Springer
London
Berlin
Heidelberg
New York
Barcelona
Budapest
Hong Kong
Milan
Paris
Santa Clara
Singapore
Tokyo

Mike Sharples and Thea van der Geest (Eds)

The New Writing Environment

Writers at Work in a World of Technology

Springer

Chapter **9**

Detecting Stylistic Inconsistencies in Collaborative Writing

Angela Glover and Graeme Hirst*

9.1 Introduction

When two or more writers collaborate on a document by each contributing pieces of text, the problem can arise that, while each might be an exemplary piece of writing, they do not cohere into a document that speaks with a single voice. That is, they are *stylistically inconsistent.* But given a stylistically inconsistent document, people often find it hard to articulate exactly where the problems lie. Rather, they feel that something is wrong, but cannot quite say why.

An example of stylistic inconsistency can be seen in the following sentence, which is from a brochure given to hospital patients who are to undergo a cardiac catheterisation. (The parenthesised numbers are ours, to refer to the individual clauses.)

> (1) Once the determination for a cardiac catheterisation has been made, (2) various tests will need to be performed (3) to properly assess your condition prior to the procedure.[1]

Clause 1 and (to a slightly lesser extent) clause 3 are in medical talk, as if in a formal communication from physician to physician; clause 2 is much more informal, and is expressed in ordinary, lay, language. The effect of the two styles mixed together in the one sentence is a feeling of incongruity – which was presumably not intended by the author or authors. This example, however, is unusual in its brevity. More often, the problem of inconsistency emerges only over longer stretches of text, especially where the granularity of the multiple authorship is at the paragraph, section, or chapter level. Moreover, while stylistic inconsistencies arise primarily in jointly written documents, we do not exclude the possibility of their occurrence in singly authored texts, especially those where different parts were written at different times or, initially, for different purposes.

Our ultimate goal in this research is to build software that will help with this problem – that will point out stylistic inconsistencies in a document, and perhaps suggest how they can be fixed. In this chapter, we report some of our initial explorations and data collection.

* Address correspondence to the second author. E-mail: *gh@cs.toronto.edu.*

9.2 Style

When we say *style* here, we are *not* speaking of lower-level copyediting concerns such as punctuation or formatting – the domain of most of the elements of contemporary 'style checkers' (although consistency in these matters is obviously important as well). Rather, we are speaking of higher-level concerns such as the author's choice of words and syntactic constructions that give a piece of writing its particular 'feel'. And we are speaking not of literary style or literary texts, but rather of everyday writing such as magazine articles, technical manuals, academic papers, business letters, and so on.[2]

Despite the influence of *genres* and *group styles* of writing, people generally develop an *individual style* within the group norms; indeed, many cultivate distinctive styles. But when writers work collaboratively, trying to merge their styles can result in time-consuming revision, frustration, and interpersonal conflict, and even then it is not always successful. Ede & Lunsford (1990) noted that 'of the disadvantages [of collaborative writing] cited, perhaps the most often mentioned involved what one engineer called "the tough task of making a common single style from numerous styles"' (p. 60).

But stylistic consistency is important. A variety of styles within one document may be distracting to readers and may lead readers to believe that the document was written in a careless or hurried manner (Farkas, 1985). Shifts in style are a cognitive burden to readers, since they force readers to change their expectations (Enkvist, 1964), and reading comprehension may be impaired by the additional load.

One way to avoid the problem of merging styles would be to impose a particular style on the writing before the collaboration begins. In practice, however, this is difficult to achieve. Many people are loth to abandon their preferred writing style, and this is not simply obstinacy; one of the strategies that writers use to reduce the cognitive burden of the task is to draw on a routine or well-learned procedure (Flower & Hayes, 1980). Changing one's well-developed writing style will make the writing task more difficult, since more attention will have to be allocated to an aspect of writing that is normally under less conscious control. Moreover, many writers do not want to impose stylistic restrictions on others. Writing together often already involves conflicts about content and procedure, which can be time-consuming and stressful to resolve, and allowing each collaborator to write in his or her own writing style avoids an additional source of potential conflict among the group members. In any case, people tend to have difficulty describing style, so even when the imposition of a specific style is acceptable, it is not always possible for the writers adequately to provide one another with the information needed to write in the agreed-on style.

It is also difficult to integrate different writing styles after a collaborative document is complete, or if the collaboration involves the assembly and editing of pieces of text written earlier, because people are generally poor at consciously recognising inconsistent style. They might be dissatisfied with the document, yet not know why. And even when people recognise that the document does not have a single style, they are often unable to articulate the specific stylistic inconsistencies they have noticed.

Thus, writing a multi-authored document with a single voice can be very difficult. It would therefore be useful if writers' aids – especially those designed for supporting collaborative writing – were able to assist in this. The task would have two components. First, the system would have to be able to discover stylistic inconsistencies; and second, it would have to present its findings in a manner that would enable the authors to correct the inconsistencies – which implies being able to articulate the problems, and any suggestions for change, in terms comprehensible to the average user.

9.3 Computational Approaches to Style

9.3.1 Qualitative and Quantitative Methods

There has been little study, even informally, let alone computationally, of style in the qualitative, quotidian sense that we use it here. Hovy (1988) coded a large number of stylistic heuristics in his language generator, PAULINE. And DiMarco, Hirst, and their colleagues (DiMarco & Hirst, 1993; DiMarco & Mah, 1994; Green & DiMarco, 1993; Hoyt & DiMarco, 1994; Makuta-Giluk & DiMarco, 1993) developed simple grammars of style for syntactic correlates of attributes such as abstractness, obscurity, and amity, and have implemented them in a number of programs.

But most of the computational research on style has been statistical. Such research often goes by the name of *stylometry*, *stylostatistics*, or *computational stylistics* (Milic, 1967, 1991; Kenny, 1982). This differs from the research cited in the previous paragraph (which has also been called *computational stylistics*) in that it is merely computer-assisted analysis of style; any qualitative analysis required must be done by a human. Applications of statistical methods include the identification of quantitative characteristics of authors' writing styles, especially for authorship attribution ('author fingerprinting'); the search for stylistic sets of markers associated with different genres; and the description of stylistic features of historical periods, including investigation of diachronic language change.

Authorship attribution studies are particularly relevant here, because the aim of such research is to identify differences between the writing styles of different authors, and to discover the style markers of particular authors. Presumably, stylistic inconsistencies are present in a document exactly to the extent that an author-identification technique could (at least in principle) determine that different parts of the document have different authors. Some author-identification studies have indeed tried to find stylistic differences within one, possibly collaboratively written, document. Moreover, the difficulty faced by collaborative writers in trying to merge writing styles may be aided by the identification of the writers' specific stylistic differences. Therefore, author identification techniques provide a starting point for a stylometric study of collaborative writing and style merging.

9.3.2 Author Identification Studies

Many different kinds of test have been proposed for use in author identification. Table 9.1 lists a number of them; the tests are grouped in the table by the degree of linguistic analysis of the data that is required for the test to be carried out. Not included in the table are those tests that are used to identify a particular genre (e.g., ratio of *thus* to *therefore*), or a particular subject (e.g., particular topic words), nor tests that do not seem likely to be of any applicability in collaborative writing (e.g., alphabetics, syllabification).

We now describe four representative studies of author identification, with an emphasis upon studies that looked for stylistic inconsistencies within a single text.

- Unanalysed text
  - Register of words used (formal, slang, technical, etc)
  - Frequent words (at least 3 per thousand)
  - Sentence length (mean and standard deviation)
  - Word length (mean and standard deviation)
- Tagged text
  - Type/token ratio
  - Distribution of word classes (parts of speech)
  - Distribution of verb forms (tense, aspect, etc)
  - Frequency of word parallelism
  - Distribution of word-class patterns (e.g., determiner + noun + verb)
  - Distribution of nominal forms (e.g., gerunds)
  - Richness of vocabulary
- Parsed text
  - Frequency of clause types
  - Distribution of direction of branching
  - Frequency of syntactic parallelism
  - Distribution of genitive forms (*of* and *'s*)
  - Distribution of phrase structures
  - Frequency of imperative, interrogative, and declarative sentences
  - Frequency of topicalisation
  - Ratio of main to subordinate clauses
  - Distribution of case frames
  - Frequency of passive voice
- Interpreted text
  - Frequency of negation
  - Frequency of deixis
  - Frequency of hedges and markers of uncertainty
  - Frequency of semantic parallelism
  - Degree of alternative word use (preference for synonyms)

Table 9.1 Some of the tests that have been proposed for use in author identification, organised by the degree of linguistic analysis required.

Both Morton (1978) and Smith (1988) studied the play *Pericles*, which is alleged to have been written by two different playwrights. It is generally accepted that acts III, IV, and V of this play were written by Shakespeare; however, acts I and II have been attributed to various authors. Morton's study found no significant differences in the preferred position of frequently occurring words, the occurrence of common collocations, or proportionate pairs of words (e.g., the ratio of *no* to *not*) between the first and second parts of *Pericles*. However, he did find significant differences between *Pericles*, selected essays of Bacon, and several plays by Marlowe. Morton therefore concluded that *Pericles* was in fact written by one author – Shakespeare. Smith, however, claimed that Morton's study is deficient in several respects. Smith therefore conducted a study that compared the rates of usage of the first words of speeches (excluding proper names) that occurred at least ten times per thousand in one or more of the plays under investigation. He separately compared both parts of *Pericles* with plays by Shakespeare, Chapman, Jonson, Middleton, Tourneur, Webster and Wilkins. The rates of occurrence of first words of speeches were often similar among Shakespeare and his contemporaries, but groups of words could be used to distinguish one playwright from another. Whereas the second half of *Pericles* was most similar to Shakespeare's other works, the first (disputed) half was most similar to Wilkins's play.

Morton also analysed *Sanditon*, a novel that Jane Austen did not complete before her death. Using a summary of *Sanditon* that Austen had written, 'Another Lady' finished the book for publication. She deliberately imitated Austen's style to try to produce a stylistically consistent novel. Morton was interested in whether stylistic differences could be detected between the two writers, despite the latter's attempt at imitation. He compared characteristic writing habits of Austen's, culled from *Emma*, *Sense and Sensibility*, and the first part of *Sanditon*, to the second part of *Sanditon*. The Other Lady was able to reproduce relatively mechanical habits such as the use of *and* following commas, semicolons, and colons. However, less-conscious habits, such as the ratio of *with* to *without*, were not successfully imitated.

Irizarry's (1991) computer analysis of *Infortunios de Alonso Ramírez* (*IAR*) attempted to discover whether the novel was collaboratively written, or had a single author. The novel purports to be the description of an illiterate sailor's life adventures written by an amanuensis, the writer Carlos de Sigüenza y Góngora, but it is believed by some to be a complete work of fiction. Irizarry investigated the plausibility of the collaboration by comparing *IAR* to three other narrative works of Góngora, all of which were written within three years of *IAR*. Five of the tests revealed significant divergences in style between *IAR* and the other works. Variation in word length was the only test Irizarry tried that was not useful in distinguishing the works. She therefore concluded that the novel was a collaborative effort.

McColly (1987) investigated the style and structure of the Middle English poem 'Cleanness or Purity' to discover whether the two parts are halves of the same whole, or whether they form two distinct texts. He compiled function-word frequencies, as well as frequencies of certain modifiers (e.g., *many*) and pronouns (e.g., *all*), discarding frequencies of less than one per thousand, for a total of 59 words. He then compared the relative frequencies of these words in the two halves of the poems, as well as in random samples from each half. The difference between the halves was significant, particularly the use of conjunctions and some verb tenses. He concluded that these differences reflect a lack of structural unity in the poem.

These studies suggest some techniques that might be appropriate for finding stylistic differences among collaborative writers, as they had the common goal of detecting such differences within a single piece of text. In section 9.4.2 we will consider how techniques such as these can be adapted to our present goals.

9.4 Explorations of Inconsistencies of Style in Collaborative Writing

9.4.1 The First Step

Let us now consider in detail our goal of helping collaborating writers achieve consistency of style. This will require advances in a number of areas in stylistics and computational methods:

- We need to know what kinds of things do and don't count as undesirable inconsistencies.
- We need to be able to detect these things computationally.
- We need to be able to articulate stylistic problems in terms that the user can understand.

- We need to be able to suggest to the user, again in simple terms, how stylistic problems can be corrected.

A catalogue of undesirable stylistic inconsistencies awaits further research. We must not simply assume that any identifiable inconsistency will necessarily be distracting to the reader, or even that such a distraction is necessarily bad; a skilled writer might deliberately use an inconsistency for effect. Moreover, identifiable stylistic differences between parts of a document might be no more than a reflection of different content or purpose. For example, a technical manual might be divided into introductory information, instructions for operation of the equipment, and technical specifications; consequently, the sections might be quite distinct by any stylistic measure, but mutually harmonious nonetheless. Similarly, in this chapter, the presentation of the statistical analyses in section 9.6.2 is, by the nature of the material, stylistically dissimilar from the preceding sections, without ill effect. But gratuitous differences in style can probably be assumed to be deleterious unless shown otherwise. For example, a seemingly random mixture of formal and informal, technical and non-technical, or static and dynamic styles would surely be a candidate for revision.

Methods and terms for explaining stylistic problems to users and helping them with improvements must also await future research. Certainly, it would not be adequate to tell a user simply that one paragraph is dynamic and the next static and that one or the other should therefore be rewritten to make them match. Even if the user understands the problem, this abstract advice gives little clue as to how to go about the task of rewriting. (In fact, the stylistic analyser of Payette & Hirst (1992) did simply tell users, sentence by sentence, whether their text was static or dynamic or whatever; but the program was intended for use in the classroom by advanced learners of a second language, and it was assumed that in such a context, this kind of analysis would be acceptable.)

The most tractable part of the problem at present is clearly the detection of stylistic inconsistencies (whether bad or benign), and it is that to which we turn our attention for the remainder of this chapter.

9.4.2 Author Identification Techniques and the Detection of Stylistic Inconsistencies

Our starting point is the reseach on author identification that was reviewed earlier. There are clear similarities between the problem of author identification and that of finding stylistic inconsistencies. In each case, we are trying to see if there are attributes of a text, or set of texts, that have one value in some areas and a different value in others.

But there are significant differences, too. In author identification, the task is generally to compare a disputed text with an attested text. The attributes of interest are those whose values are expected to be relatively constant for a single writer and yet vary from person to person; they may be purely quantitative, and need not be correlated with the 'feel' or qualitative style of the text at all. In finding stylistic inconsistencies, on the other hand, there is no attested text as such, and the task is to compare fragments of a single text with one another. The attributes of interest are those whose variation would be deleterious to the quality of the paper, regardless of their expected inter- or intra-individual variability, and it must be possible to characterise their qualitative effect upon the 'feel' of the text. Also, the granularity of the analysis is different in the two problems. Author identification generally involves the analysis of corpora of tens of thousands of words. In an analysis to assist collaborating writers, by contrast, the whole document might be only a few thousand words, and the area of analysis could be as small as a couple of paragraphs.

So we decided to explore the question of how well we could adapt to identifying stylistic inconsistencies in shorter writing samples the kinds of methods that are used for author identification. This requires an immediate defence, for, despite the similarities between the tasks, any purely quantitative method seems, *a priori*, to be inherently inappropriate for a goal that emphasises automatic qualitative analysis. It is of little use to the writers to be told, for example, that two sections of text differ in their proportion of three-letter words or their distribution of prepositions. Even if we don't yet know how best to present results to the users, it is intuitively clear that quantitative terms, although concrete, are likely to be even worse than the qualitative but abstract terms that we scorned earlier. One would expect, therefore, that the qualitative approach of DiMarco & Hirst (1993) would be a more appropriate foundation for our work, as it was for Payette & Hirst (1992).

Nevertheless, quantitative methods have their advantages. First, they are relatively well understood, and are easy to implement and fast to run, compared to grammars of style. Second, some qualitative measures of style can indeed be easily correlated with quantitative measures – in particular, some inconsistencies in stylistic register might be obvious just from counts of lexical indicators (such as the use of slang, technical jargon, or highfalutin words) in different parts of the text. So we decided that it would be useful to explore quantitative methods, just to see how far we could take them.

9.4.3 The Design of an Exploratory Study

Our hypotheses To obtain data for our study, we devised a task in which subjects would write a text of several hundred words in two parts (see section 9.5). The assumption is that each writer's second part will be stylistically more like their own first part than like anyone else's; and hence also more like their own first part than anyone else's is. Then, by pairing each first part with second parts by other writers, we would be able to construct for analysis a set of 'collaborative' documents, with possible stylistic inconsistencies, that were controlled for content. In addition, we would be able to take the parts separately, and compare each first part with each second part, with various stylistic tests, to see if we could match them up correctly.

A second question we had was whether people write consistently over time. One of the premises of author fingerprinting is that adult writers have developed a stable style, and that in fact it is almost impossible to significantly change one's mature style, even consciously (Cluett, 1976). However, our personal writing experiences suggested that this might not be the case. As writers, we have had the frustrating experience of adding to our own previously written work and finding it difficult to continue the writing in a consistent style. We therefore decided to have a group of subjects for whom a week would elapse between the writing of their first half and that of their second half, so that we could investigate whether the two parts were less consistent for these subjects than for those who wrote both parts on the same day.

Thirdly, we were interested in whether people adapt their style to the other author's when they add to a previously written document. Specifically, we wanted to know whether reading a co-author's writing affected one's own writing. If so, separate writers who pass the document from writer to writer, rather than partitioning the document, might create fewer stylistic inconsistencies. To investigate this question, we decided to have some subjects write only the second half of a text, doing so after reading another subject's first half, to find out whether their writing would exhibit more consistency with the first half that they had read than two halves written independently by different subjects.

Tests selected We investigated as large a pool of stylistic features as possible. There were three reasons for this. First, since texts tend to have more features in common than not, and the features that are inconsistent will be at least partially dependent on the particular texts being compared, finding these features might be difficult (Crystal & Davy, 1969). Rather, writers probably each have a set of characteristic stylistic features. These sets might overlap to different extents and in different ways. Systematically exploring many variables is more likely to yield results than investigating only one or two, as seemingly improbable but significant features might be discovered (Mosteller & Wallace, 1964). Moreover, since stylistic features are not independent of content, authorial attitude, and rhetorical stance, using a variety of tests might reduce the effect of this dependence (Dixon & Mannion, 1993).

Despite our remarks in the previous section, we did not count lexical indicators of stylistic register. First, this would have required an extensive lexicon of stylistic connotations, which we did not have. Second, given that our data were to be from a simple writing task deliberately controlled for content, we did not expect significant variation in this area.

Rather, grammatical aspects of style seemed to be the best candidates. Part-of-speech assignment is relatively unambiguous and, although writers often carefully select the particular word they want, they generally do not consciously select the part of speech they want to use, which, rather, seems to come from their personal style. That is, syntactic preferences tend to remain static in the adult writer. The difficulty that most writers experience when attempting to reformulate syntactic constructions provides further evidence that their preferences are largely unconscious, since such conscious analysis is rarely done. And finally, unlike vocabulary, syntax is not highly variable across different domains of discussion (Milic, 1967). Therefore, although the majority of the author identification studies reviewed in section 9.3.2 largely investigated word choice, we chose to focus on syntax.

Stylistic tests were chosen according to three criteria. First, they should have been used successfully in previous work on stylistic analysis; we did not want to start inventing new tests without first trying existing ones in this new application. Second, they should be appropriate for short writing samples – unlike type/token ratios, for example, which require much larger sample sizes. Third, they should be possible to carry out on either unprocessed or tagged text, rather than parsed text, since automatic and robust parsing of unrestricted text is still a difficult and time-consuming task. We decided to use the following measures:

word length distribution;
sentence length: mean, range, and standard deviation;
percentage of two- and three-letter words;
part-of-speech distribution (including punctuation);
relative proportion of each part-of-speech class in sentence-initial and sentence-final position.

9.5 Experiment

Before carrying out the experiment, we did a pilot study to determine approximately how long the experiment would take, and whether 500 words was an appropriate target sample length. Ten students participated in the study; they included undergraduate and graduate students, and native and non-native speakers of English. Due to the poor writing quality of some of these samples, we decided to impose restrictions on whom we would accept as subjects in the actual experiment.

9.5.1 Subjects

Subjects ($N = 20$) were solicited by electronic bulletin board and by poster. They were paid either $15 or $25 for their time, depending on whether they were required to make one or two visits. They were told that the experiment involved writing, but were not informed that writing style was being investigated until after they had completed the experiment.

Subjects were mainly graduate students from various departments at the University of Toronto. We required that they be native speakers of English, in an attempt to reduce the probability of syntactic errors that could confound the stylistic analysis. And we required that they be graduate students or hold a graduate degree, to ensure that subjects had had enough experience in writing to have developed a personal writing style.

Subjects were given an optional questionnaire on their gender, age, level of education, and occupation or field of study. Most people answered all questions, providing us with the following information. There were nine female subjects and eleven male subjects. They ranged in age from 21 to 47, sixteen of whom were in their twenties. Subjects were studying or had studied in the following areas: business administration, computer architecture, computer science (2), education, engineering, English, genetics (2), literature, mathematics, neuroscience, organisational behaviour, psychology, sociology (4), and zoology (2).

9.5.2 Procedure

The basic task of the subjects was to watch a 25-minute episode of the television program *The Twilight Zone* entitled 'Kick the Can', and then summarise it in approximately 500 words.

There were groups of two to five subjects in each session. They were given pen and paper when they arrived. Once everyone was assembled, they were given written viewing instructions that differed according to which experimental condition was being run. The subjects were instructed not to talk to each other during the experiment, but were allowed to approach the experimenter with any questions. Next, they watched approximately half of the television episode. The tape was stopped at a natural breakpoint about twelve minutes into the show. Subjects were then given the next set of instructions:

- In condition A, subjects ($N = 9$) wrote summaries of what they had seen, and then watched the second half of the episode immediately after the completion of writing.
- In condition B, subjects ($N = 9$) wrote summaries of the first half, but were required to return the following week to complete the experiment. (The data

from a tenth subject, who did not return, were excluded.)

- In condition C, subjects ($N = 2$) viewed the first half, but did not write about it. Instead, they were given someone else's description to read (gathered from a previous session). After reading this, they watched the second half of the video. They were then asked to complete the description of the video, the first half of which they had been given.

All subjects wrote a summary of the second half of the video after viewing it. The viewing and writing instructions for the second half were the same as those for the first half.

At the conclusion of the experiment, subjects were given a handout that outlined the purpose of the investigation. Any questions that were not answered by the handout were answered by the experimenter upon request.

Despite our running the experiment for several weeks, our objective of 30 subjects (10 per condition) was not met because the response rate was low. Since a first look at the subjects' writing samples revealed a much less diverse range of styles than we had expected and a generally poor quality of writing, we decided to discontinue the solicitation of subjects until further investigation had been carried out.

9.6 Analysis of the Data

9.6.1 Tagging the Data

Although subjects in the experiment were instructed to write 500 words in total, their summaries ranged from 476 to 1177 words. Examples of subjects' responses are given by Glover (1996).

Each summary was transcribed into a file. Obvious spelling errors were corrected, but no other changes were made to the writing. In the case of illegible words, the best guess was made. Each word was then tagged with its syntactic category by a *part-of-speech tagger*, a program that determines the appropriate part of speech of each word in an input text. (In the rest of this chapter, the word *tag* will be used metonymously to refer both to the part of speech of a word and to the tag that is used to label it.)

We used two different taggers (on separate copies of each file) in order to compare their results: the POST part-of-speech tagger (Weischedel *et al.*, 1993) and the Brill tagger (Brill, 1992, 1994). Both taggers use the same set of syntactic categories – the *University of Pennsylvania tag set*, but because they use different methods, they might not agree on the appropriate tag for any given word. POST uses a probability model, whereas Brill's tagger uses a learning paradigm called transformation-based error-driven learning.

Despite their different approaches, the two taggers have similar error rates (Brill, 1994). Since neither tagger was retrained on text similar to the writing in the experimental task, it is likely that the error rates in our data are higher than the error rates reported for text similar to the taggers' training data.

9.6.2 Statistical Analysis of the Data

We have now reached the main question of the study. Can we use author-identification techniques to find stylistic inconsistencies within the singly and multiply authored tagged texts?

The data listed at the end of section 9.4.3 were collected from the writing samples:

- word length distribution;
- sentence length: mean, range, and standard deviation;
- percentage of two- and three-letter words;
- part-of-speech distribution (including punctuation);
- relative proportion of each part-of-speech class in sentence-initial and sentence-final position.

Here we report only on the last three of these measures; complete details of the analysis are given by Glover (1996).

9.6.3 Method

Once the stylistic data were collected, the information was analysed statistically to find out whether any of the stylistic tests could be used to match each first half from the writing samples with the corresponding second half. No assumptions of any theoretical frequency distribution (e.g., the Poisson distribution) were used to predict the distribution of the features investigated because few studies have shown such distributions, and the majority of these have investigated function word frequency (e.g., Mosteller & Wallace, 1964), rather than part of speech.

Statistical test used We chose to use the chi-square test for homogeneity, which is appropriate for testing the heterogeneity among a number of different samples, or the likelihood that they were drawn from the same population (Brainerd, 1974). Moreover, the chi-square test is commonly used in stylometric investigations, and we were interested in a new application of existing techniques, rather than in developing a new method of analysis. The null hypothesis is that the sample proportions are equal – that is, both halves were written by the same person, or, at least, are stylistically indistinguishable by the test.

Comparing taggers To see whether the differences between POST and the Brill tagger were large enough to lead to different results in the stylistic tests, we carried out the comparisons of tag frequencies on the separate data from each tagger.

Method of analysis Once obtained, the comparisons were placed in a ranked order of decreasing homogeneity in the features tested. The smaller the value, the greater the homogeneity of the two samples being compared. The more homogeneous the samples are, the more likely it is that they were written by the same person. We also examined the differences between the expected and observed values whenever the chi-square result was surprising (i.e., 'significant') to find out where the source or sources of difference lay.[3]

9.6.4 Results: Matching Pairs

We first examined the matching parts one and two – that is, pairs in which either both parts were written by the same person or part two was written as a conclusion to someone else's part one. These are the cases in which we expected stylistic homogeneity across parts.

Two- and three-letter words In the first test, we compared texts for their ratio

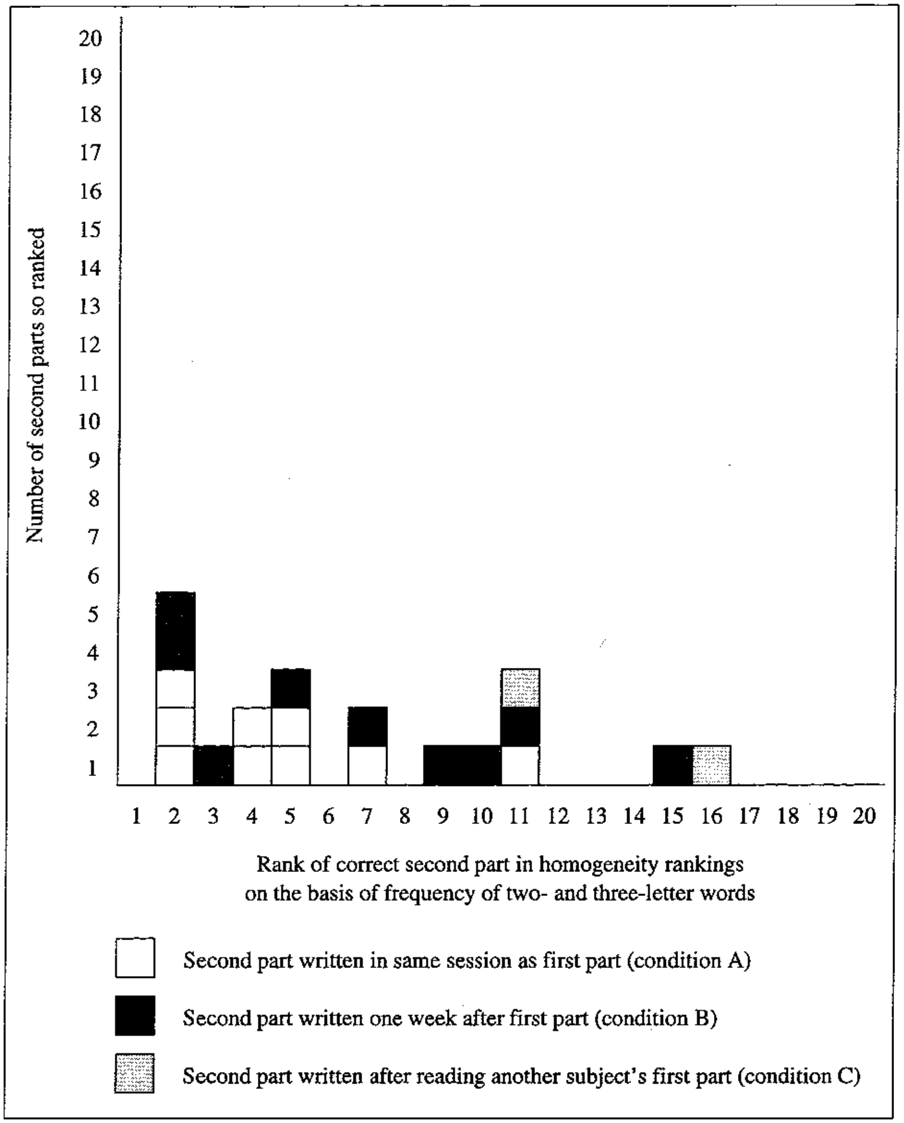

Fig. 9.1 Histogram of homogeneity ranking of each part two with respect to its part one, by comparing the relative frequency of two- and three-letter words to other words. Each block represents one sample of writing.

of two- and three-letter words to other words. Figure 9.1 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.

We found that over half of the parts one and two that matched were ranked in the top five, and fifteen were ranked in the top ten, and no part one and its matching part two were significantly different. Thus, matching parts were consistent by this measure, suggesting that it may be a useful measure of stylistic consistency.

Tags overall In the next two tests, we compared the matching parts one and two of texts for the frequency distribution of the different types of tags over the complete text. In only two cases were the second parts correctly ranked highest. But for half of the texts, the correct part two was ranked in the top four, and only four matches were not ranked in the top ten. So frequency of parts of speech, as indicated by the taggers, seems an accurate means of testing for stylistic consistency.

Interestingly, although the test of tags overall was the most successful method of matching the parts one and two, the significance levels varied more than for any of the other comparisons. At a significance level of 0.05, ten part twos (half of them) were found to be significantly different from their matching part ones (including one written by a different author after reading part one), and five of these (including the one by a different author) demonstrated heterogeneity at the 0.001 level. When we examined these ten texts, we found that only six tag categories were responsible for the differences observed. The most common cause of the differences between part one and part two was that the subjects obviously didn't recall the names of all of the main characters of the story 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*).[4]

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 two 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 problem not uncommon for writers. One of these part twos had been 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, we 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 of the video compared to the second half.

Sentence-initial tags Next, we compared the matching parts one and two of the texts for the frequency distribution of the different types of tags in sentence-initial position. We had five categories in the analysis of sentence-initial parts of speech – determiners, modifiers (adjectives and adverbs), nominals (nouns, proper nouns, and cardinal numbers), pronouns, and conjunctions. In two texts, the same category was absent, and so they could not be compared.

The rankings of the sentence-initial tags were similar to the rankings of the tags from the complete texts. At 0.001 significance, none of the matching parts one and two, including those part twos that completed a part one written by a different author, were significantly different. At the 0.05 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

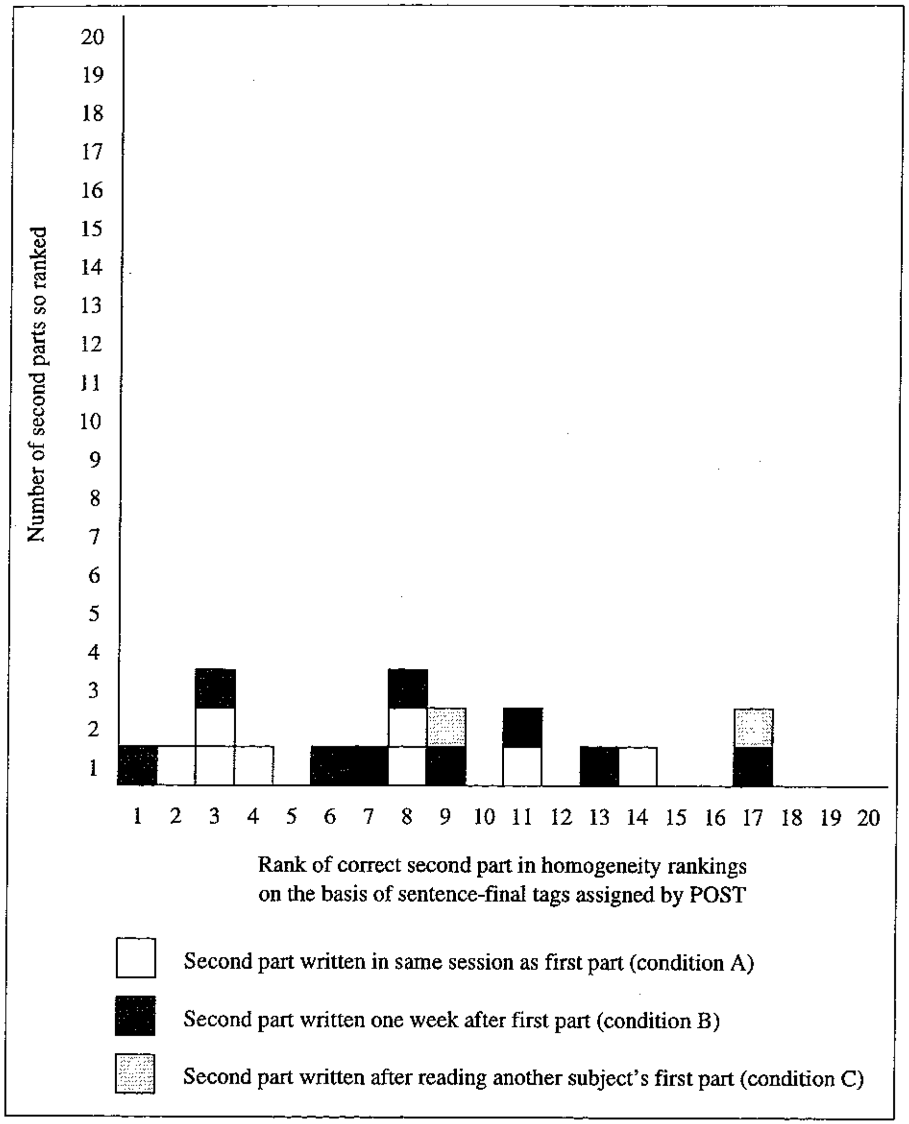

Fig. 9.2 Histogram of homogeneity ranking of each part two with respect to its part one, by comparing the frequency distribution of sentence-final tags, as assigned by the POST tagger. One comparison (in condition A) was not possible.

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.

Sentence-final tags Last, we compared the matching parts one and two of the texts for the frequency distribution of the different types of tags in sentence-final position. We had four categories in the analysis of sentence-final parts of speech – modifiers (adjectives and adverbs), nominals (nouns, proper nouns and cardinal numbers), pronouns, and verbs and particles. There was one matching pair for which no comparison could be made because the expected value of two tags was zero. The rankings were more spread out for the sentence-final tags than for any of the other rankings (see Figure 9.2). Also, the texts that were ranked low were not the same texts that were ranked low in the sentence-initial-tag and tags-overall data. Further, very few comparisons showed statistically significant differences. Combined, these results indicate a lack of variability between pairs. These findings also suggest that the use of sentence-final tags is distinct from the use of sentence-initial tags or all tags. However, further investigation is required before any firm conclusions can be drawn.

Differences between writing conditions Since we had only two writing samples in condition C, it is not possible 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. Overall, second halves that were written a week later did not show 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.

9.6.5 Results: Non-matching Pairs

We next examined paired parts one and two that were **not** written by the same person. 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). Here we discuss significant differences that we found and remark on the consequences of the observation in stylistic support for writers.

Two- and three-letter words Fifty (out of 340) of the chi-squares indicated significant differences at the 0.05 level; only five were significantly different at the 0.001 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 (e.g., conjunctions, pronouns, and prepositions) were associated with high percentages of two- and three-letter words, whereas other parts of speech (e.g., modifiers) 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*) and colloquial (e.g., *kid* rather than *child* or *youth*). In general, then, this test appears to differentiate between a simple and a more descriptive style.

Tags overall Of the 339 possible comparisons, 327 pairs were significantly different at the 0.05 level; 198 were significantly different at the 0.001 level. One comparison could not be made due to an observed frequency of zero for the category 'other punctuation'. According to stylometric theory, writing samples by different people are likely to be significantly different, so these high numbers are not surprising.

A preliminary analysis revealed certain clusters of differences between the pairs, some of which have been observed in the analyses of matching pairs (above). The main patterns of differences were found in preferences for verb tenses and nominals, the number of coordinate conjunctions, and the frequency of commas, sentence-final punctuation (i.e., periods, question marks, and exclamation marks), and other punctuation. The frequencies of certain tags (e.g., possessives, *wh*-words, modals, and subordinate conjunctions) were relatively invariant and their usage rarely or never varied greatly from sample to sample. Perhaps these parts of speech are generally invariant 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 invariant in most texts. Other tags (such as modifiers) occasionally varied a great deal from one sample to the next, but were not often large contributing factors to the differences between writing samples.

The inconsistencies in verb tense, and the differences in nominal usage (due to the avoidance of the use of characters' names) have already been discussed. Although most writers did use tense consistently, there was a higher percentage of verb-tense inconsistencies in the non-matching pairs. 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. These kinds of difference are not difficult to notice when editing texts, but obtaining such information automatically would alert an editor to such problems, thus potentially speeding up and improving the accuracy of copy-editing.

The differences in frequency of periods (full stops) 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. Unfortunately, although wide differences 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). If the different types of sentence-final punctuation could be distinguished,[5] and each type occurred in each sample, this information might be more useful to the user than data about sentence-length variations.

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. The samples usually had many parentheticals and adverbials, which were often placed in the middle of a sentence, thus requiring two commas rather than 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 a lot of 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 is not required for disambiguation). These last two are copy-editing issues, but flagging differences in comma usage may help writers eliminate unnecessary commas and ensure that they are consistent in their comma placement.

Although we had to collapse all punctuation except periods and commas into an 'other punctuation' category, quotation marks were the punctuation marks mainly associated with high usage of other punctuation. Of the significant differences, most were a result of comparisons with two parts that did not have any punctuation except periods and commas. Parts with a high percentage of other punctuation were all characterised 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 or parentheses for parentheticals), and flagging their usage may help writers to increase consistency. In larger text samples, it is more likely that comparisons of individual punctuation marks would be possible.

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 *ands* 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*). The overuse of *and*, run-on sentences, and a lack of variation in sentence structure are generally considered to be deleterious to 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.

Sentence-initial tags Of the 322 possible comparisons of pairs, 67 were significantly different at the 0.05 level; seven comparisons had chi-square values that indicated a lack of homogeneity between the samples at 0.001 significance. In the 67 comparisons that showed lack of homogeneity, the differences were caused almost exclusively by nominal and determiner usage.

Mismatches in nominal and pronominal usage were associated with the problem discussed earlier: the writer of one part knew the name of the protagonist, whereas the writer of the other part did not. Mismatches in determiner and nominal usage were also often due to the same problem.

Although it appears that modifiers did not play a large role in the differences in sentence-initial tags, a second factor that influenced sentence-initial tag occurrences was the placement of adverbial elements, such as adverbs, prepositional phrases, and adverbial clauses. Since many adverbial elements begin with a preposition rather than an adverb, these differences would probably have been observed in the preposition counts if they had been included in the analysis.[6] Since they weren't, the difference showed up only in the preferred sentence-initial tag of the samples that did not use many adverbial elements.

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 that the writer intends, and not all positions are possible for all adverbials (e.g., adverbial clauses cannot be placed sentence-medially), many writers show a definite predisposition towards where they put moveable elements. Information about collaborative writers' preferred adverbial placement might be helpful to them when they are trying to make their documents more consistent.

Finally, related to adverbial placement, but associated with what appears to be a selected style of writing rather than preferred (and probably less conscious) adverbial placement, is the use of a reporting style that emphasises time and location. Writing characterised by such a style had more adverbials throughout the writing, but especially sentence-initially, as this is a salient position in the sentence (e.g., *Next day...*; *Upon returning to the residence...*).

Sentence-final tags Of the 331 possible comparisons, only 18 comparisons showed significant differences at the 0.05 level. No comparisons showed significant differences between parts one and two at the 0.001 level. There were nine non-matching pairs for which no comparisons could be made because the expected value of some tag was zero.

In seven cases, the difference 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 characterised by the use of intransitive verbs and passives, and the placement of prepositional phrases sentence-initially, all of which contributed to having verbs predominate in the final position. It was also the shortest sample, so the variability in sentence structure was probably lower than usual. The samples it differed from were characterised by transitive verbs, absence of the passive voice, prepositional phrases at the end of the sentence, and other optional elements placed sentence-finally (e.g., *now*).

In three cases, the nominal category accounted for the main difference; in one case both the nominal and pronominal did; and in the remaining seven, modifiers and nominals were responsible for the difference. Five of these cases involved comparisons with the same part two, which had many modifiers and relatively few nominals sentence-finally. This part two and the other such samples had a lot of temporal and locative information (e.g., *now, there*), which was most often placed at the end of the sentence.

Since there were few significant differences, and the majority of these involved comparisons with only two of the 40 samples, it is difficult to draw any general conclusions about stylistic inconsistencies that are flagged by sentence-final tags. However, as in the sentence-initial-tag test, preferred adverb placement has an impact.

9.6.6 Discussion

Examination of parts one and two that were written by the same person (or concluded by another person after reading part one) suggests that, when stylistic inconsistencies are detected in a single writer's work, they do not seem to reflect habits of writing, and thus the stylometric assumption that writers' styles are stable is not refuted. Rather, one reason that the differences arose was the writers' initial lack of knowledge (of the characters' names), and their attempts to circumvent this problem (by using noun phrases to refer to them). A second problem was one that many writers have: maintaining consistent verb tenses. Finally, in one case, the use of adjectives changed from one section to the next, influenced perhaps by self-imposed time-pressure and a more action-oriented second half of the story.

Comparisons of parts one and two that were written by different people and that showed statistically significant differences between them revealed a wider variety of stylistic inconsistencies. Some of these differences appear to be due to the individual writers' unconscious preferences, which is what stylometric theory predicts.

At the copy-editing level, inconsistencies in verb tense and noun usage were similar to those found in the comparisons of the matching writing samples. Another such inconsistency was detected in the use of punctuation. Although these inconsistencies are not the type that we set out to find, the automatic detection of such inconsistencies would be helpful to writers, particularly in long, multi-authored documents.

The type of higher-level stylistic inconsistencies that we were looking for were also found: a 'simple' style associated with a high percentage of two- and three-letter words, a syntactically 'boring' style associated with the overuse of coordinate conjunctions, and a 'reporting' style associated with the placement of adverbials at the beginning of the sentence. The placement of optional elements (e.g., commas, adverbials), the use of quotation marks, and preferences for transitive rather than intransitive verbs, were also stylistic differences that were revealed by this analysis.

Comparisons of the samples suggest that they are in some respects relatively homogeneous, a fact that might be due to the writing task. Further investigation is required to answer this question. Of the four measures we used, part-of-speech distribution seems to be the most promising; it revealed the most information about stylistic inconsistencies. Perhaps comparisons of the tags used in sentence-initial and sentence-final position would provide more information in longer samples. Longer samples would be more likely to contain at least one instance of all typical initial and final tags, thus allowing a more complete analysis. Sentence-initial and -final tag patterns provide different information from that of the part-of-speech distribution, since they reveal some of the preferences that writers have as to where to place optional elements in a sentence. Comparisons of 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.

9.7 Conclusion

In this paper, we have analysed the problem of deleterious inconsistencies of style in collaborative writing, and laid out an approach to research on the topic. Our work was intended to be exploratory, and our results barely make a start at solving the problems. We have described an experiment aimed at collecting data for the research, and some of the limitations and problems that arose. We have shown that some stylometric tests can match up different parts of a writer's text fairly well. Also, some of these tests flag inconsistencies that are likely to occur when different sections of a document are written by different people.

We will continue our research by looking at some of the questions raised. How do people perceive documents that stylometric tests flag as stylistically inconsistent? Do significance levels identify stylistic inconsistencies in texts that people also perceive to be stylistically inconsistent? If not, what criteria can we use to decide that perceptible inconsistencies are present? How can we use the results of stylometric tests to give writers advice on how to improve their documents? Does the information about where the inconsistencies are located in the text help writers to merge different writing styles?

Acknowledgements

We are grateful to CBS Incorporated for permission to use the video of 'Kick the Can' in our experiment; to BBN Systems and Technologies (in particular, Ralph Weischedel) and Eric Brill for the use of their part-of-speech taggers; to Tom Bellman for his assistance in running the experiments; to Linda Peto for writing programs for the statistical analysis of the data; to Marilyn Mantei and Marlene Scardamalia for discussions and helpful suggestions; and to Mike Sharples and Thea van der Geest for their comments on an earlier draft of this paper. The preparation of this paper was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Notes

1 Massachusetts General Hospital, Knight Cardiac Catheterization Laboratory (1993). 'Your guide to cardiac catheterization.' Page 1.

2 This chapter is a good example of its own subject matter. The initial sections of the chapter were largely written by Hirst. The later sections, and some parts of the earlier ones, were taken from text originally written by Glover for her thesis (Glover, 1996); they were abridged and edited for both content and style by Hirst, with review and revision by Glover. Although we tried to remove most of the stylistic inconsistencies from the paper, we have allowed a few to remain in order that locating them may be an instructive exercise for the reader.

3 We simply compared the differences between the values to find out which categories contributed a disproportionate amount to the chi-square value, i.e., the largest differences.

4 A few of the sentences were so ridiculous that this lack of knowledge was obvious (e.g., *Ben recognised 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, we had not anticipated such a problem.

5 Neither of the two taggers that was used tags periods, question marks, and exclamation marks with distinguishing tags, but this would not be difficult to do.

6 Two relatively common initial tags, prepositions and verbal elements, had to be left out of the analysis because they did not occur in enough samples. We believe that they would have occurred often enough to be included in the analysis if the sample sizes had been larger, since it was primarily the shorter samples that did not have them. In longer samples, constructions that a writer uses infrequently are more likely to be used.

References

Brainerd, Barron (1974). *Weighing evidence in language and literature: A statistical approach.* Toronto: University of Toronto Press.

Brill, Eric (1992). 'A simple rule-based part of speech tagger.' *Proceedings of the Third Conference on Applied Natural Language Processing, Trento*, 152–5.

Brill, Eric (1994). 'A report of recent progress in transformation-based error-driven learning.' *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), Seattle*, 722–7.

Cluett, Robert (1976). *Prose style and critical reading.* New York: Teachers College Press.

Crystal, David and Davy, Derek (1969). *Investigating English style.* London: Longmans, Green & Co.

DiMarco, Chrysanne and Hirst, Graeme (1993). 'A computational theory of goal-directed style in syntax.' *Computational Linguistics*, 19(3), 451–99.

DiMarco, Chrysanne and Mah, Keith (1994). 'A model of comparative stylistics for machine translation.' *Machine Translation*, 9(1), 21–59.

Dixon, P. and Mannion, D. (1993). 'Goldsmith's periodical essays: A statistical analysis of eleven doubtful cases.' *Literary and Linguistic Computing*, 8(1), 1–19.

Ede, Lisa S. and Lunsford, Andrea A. (1990). *Singular texts/plural authors: Perspectives on collaborative writing.* Carbondale, IL: Southern Illinois University Press.

Enkvist, Nils Erik (1964). 'On defining style: An essay in applied linguistics', in John Walter Spencer (ed.), *Linguistics and style*, 1–56. London: Oxford University Press.

Farkas, D. K. (1985). 'The concept of consistency in writing and editing.' *Journal of Technical Writing and Communication*, 15(4), 353–364.

Flower, Linda S. and Hayes, John R. (1980). 'The dynamics of composing: Making plans and juggling constraints', in Lee W. Gregg and Erwin R. Steinberg (eds), *Cognitive processes in writing*, 31–50. Hillsdale, NJ: Lawrence Erlbaum.

Glover, Angela (1996). *Automatically detecting stylistic inconsistencies in computer-supported collaborative writing.* Master's thesis, Ontario Institute for Studies in Education, February 1996, published as Technical Report CSRI-340, Department of Computer Science, University of Toronto. *ftp://ftp.csri.toronto.edu/csri-technical-reports/340.*

Green, Stephen J. and DiMarco, Chrysanne (1993). 'Stylistic decision making in natural language generation.' *Proceedings, Fourth European Workshop on Natural Language Generation, Pisa, 1993*, 155–8.

Hovy, Eduard Hendrik (1988). *Generating natural language under pragmatic constraints.* Hillsdale, NJ: Lawrence Erlbaum.

Hoyt, Pat and DiMarco, Chrysanne (1994). 'A goal-directed multi-level stylistic analyzer.' *Proceedings, 10th Canadian Conference on Artificial Intelligence, Banff, May 1994*, 23–30.

Irizarry, Estelle (1991). 'One writer, two authors: Resolving the polemic of Latin America's first published novel.' *Literary and Linguistic Computing*, 6(3), 175–9.

Kenny, Anthony (1982). *The computation of style: An introduction to statistics for students of literature and humanities.* Oxford: Pergamon Press.

Makuta-Giluk, Marzena and DiMarco, Chrysanne (1993). 'A computational formalism for syntactic aspects of rhetoric.' *Proceedings, First Conference of the Pacific Association for Computational Linguistics, Vancouver, April 1993*, 63–72.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993). 'Building a large annotated corpus of English: The Penn Treebank.' *Computational Linguistics*, 19(2), 313–30.

McColly, William B. (1987). 'Style and structure in the Middle English poem *Cleanness*.' *Computers and the Humanities*, 21, 169–76.

Milic, Louis Tonko (1967). *A quantitative approach to the style of Jonathan Swift.* Studies in English literature 23. The Hague: Mouton & Co.

Milic, Louis Tonko (1991). 'Progress in stylistics: Theory, statistics, computers.' *Computers and the Humanities* 25, 393–400.

Morton, Andrew Queen (1978). *Literary detection: How to prove authorship and fraud in literature and documents.* Bath: Bowker.

Mosteller, Frederick and Wallace, David L. (1964). *Inference and disputed authorship: The Federalist.* Reading, MA: Addison-Wesley.

Payette, Julie and Hirst, Graeme (1992). 'An intelligent computer-assistant for stylistic instruction.' *Computers and the Humanities*, 26(2), 87–102.

Sanford, Anthony J. and Moxey, Linda M. (1989). 'Language understanding and the cognitive ergonomics of style', in Patrik Holt and Noel Williams (eds), *Computers and writing: Models and tools*, 38–49. Oxford: Intellect.

Smith, M. W. A. (1987). 'The Revenger's Tragedy: The derivation and interpretation of statistical results for resolving disputed authorship.' *Computers and the Humanities*, 21, 21–55.

Smith, M. W. A. (1988). 'The authorship of acts I and II of *Pericles*: A new approach using first words of speeches.' *Computers and the Humanities*, 22, 23–41.

Weischedel, Ralph, Meteer, Marie, Schwartz, Richard, Ramshaw, Lance and Palmucci, Jeff (1993). 'Coping with ambiguity and unknown words through probabilistic models.' *Computational Linguistics*, 19(2), 359–82.