was due to the verb category. All seven differences involved comparisons with the same part two, which had no modifiers and no pronouns in sentence-final position. This sample was characterized by the use of intransitive verbs and passives, and the placement of prepositional phrases sentence-initially, all of which resulted in having verbs placed in sentence-final position. It was also the shortest sample, which might have contributed to the lack of variability in sentence structure. The samples it differed from were characterized by transitive verbs, absence of the passive voice, and the placement of moveable elements at the end of the sentence (e.g., *now*).

In twelve cases the nominal category accounted for the main difference, in twelve cases modifiers and nominals were jointly responsible for the dissimilarity, in one case modifiers were responsible for the contrast, and in one case both nominal and pronominals accounted for it. One subject's writing was involved in eighteen of these 26 comparisons. This whole writing sample and one other part two had many modifiers, but relatively few nominals in sentence-final position. These samples were characterized by a lot of temporal and locative information (e.g., *now*, *there*), which was most often placed at the end of the sentence. Five comparisons involved the same part one, which had few modifiers and many nominals in sentence-final position. It had many prepositional phrases in sentence-final position and many modifiers placed sentence-initially.

In the comparison in which the pronominals also contributed to the difference, the sample with many pronominals was characterized by speculation about the characters' motivations and their understanding of other characters' motivations. This pondering resulted in more pronouns as objects (e.g., *Mr. Whitely wonders what changed <u>him</u>.*) and reflexives (e.g., *childlike actions were the only way to save <u>himself</u>.*). The sample with few pronominals was more action-oriented, and had many locative and temporal adverbs, and prepositional phrases in sentence-final position.

Differences between taggers There were nine non-matching pairs for which no comparisons could be made in the Brill-tagged text. Six of these were independent of the tagger used.

In four cases, comparisons were not possible for the POST-tagged text, but were for the Brill-tagged text. In two of these cases, the impossible comparisons were due to the absence of pronouns in a part two that, due to input error and tagging error, had two pronouns in the Brill-tagged text.[36] In the other two cases, the impossible comparisons were due to the lack of modifiers in two part two samples. In one case, where POST assigned RP (particle), the Brill tagger assigned RB (adverb); in the other case, POST assigned NN (noun) where the Brill tagger assigned RB. Both cases resulted in the modifier category having a zero value in the POST-tagged data, thus eliminating the comparison.

In three cases comparisons were not possible for the Brill-tagged text, but were for the POST-tagged text. All three involved a comparison with a part two, which, due to a combination of input error and tagging error, had no pronouns and no modifiers when tagged by the Brill tagger. Therefore, it could not be compared with two part ones that had no pronouns and one that had no modifiers in sentence-final position.

There were fewer comparisons that showed significant differences when the samples were tagged by the Brill tagger—25 compared to 33.

Sixteen of the 33 comparisons that were significant for the POST-tagged text did not show significant differences when tagged by the Brill tagger. In seven of these cases, the comparisons had the highest non-significant chi-square values for the part twos being compared with the particular part one.

In two comparisons, the difference between the Brill-tagged and POST-tagged texts was slight—only one sentence-final word had a different tag assigned to it. Because the chi-square values were very close to the significance level, in each case, the difference was enough to lower the level below significance.

In four cases, a part one had its modifier category lowered and its verbal category raised due to tagging differences and errors (e.g., POST tagged *thinking* as a noun, whereas the Brill tagger tagged it as a gerund in the phrase *way of thinking*). The four part twos had their modifier category raised and their verbal category lowered due to the non-tagging of particles (RP) by the Brill tagger, since in cases where POST assigned the tag RP, the Brill tagger assigned the tag RB, adverb.

In another three cases, part ones with a low number of modifiers and/or a high number of verbs had these categories more equalized primarily because of the RB versus RP tag assignment. When compared to a part two with a lot of modifiers and not many verbs, the differences between the samples were moderated.

In two other cases, the lowering of the verbal category made the comparison with a part one that had no verbs in sentence-final position non-significant.

In two cases, the difference between the Brill-tagged and POST-tagged texts was due to a relatively large number of sentence-final words tagged as verbs by the Brill tagger (in one sample, the Brill tagger was accurate in assigning more verb tags, whereas in the other, many of the verb assignments were errors). When the Brill-tagged part ones were compared with part twos that had high percentages of verbs sentence-finally, the divergences between them were moderated.

In the last three cases, a part two that had relatively few nominals had its nominal count slightly modified due to one different tag assignment by the Brill tagger, and the three part ones with high nominal counts had theirs slightly lowered through a combination of factors.

---

[36] Because the Brill tagger required pre-processing, there were more spurious errors introduced into the text.

In eight cases, there were significant differences found in the Brill-tagged text, but not in the POST-tagged text.

In two cases, the difference was due to the nominal category. In the first case, there were no differences in the nominal counts, but there was a difference in the sentence counts because of the use of dialogue in the part one: there were three sentence fragments that were analyzed as sentences by POST, but analyzed by me as part of the spoken sentence. The difference in counts overall lowered the ratio of nominals to other categories in the part one enough to make them significantly different from the part two. In the second case, there were three errors in the POST tagging, which added three to the nominal count of the part two, thus making its nominal count more similar to the part one's count. Two of these errors were due to the presence of exclamation marks. As mentioned in section 4.5.2, POST did not separate the exclamation mark from the preceding word and also mistagged that word. In this sample, a pronoun and an adverb were both mistagged as cardinal numbers. The third error was the mistagging of a present participle as a noun in: *He saw a boy under a tree counting*.

The next two cases involved a part one that had many sentence-final words erroneously tagged as verbs by the Brill tagger (e.g., *to bed.*) In the first case, the RB versus RP tagging also lowered the verbal count and increased the modifier count of the part two, thus increasing the difference between the two. In both cases, the main difference was in the modifier category, which also showed large differences in the POST comparisons, but since the other differences were smaller, the discrepancy did not reach significance.

In four cases, the difference was due to the verb category. These comparisons involved the part two, mentioned above, which had many verbs, but neither modifiers nor pronouns in sentence-final position. Once again, the change from RP to RB made a difference. Since verbs dominated in sentence-final position in this particular part two, the loss of the particle tags to the verb category in the part ones made a big difference.

There were more differences between the taggers for sentence-final tags than for any other tests. The main reason for this divergence was a combination of a high error rate in the tagging of particles (RP) by POST and the non-tagging of particles by the Brill tagger. In the Brill-tagged comparisons, the first category (modifiers) was usually higher than in POST, since many of the words tagged as particles by POST were tagged as adverbs. The last category (verbals) was usually lower than in the POST-tagged comparisons since the category includes particles. Given the error rates, perhaps particles should not have been included. However, the effect on the modifier category would have remained, since the Brill tagger assigned the adverb tag to most particles.

5.2.6 Comparison of means

Sentence length There were 76 comparisons that were found to be significantly different at the .05 level using the comparison of means test. Thirty of these were also significant at the .001 level. In nearly all cases of significantly different comparisons, it was the part one that had a longer average sentence length and a larger standard deviation. As mentioned in section 5.1.6, perhaps writers included more description in the first half to set the scene, which resulted in longer sentences, or perhaps writers were in a hurry to finish the second part so that they could leave, which resulted in shorter sentences. Overall, when the significantly different parts were compared, the writing in the parts with longer average sentence lengths had more compound and more complex sentences. Many of them had a low number of adverbs and/or adjectives, because the writers used adjectival and adverbial phrases or clauses instead. They also tended to have many parentheticals (e.g., *His friend…goes to the home's doctor (whether psychiatrist or general practitioner is somewhat unclear)…*).

Of the 76 comparisons, 43 for POST and 42 for Brill (39 of which were the same) did not show large differences between the observed and expected values in the sentence-final punctuation count. There were no comparisons that had a large difference between the expected and observed values, that did not also show significant differences in the comparison of means test.

There are several factors that contributed to the difference in number of comparisons identified as significant by the comparison of means test and the chi-square test.

First, in 27 of the POST and 26 of the Brill comparisons, the counts of the sentence-final punctuation differed from the number of sentences that had been calculated using the **UNIX Writer's Workbench** program **style**. In most cases, the counts differed by only one or two, but the amounts were as high as five. Upon examining these comparisons, I found that those with a small count difference tended to have higher differences between the observed and expected values, so the small discrepancy in the sentence counts might have changed the outcome. The differences in counts were due to a variety of reasons. One reason was that **style** did not include sentence fragments (e.g., *Poor Charlie.*) or sentences that were interpreted as fragments (e.g., *Wake up!*) in its sentence count. Another reason was that in the chi-square test, tags were being counted, rather than sentences. Since some writers used intersentential punctuation (e.g., *Ben is trying to convince the main man (Charlie?) to stop acting foolish*), or repeated sentence-final punctuation (e.g., *The tenants all thought he had gone loony!!*), there were more sentence-final punctuation tags in those samples because the taggers do not distinguish such uses of periods, question and exclamation marks from sentence-final punctuation. There were also a few errors on the part of the POST tagger in which abbreviation periods were deemed to be sentence-final punctuation (e.g., *i.e.*). Since the Brill tagger required preprocessing into sentences by hand, a few oversights (e.g., a period I forgot to separate from the previous word with a space) caused count errors.

Second, since I chose my cut-off point for which divergences between the observed and expected values seemed to be high enough to contribute a reasonable amount to the significant differences simply by looking at the data and comparing values, my cut-off point might not be appropriate. More investigation is needed to find out at what level readers themselves detect a difference to find an accurate measure.

Finally, the comparison of means is a more sensitive test than simply comparing the number of sentence-final punctuation marks because it not only evaluates the difference in average sentence length, but also takes into account the standard deviation. Thus, if two samples have similar average sentence lengths, but one has sentences that are close to the same length throughout, while the other has sentence lengths that vary widely throughout, the comparison of means test will find the difference, whereas the comparison of the expected and observed values for number of sentence-final punctuation marks will not.

5.3 Discussion

5.3.1 Summary of results

Some interesting stylistic differences were revealed by the statistical tests that were performed. Examination of matching part ones and twos suggests that when stylistic inconsistencies are detected in a single writer's work, they do not seem to reflect habits of writing. Therefore, the stylometric theory that writers' styles are stable was not refuted. Rather, one reason that the differences arose was due to some writers' initial lack of knowledge (of the characters names), and attempts to circumvent this problem (by using noun phrases and pronouns to refer to them). A second problem was one that many writers have: maintaining consistent verb tenses. Finally, in two cases, the use of modifiers changed from one section to the next, influenced perhaps by a self-imposed time-pressure and a more action-oriented second half of the story. Also consonant with stylometric theory was the lack of any apparent time effect on writing style, at least after a one week interval.

Comparisons of non-matching part ones and twos that showed statistically significant differences revealed a wider variety of stylistic inconsistencies, as stylometric theory predicts. Some of the inconsistencies were of the type that copyeditors are expected to detect, some inconsistencies did not reveal anything interesting about writing style, while others were influenced by the content or purpose of the text. Finally, some of the inconsistencies were of the type that I was most interested in: they were associated with the style, or feel, of the writing.

Copyediting inconsistencies Inconsistencies in verb tense and noun usage were similar to those found in the comparisons of the matching writing samples. Other such inconsistencies were detected in the use of commas and quotation marks.

Uninformative inconsistencies Perhaps certain stylistic factors are not as salient to readers and so they have higher thresholds for tolerating differences. Or, the high error rate in tagging some parts of speech, such as prepositions, might have invalidated the results of the tests. In cases in which the roles of the parts of speech are very similar (e.g., adverbs and adjectives) they might have to be combined into one category to provide useful stylistic information.

Text-based inconsistencies Introductions tended to have more indefinite articles than did other parts of the text. When setting the scene, writers used more adjectives, but when describing action, they used more adverbs. The use of particular vocabulary, due to content (although alternative vocabulary was usually available), influenced many of the word-length differences. Another such inconsistency was in the use of two- versus four-letter words: when the individual protagonist teamed up with a group of people in the second part, the predominant pronoun *he* changed to *they*, which affected samples with many pronouns. The prevalence of male characters lead to low pronominalization when writers noticed and tried to avoid the ambiguity of *he*. Rate of pronominalization was also influenced by genre: narrative sections contained many pronouns, whereas descriptive and interpretative text had few.

Stylistic inconsistencies Some of these, such as the preferred placement of moveable elements and preferences for transitive rather than intransitive verbs, are probably relatively unconscious writing habits of the type that authorship studies seek. Others, such as a "reporting" style associated with the placement of locative and temporal adverbials at the beginning of the sentence and elevated vocabulary associated with a high percentage of long words, are more likely to be deliberate. Still others reveal writers' weaknesses: for example, a syntactically "boring" style characterized by the overuse of coordinate conjunctions. Other stylistic differences that were revealed by this analysis may fall into more than one of these categories: a "simple" style characterized by a high percentage of two- and three-letter words, high pronominalization, and short sentences; a "connective" style with many adverbs of cause placed sentence-initially; highly descriptive or sparse samples that had correspondingly high or low percentages of modifiers.

5.3.2 Differences between taggers

As can be seen by comparing the histograms for the matching pairs (see section 5.1), there was little disparity between the rankings obtained using the Brill-tagged text and those from the POST-tagged text, particularly for the sentence-initial tags and the tags overall. Only two of the 60 chi-square comparisons had opposite results.

In the tests for the non-matching pairs, there were fewer significant differences in the Brill-tagged text than in the POST data for sentence-initial and final tags, but the taggers had the same number of significant differences in the tags overall. Examination of the divergences revealed that there were several contributing factors: errors on the part of both taggers, text preprocessing errors, and differences in sentence analysis.

Overall, neither tagger stood out as noticeably better—each had its strengths and weaknesses. The main cause of opposite results obtained from tests performed on the Brill data compared to those done on the POST data seems to be the length of the writing samples. The test that had the largest amount of data, tags overall, had the fewest divergences, whereas the sentence-initial and sentence-final tag comparisons had a relatively high number of divergences due primarily to minor tagging differences. In some cases, even lowering a part-of-speech count by one meant the difference between significance and non-significance. These results suggest that, given the taggers of today, longer writing samples are needed to obtain more reliable information.

5.3.3 Conclusion

Of the six measures I used, overall part-of-speech distribution revealed the most information about both low- and high-level stylistic inconsistencies. Detection of low-level inconsistencies would be the likeliest candidate for incorporation into existing style checkers because the relationship between the inconsistency detected and how the user interprets and acts on this information is relatively straightforward. For example, if writers are told that present tense usage is high in part one, but low in part two, they know that the use of tense is not consistent and they must therefore decide whether there is a valid reason for the inconsistency. If there is not, they must decide which tense to use, then change instances of inappropriate tenses. Although inconsistencies at the copyediting level were not the type that I set out to find, the automatic detection of such inconsistencies would alert editors to such problems, analogous to good spellers' use of spelling checkers to find errors they may have otherwise overlooked themselves, thus potentially speeding up and improving the accuracy of copyediting. This type of support would be especially useful when editing long, multi-authored documents. The relationship between inconsistencies and the information they reveal, however, is not so straightforward when high-level inconsistencies are detected. The causes are not the same in all cases, and often there is more than one contributing factor to such inconsistencies. In order for the tests that reveal high-level inconsistencies to be used by writers, there would have to be a great deal more linguistic analysis of the patterns revealed by part-of-speech distribution to determine how this information should best be interpreted.

Comparisons of the tags used in sentence-initial and sentence-final position provided some information that was different from that of part-of-speech distribution overall. For example, they revealed some of the preferences writers have as to where to place moveable elements in a sentence. However, the results of these two tests were not as robust. Comparisons of sentence-initial and final tags would provide more, and more reliable, information in longer samples for two reasons. First, longer texts would likely contain at least one instance of all typical sentence-initial and final tags, thus allowing for a more complete analysis. Second, larger samples would not be affected to the same extent by minor tagging and preprocessing errors and discrepancies.

Comparisons of the percentage of two- and three-letter words indicated that there was not much variability, at least in this set of writing samples, but an interesting cluster of differences that distinguished high from low percentages was revealed. More research as to the validity of these results, though, would have to be done before such information was given to writers.

Sentence-length comparisons are a standard part of current style checkers, but adding information about standard deviation would provide writers with a more complete picture of the differences. However, the usefulness of such information is questionable. Although large deviations in sentence length are indicative of stylistic differences, people are often unsure about what to do with this information, and it is not clear what side effects result when people do try to alter their average sentence length (Sanford & Moxey, 1989). McGowan (1992) found that several of his subjects commented on the difficulty of reducing sentence lengths in a test file given to them. Also, there have been some studies indicating that writers' attempts to reach a target sentence length actually results in worse, rather than better, prose (see Oliver, 1985). Therefore, the advantage of including information about average sentence length remains in question.

Word-length distribution seems to be the most promising test at this time. Not only did it reveal an appreciable amount of information, but much of this information would not require linguistic analysis before writers could use it. The words that contributed to the significant differences could simply be listed in ranked order (e.g., use of nine-letter words was significant: *geriatrics*: ten times; *residence*: nine times; *residents*: six times, etc.), and writers could decide whether the grammatical or vocabulary differences that had been revealed were detrimental to the writing. It remains to be seen, however, whether this information would be useful to writers or editors. Differences in vocabulary are more salient to readers than are grammatical differences, and presenting data about vocabulary might involve displaying the obvious.

The main question of my investigation was how well stylostatistical methods that are used for identifying authors could be adapted to the problem of detecting stylistic inconsistencies. Overall, the results are encouraging. Examination of most of the statistically significant results revealed distinct stylistic differences between the samples being compared. Further, a wide variety of inconsistencies on various levels were revealed, particularly between writing samples written by two different writers. Moreover, the majority of these inconsistencies had not been immediately obvious to me on perusal of the samples before performing the stylostatistical tests, which suggests that the results of such tests may indeed accelerate and improve people's detection of stylistic inconsistencies.