AUTOMATICALLY DETECTING STYLISTIC INCONSISTENCIES IN COMPUTER-SUPPORTED COLLABORATIVE WRITING

by

Angela Dorothy Glover

A thesis submitted in conformity with the requirements
for the Degree of Master of Arts
Graduate Department of Education
University of Toronto

AUTOMATICALLY DETECTING STYLISTIC INCONSISTENCIES IN COMPUTER-SUPPORTED COLLABORATIVE WRITING

Angela Dorothy Glover

Master of Arts 1996

Graduate Department of Education

University of Toronto

Abstract

Collaborative writing is increasingly common in both professional and academic fields. One difficulty that collaborative writers face is trying to produce a consistent style, as each writer may bring a distinctive style to the collaborative writing task. I investigated the viability of using stylostatistical techniques to discover describable, computationally tractable stylistic tests to help collaborative writers eliminate such differences.

Writing samples were collected by having graduate students watch two halves of a television episode, then write a summary of each half. Automatically generated syntactic information was used in statistical analyses to ascertain which halves differed significantly. Examination of the statistically significant results revealed a wide variety of inconsistencies on various levels. Many of these inconsistencies were not immediately obvious before the stylostatistical test results were known. I therefore conclude that stylostatistical techniques provide a promising approach for creating a computer tool to accelerate and improve people's detection of stylistic inconsistencies.

Acknowledgements

Although only one name appears as author of this work, writing a thesis is indeed a collaborative effort. I would like to express my sincere thanks to the many people who made this possible.

To Graeme Hirst for insightful observations, detailed comments, amazing turn-around time, and delightful conversation.

To Marlene Scardamalia for her useful comments and for always squeezing me in to her busy schedule.

To Marilyn Mantei for her assistance in designing and running the experiment.

To Tom Bellman for helping me run the experiment.

To Linda Peto for her thoughtful advice, programming skills and equanimity in the face of seemingly endless re-designing.

To Tahany Gadalla from OISE’s Statistical Consulting Service for statistical advice

To Chris Fung, my liaison officer, for her patience and help.

To the Ontario Institute for Studies in Education and to the Natural Sciences and Engineering Council for financial assistance.

To fellow OISE-ites, Alexandra, Andrew and Seema, who “showed me the ropes”.

To my friends who provided sympathetic ears, well-timed distractions and a cheering section.

To Chris Beck whose offer to “read a draft” got him more than he bargained for.

To my family who showed their unwavering support in so many ways—both large and small.

Table of Contents

1. Introduction

Computers have changed the way people write. Word processors have become the preferred writing tool in many offices, classrooms and homes. The popularity of the word processor can be attributed to the fact that it makes certain aspects of writing easier. However, it by no means solves all of the difficulties involved in writing, and indeed exacerbates some, as well as creating new ones. Hence, the proliferation of writing software, as designers attempt to support writers in more, and better, ways. To ensure the development of software that people will use, writing practices are being carefully examined to determine what types of tools could most benefit writers. One practice that has received increasing attention in recent years is *collaborative writing*. The current impetus to produce collaborative writing software comes from the desire to support a writing practice that has only recently begun to be regarded as prevalent and legitimate.

Since people tend to bring their individual practices to collaborative work, writers accustomed to using a word processor when writing alone will doubtless want to use one when writing with others. Although writers can (and do) collaborate using software designed for individual writers, collaborative writing tools would allow them to bring not only their preferred writing practices, but also their preferred collaboration practices to the collaborative writing situation. For writers who work collaboratively at a distance, a practice which is increasing among academics and within large corporations (Sharples, 1993), computer support for collaborative writing is crucial. The ultimate goal of computer support for collaborative writing, therefore, is to make it possible for writers to go about their task even when they are unable to physically meet. To achieve this goal, designers of collaborative writing support software need to be concerned not only with facilitating particular practices that occur only when writers collaborate, but also with alleviating particular difficulties that arise due to the collaborative situation.

One such difficulty is the merging of individual writing styles[1] to produce a multi-authored document that has a single voice. Although collaborative writers might individually produce exemplary pieces of writing to contribute to their collaborative document, the pieces might be *stylistically inconsistent*. An example of stylistic inconsistency can be seen in the following sentence, which is from a brochure given to hospital patients before they undergo cardiac catheterization.[2]

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

Clause 1 and, to a slightly lesser extent, clause 3 are in medical language, as if in a formal communication between physicians, whereas clause 2 is much less formal, and is expressed in ordinary lay language. The effect of these two styles mixed together in one sentence is a feeling of incongruity—which was presumably not intended by the authors of the brochure. 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. Further, although people might notice that something is wrong when

[1] *Style* has a number of different meanings, and thus the term *writing style* tends to be used in several ways: the method by which someone carries out a writing task; formatting guidelines (e.g., APA style); literary style. In my discussion of style, I am referring to the author's choice of words and syntactic constructions that gives a piece of writing its particular "feel". A more formal defunition will be given in section 2.4.10.

[2] The parenthesized numbers are mine, to refer to the individual clauses.

[3] Massachusetts General Hospital, Knight Cardiac Catheterization Laboratory (1993). "Your guide to cardiac catheterization." Page 1.

reading a stylistically inconsistent document, they are often unable to articulate exactly where the problems lie. If writers or editors are unable to perceive why a document is inconsistent, they will not know how to go about solving the problem.

The ultimate goal of this research is to build software that will help with the problem of stylistic inconsistency. There are two steps to this task: first, the system must identify stylistic inconsistencies in a document; second, the findings must be presented and changes suggested in a manner that is easily comprehensible to the average user. My immediate goal is to investigate the viability of the first step using a method adapted from *stylostatistics* that might help writers identify, and therefore more easily solve, instances of stylistic inconsistency. To accomplish this goal, I first gather writing samples by carrying out an experimental task in which subjects produce documents written in two parts. By pairing each first part with each of the second parts, I construct a set of "collaborative" documents, with possible stylistic inconsistencies, that are controlled for content. Next, I analyze these samples using relevant stylostatistical tests that can currently be carried out on text that has been processed automatically. I then manually examine the results of these tests to determine what they mean qualitatively. Finally, I conclude with suggestions for future work that might help answer the many questions raised by this exploratory study.

The contributions of this thesis are:

1) The application of stylostatistical techniques to a new problem: the identification of stylistically inconsistent writing.

2) The linguistic interpretation of the results of the stylostatistical tests.

3) The demonstration that the results of stylostatistical tests provide information about writing style that is potentially more useful to writers than the advice from existing style checkers.

2. Collaborative writing

2.1 Why the interest in collaborative writing software?

Two trends have provided the impetus for the recent interest in the development of computer support for collaborative writing: collaborative writing is becoming increasingly common both in the workplace and in the classroom (Duin, 1991); and writers are relying more and more on computer technology to support their writing practices (Dorner, 1992).

Although published work would seem to indicate that most writing is produced by a single author, a belief reflected in much of our talk about writing, surveys indicate that collaborative writing occurs far more frequently than most people think it does—in many workplaces, in diverse occupations, involving a wide range of writing activities (Ede & Lunsford, 1990; Couture & Rymer, 1991; Beck, 1993). For example, Ede and Lunsford (1990) found that 87 percent of their survey respondents, approximately 700 members from seven large professional organizations in the United States, reported writing as members of a group or team at some time on the job. Some of these people collaborated on almost every document they produced in the workplace. Collaborative documents ran the gamut of length and formality from memos to published books.

Collaborative writing has also been gaining in popularity as a method of instruction at all levels of the educational system over the last ten years (Forman, 1992) for a variety of reasons. As well as being considered socially beneficial, collaborative writing and other collaborative activities may benefit students cognitively because they encourage students to reason overtly about their mental activities (Brown & Campione, 1990). Further, although schools primarily view intelligence as an individual's possession, cognitive research on *distributed cognition* (see section 2.2.2) is expanding the school's notion of intelligence, where it resides and how to foster it (e.g., Brown & Campione, 1990; Scardamalia & Bereiter, 1994). Finally, collaborative writing experience appears to be valuable preparation for future employment, given the above findings.

2.2 Problems with computer-supported collaborative writing tools

As reliance on technology grows in the workplace and in the schools, it becomes even more important for system designers to be aware of occupational and educational practices so that they can provide appropriate support to help people work and learn more effectively. Despite the general enthusiasm for the word processor, other writing tools, such as grammar checkers, are not widely used by writers (Rimmershaw, 1992). Collaborative writing tools are particularly underused. So, although the impact of computer technology on writing has been significant, it has not been as influential and helpful as it promises to be, especially for collaborative writers. The lack of success of many computer writing tools among word processor converts indicates that there are some serious design flaws in much of the writing software. I will review some research that suggests where the main problems lie in the currently available collaborative writing software.

Attempts to provide collaborative writers with computer support began more than twenty years ago and development has received a marked increase in attention over the last ten years, but synchronous, remote on-line collaborative writing remains a rarity. Investigations of writing in the workplace indicate that collaborative writers use the computer primarily for: email communication; sharing texts on-line; sharing templates, style sheets etc.; and retrieving information from databases (Van Pelt & Gillam, 1991). When authors do use available collaborative technology to write together, it is usually out of interest in exploring new media, rather than because they believe that the system will aid their collaborative writing (Newman & Newman, 1992). The low uptake of collaborative writing software points to an obvious need to rethink the development of collaborative writing tools to ensure that the next generation of software will be embraced by joint authors.

Why has collaborative writing software been so pointedly ignored by collaborative writers? Given the size and diversity of the collaborative writing population, and individual writers' increasing reliance on the word processor, the reasons appear to be associated with the software itself, rather than with the users. Despite the recent proliferation of group software, most attempts to use it outside of experimental situations fail because collaborative writing is not well supported by current computer-based aids (Duin, 1991). Many of these writing aids embody assumptions about collaborative writing that do not necessarily reflect the reality of the collaborative experience, forcing writers into some practices that are at worst harmful, at best useless (Rimmershaw, 1992), demand too much effort from the user (Sharples, Plowman & Goodlet, 1993) and often have unexpected side effects (Beck, 1992).

Duin's (1991) review of collaborative writing software indicates that much of it has primarily focused on providing on-line access to other writers and their writing, rather than providing writing tools. In Posner's (1991) review of six computer-supported writing systems (Aspects, GROVE, ForComment, PREP, Quilt and ShrEdit), she observed that none of the systems comprehensively supported collaborative writing, and that almost none of them provided what she considered "good" support for thirteen requirements for collaborative writing that she identified. Beck (1992) criticized three existing collaborative writing software systems (ShrEdit, PREP, and CoAuthor) on the grounds that they limit the user to a single way of writing collaboratively. Sharples et al. (1993) found that two existing software systems designed to help collaborative writers (The Coordinator and Quilt) require too much effort on the user's part, and are not well integrated with computer practices. Despite the fact that the main thrust of most current CSCW is to provide more flexible systems to support different approaches to collaborative writing (e.g., Neuwirth, Kaufer, Chandhok, & Morris, 1994), Sharples et al. (1993) discovered in their evaluation of four systems (ShrEdit, StorySpace, PREP, and MUCH) that few of the systems are meeting their stated goals, and they question the issues these systems address.

Although the central issue in studying computers and writing is finding the relationship between the writing process and technology designed to support it (Holt, 1992), these software reviews indicate that few collaborative writing software designers appear to have a solid understanding of this relationship. There are four significant and interrelated problems that have hindered them in this respect: the lack of a common definition of collaborative writing; the need for a model of collaborative writing; the paucity of research on collaborative writing; and the difficulty of testing collaborative writing software. I will discuss each one, with reference to its impact on the development of collaborative writing software.

2.2.1 A matter of definition

There is no single, commonly accepted definition of collaborative writing (Harris, 1994). Proposed definitions are often vague, because collaborative writing is difficult to describe since its boundaries are fuzzy. In order to get as much information as they can, researchers who conduct surveys of collaborative writing usually choose to make their definition broad, thus encouraging respondents to discuss approaches to writing which might not be perceived as methods of collaborative writing by all people. For this reason, Ede and Lunsford (1990) used the following definition in their survey about writing in the workplace: "any writing done in collaboration with one or more persons" (p. 15).

Rimmershaw (1992), on the other hand, generated her all-encompassing (and poetic) definition of collaborative writing *after* conducting her interviews of academic collaborative writers, because she noticed how widely their personal definitions varied:

> any piece of writing,
> published or unpublished,
> ascribed or anonymous,
> to which more than one person has contributed,
> whether or not they grasped a pen,
> tapped a keyboard,
> or shuffled a mouse. (p. 16)

Many researchers believe it is important to make a distinction between at least two types of collaborative writing. Van Pelt and Gillam (1991), for example, distinguish between *team-work collaboration* in which an author receives input from a team, but retains authority over the document, and *shared-document collaboration* in which authors share responsibility for all important decisions made during the writing of the document. Shared-document collaboration is closer to what most people think of when they refer to collaborative writing, and is rarer than team work. Team-work collaboration encompasses almost all writing, or at least most published writing. Although other researchers use different terms to describe these two types of collaborative writing, the distinction made is similar (e.g., Couture & Rymer, 1991).

Acknowledging the fact that most writers do not work entirely alone emphasizes the ubiquity of collaborative writing, despite the fact that it is often not accredited in the final product, thus revealing a practice that is “hidden in plain sight” (Ede & Lunsford 1990). However, there are two risks Ede and Lunsford (1990) identify that may result from defining collaborative writing in a broad way. First, the definition may fail to distinguish writing from other intellectual activities, because it is difficult to pinpoint where working together ends and writing together begins. For software developers of computer support for collaborative writing, however, distinguishing writing from other intellectual activities might not be an important concern, but rather, what is reasonable and possible to support will drive development. Second, a broad definition may conflate individual writing and collaborative writing because the boundary between writing alone and writing collaboratively is not clear. However, a good collaborative writing tool should support both individual and group writing, for two reasons: first, such systems will allow writers to write alone or with others, without having to switch tools; second, Posner (1991) has identified the “single-writer strategy” (see section 2.3) as one of the ways in which collaborative groups write.

Interestingly, although the definition of collaborative writing forms a significant part of the discussion in the survey and literary research on collaborative writing practices, one never encounters an explicit definition of it in the collaborative writing software research. The issue of definition is not irrelevant for software designers, however. Although a broad definition should have a positive, rather than a negative, influence on software development, the lack of a common definition of collaborative writing may be affecting collaborative writing systems in detrimental ways. First, some designers only look at one aspect of collaborative writing, rather than at the whole picture. Beck (1992) found that although collaborative writing is a complex task that can be carried out in a myriad of ways, most current collaborative writing software developers have implemented a single view of what collaborative writing is and what constitutes support for collaborative writing, thus restricting users unnecessarily. Beck believes that a broader view of what factors might affect collaborative writing will lead to a better understanding of what kinds of computer aid joint authors could benefit from. Second, erroneous conclusions about what types of support collaborative writers need may be drawn from incorrect assumptions about what is meant by collaborative writing. Couture and Rymer (1991) point out that the loose way in which the term *collaborative writing* is used may be causing some researchers to overestimate the amount of shared-document collaboration that is actually being done in the workplace. In their survey of 400 professionals from a wide range of organizations, they found that although 78 percent of respondents said they *sometimes* or more often get feedback on their writing, only 24 percent contributed to a multi-authored document *sometimes, often,* or *very often*. Overestimation of shared-document collaboration seems to have been the major contributing factor to Sharples et al.'s (1993) finding that some of the features that are popularly offered by current software are not in high demand by collaborative writers. These features, such as simultaneous editing capabilities, are mainly ones associated with shared-document collaboration. Although an ideal system will support all collaborative writing practices, Sharples et al. (1993) recommend working from a principle of supporting common existing practices, to ensure a useful tool for collaborative writers.

If software designers do not pay attention to how collaborative writing is defined, they may make assumptions about what collaborative writing practices need to be supported on the basis of their own, possibly narrow, views of what collaborative writing is. To avoid producing useless or underused tools, designers need to be aware of what collaborative writers, researchers, and they themselves mean when they use the term *collaborative writing*.

2.2.2 Towards a sociocognitive model of writing

The second contributing factor to the problems surrounding collaborative writing systems is the absence of a comprehensive *model* of collaborative writing. Sharples (1991) believes that many of the problems associated with collaborative writing systems may largely be due to the fact that none of these systems is founded on a well-formulated model of the writing process, collaborative or otherwise. As Sharples and Pemberton (1992) point out, a writing tool is constrained by the assumptions and limitations of the model, whether explicit or implicit, on which it is based. Although a common model of collaborative writing has yet to be established, cognitive research on writing provides a foundation for such a model.

The scientific study of writing is little more than twenty years old. Although there has been a great deal of research interest since the early 1970's (Hayes & Flower, 1987), writing research is still relatively new. Beginning in the 1970's, writing researchers began to challenge commonly held assumptions about writing by examining the processes writers actually went through in their attempts to translate their thoughts into writing. There was a major shift in writing research away from the analysis of the quality of the written product, towards an understanding of the process of writing, and the relationship between this process and the written product (Cochran-Smith, 1991). Previously, writing models were based on the end product—the text. Two such product-based models that dominated the "folk theory" of writing have been largely debunked by cognitive research: the muse-inspired writing of the professional artist-writer, and the stage model of writing taught by pedagogues according to the cookbook method (Flower & Hayes, 1980). In contrast, cognitive researchers today view writing as a dynamic process. Flower and Hayes (1980) describe writers in the act as thinkers on "full-time cognitive overload" (p. 33), juggling knowledge, language, and rhetorical constraints as they try to achieve their writing goals.

Flower and Hayes's (1980) research on writing has been very influential, despite several limitations (see Hartley, 1991). They investigated the cognitive processes of writing using a technique called *protocol analysis*. Protocol analysis involves asking and prompting subjects to externalize their thoughts while they are performing a cognitive task, in this case, writing. The data that is collected is the verbatim report of what the subjects said (including pauses, etc.), along with the text and notes they wrote during the experiment. The transcript itself is the protocol. The protocol and writing are then examined to find out what kinds of cognitive processes appear to have been used by the writer, and a model of these processes is inferred. The writing model that Hayes and Flower (1987) developed from their work has been widely used by researchers and educators.

Flower and Hayes's research suggests that there are several critical features of the writing process: writing is goal-directed; these goals are hierarchically organized; and writers accomplish their goals using three major processes. In their 1987 paper, they refer to these processes as *planning*, *sentence generation*, and *revision*.[4] Planning involves the retrieval and shaping of knowledge to fit the textual and audience constraints. Sentence generation is what we often think of as writing—putting ideas on paper in coherent sentences. During revision, the writer evaluates and attempts to improve the draft. These processes are similar to the three components (pre-writing, writing, and re-writing) of the stage model, but what distinguishes their model of writing is the recognition that these processes are interleaved—sometimes iteratively, sometimes recursively. The interaction of these processes provides some explanation for the variety of strategies used by writers, as well as for certain problems writers encounter as they go about their task.

Another influential model, particularly in the educational domain, is Bereiter and Scardamalia's (1987), which addresses one of the limitations of the Flower and Hayes model—it accounts for the difference between novice and expert writers. Rather than proposing a single writing model, they suggest that expert writers actually have access to an additional composing process. Their *knowledge telling* model, which is similar to Flower and Hayes's model, accounts for a process that relies mainly on skills gained in everyday interaction, whereas their *knowledge transforming* model accounts for a more studied ability that involves deliberate control over parts of the writing process that are not attended to in knowledge telling. While novice writers rely only on knowledge telling, expert writers have access to both writing processes.

For researchers interested in collaborative writing, however, there is a serious limitation to these cognitive models: they describe only the single writer writing alone (Hartley, 1991). There exist, as yet, no widely accepted models of collaborative writing, cognitive or otherwise, although attempts to fill this gap are being made (e.g., Rose, 1994). My experience leads me to believe (and this is implicit in much of the computer-supported collaborative writing literature), that the single-writer model applies equally well to multiple writers in many respects. People tend to bring their methods of doing things as individuals to their work as part of a group. When people write together, however, the cognitive processes of writing cannot be considered apart from the social processes of collaboration (Sharples et al., 1993). The importance of this point is highlighted by the evolving perception in cognitive psychology that cognitions are situated and distributed, and that social and other situational factors do not simply affect cognition, but should be treated as cognitions (Salomon, 1993). In this constructivist point of view, intelligence is regarded as something which is distributed among the participants and the technologies they use to accomplish a given task in a particular environment. Since the distributed cognitive abilities may be greater than the sum of the individual parts, the system must be examined as a whole, rather than simply looking at components in isolation (Salomon, 1993). In addition to the primary interest in how cognitions interact, one focus of much of this research is on developing enabling technologies that can be embedded within a social context to facilitate collaborative work that is aimed at advancing knowledge (e.g., Scardamalia and Bereiter's (1994) *knowledge-building society*; Brown and Campione's (1990) *community of learners*). Certainly, such a goal is (or should be) shared by developers of collaborative writing software.

Although no well-defined model of distributed cognition has yet been developed, cognitive models provide software developers with important information about the mental processes involved in writing, and research on distributed cognition presents a broader view of how people think. Using these theoretical underpinnings as a starting point will make clearer the assumptions and limitations of the tools developed, allowing better evaluation and subsequent development.

[4] In their earlier work, they called these processes *planning*, *translation*, and *reviewing* (Flower and Hayes, 1980). I do not know why they chose to change their terminology.

2.2.3 Collaborative writing research

Beck (1992) attributes problems with collaborative writing software primarily to the general lack of research on collaborative writing. The main reason for this lack is the relative recency of research interest on collaborative writing. Although collaborative writing is currently the subject of much attention in a wide variety of disciplines, research on this topic is even newer than cognitive writing research; it was not until the period of 1982–1987 that academics began to seriously investigate collaborative writing (Batschelet, Karis & Trzyna, 1991). I will briefly summarize the current state of collaborative writing research, then suggest the next step in the research agenda.

Although the research interest is new, writing together is not a recent phenomenon (Sharples, 1993). Surveys on collaborative writing (e.g., Beck, 1993; Couture & Rymer, 1991; Ede & Lunsford, 1990) have actually brought to light (and thereby given credibility to) a widespread practice that has been largely ignored, not only by researchers, but also by society in general. Interviews (e.g., Rimmershaw, 1992; Posner, 1991; Ede & Lunsford, 1990) have provided further insight into the complex ways in which collaborative writers carry out their writing tasks. However, although this body of research has contributed information about general trends involved in collaborative writing practices, the fact that collaborative writing research is relatively new and small, and that the object of investigation is immense, has led to more questions being raised than answered about collaborative writing. Now that the general nature of collaborative writing has been explored, a different type of research is required to find out how the writing process is adapted and negotiated by collaborative writers in order to successfully write together: the case study. Case studies supply more specific information about practices than is generally gained in surveys and interviews. Since social interactions among collaborative writers and available technological support will affect the writing process, studying collaborative writing groups at work is the best way to discover their actual practices. Unlike writing alone, the very nature of collaborative writing forces writers to externalize their writing processes to a greater degree as they communicate about their project, thus providing the researcher with explicit information on the writing process (Plowman, 1993). Also, case studies may reveal details about collaborative writing that surveys and interviews miss because the respondents do not realize the importance of them, and therefore, fail to report them.

There are several drawbacks to conducting case studies. They are both time-consuming and labour-intensive (Sharples et al., 1993). In addition, a large number of studies is required to ensure that a full range of collaborative writing practices is explored, because the context and the purpose of the writing, as well as the make-up of the group, will have a significant impact on the strategies used. However, case studies promise to provide more detailed information about collaborative writing practices which software designers can better incorporate into collaborative writing systems.

2.2.4 Testing collaborative writing software

Duin (1991) suggests that software developers have underestimated the difficulty of evaluating collaborative writing software. Baecker, Nastos, Posner and Mawby (1993), Sharples (1992), and Duin (1991) all agree that *user-centred iterative design* is the best way to test collaborative writing software. *User-centred* implies the prototype is based on behavioural research. The iterative design consists of an iterative cycle of design, implementation, and evaluation by collaborative writers, which is repeated until a satisfactory product is produced (Baecker et al., 1993).

Reference to Baecker et al.'s (1993) work demonstrates the efficacy of the user-centred iterative design process in eliminating some of the problems associated with collaborative writing software design. After reviewing the usability studies at each stage in the development of their system, SASSE, Baecker et al. were forced to seriously reevaluate the types of computer assistance they intended to develop. Support for some practices that they initially thought were important were dropped, whereas work on supporting previously ignored practices shaped their later design.

There are several drawbacks to iterative design, however. It is costly and time-consuming. The evaluation phase is often not possible to conduct in the workplace, forcing developers to rely on experimental rather than actual writing situations (Duin, 1991). Finally, due to the immense number of variables associated with writing, using software, and especially group interaction, Duin (1991) believes that a good product can never be guaranteed, even with extensive iterative testing.

2.2.5 Conclusion

The problem of defining collaborative writing, the need for a sociocognitive model of how people write together, the lack of research on collaborative writing and the difficulty of evaluating collaborative writing software are interrelated issues that have contributed to the poor design of most current collaborative writing software. Much work remains to be done before these issues are resolved and comprehensive technology which is embedded in a broader social context can be designed to support collaborative writers.

Such work is beyond the scope of this thesis, however. Instead, I will focus on how software developers can proceed with the design of collaborative writing software given recent research in information technology and what we now know about collaborative writing. In particular, I am interested in scaffolding tools for the facilitation of practices and the alleviation of difficulties that occur in collaborative situations.

2.3 A taxonomy of collaborative writing practices

Given the wide variety of collaborative writing practices, a systematic way to implement support for these practices is required. Posner (1991) has developed a taxonomy of collaborative writing practices that is based on the research of others as well as on her own interviews of ten people who described a total of twenty-two collaborative writing experiences in detail. Although not deeply rooted in formal theory, it is a useful guide for software designers for two reasons. First, it provides a vocabulary to describe what people do when they write collaboratively. As Lunsford and Ede (1986) point out, a vocabulary is important because “what we do not have a name for, we simply do not recognize.” (p. 74). Second, it presents the various ways in which collaborative authors write, in a concise way, making clear some of the implications these practices have for collaborative writing software design. I will outline Posner’s taxonomy, then discuss some implications for collaborative writing software design.

Posner identified four roles which participants in a collaborative writing group may take: *writer*, *consultant*, *editor* and *reviewer*. Although the functions of these roles are generally agreed upon, the kinds of changes to the document that are permitted within these roles is often defined by the group. For example, the editor’s role may range from being allowed to only correct typographical, spelling, and syntax errors, to being permitted to rework the document extensively.

Posner distinguished six collaborative writing activities:

- *brainstorming* (generating ideas)
- *researching* (gathering information)
- *planning* (defining and dividing the work)
- *writing* (transforming the ideas into text)
- *editing* (making changes to the text)
- *reviewing* (commenting on the document).

She found that these different activities occurred in various combinations and sequences during a collaborative writing project. This observation is similar to Hayes and Flower's (1987) description of the iterative and recursive nature of individual writing.

Posner classified four different document control methods used by collaborative writers: *centralized*, *relay*, *independent*, and *shared*. Centralized control involves one person being responsible for the document, while others play a more peripheral role. Centralized control can be maintained throughout the document production, or can occur in the final stages, when one person takes what has been written and integrates it. Relay control indicates that document control is passed from writer to writer during the document development. Independent control is the partitioning of a document among the writers, each of whom control a different portion. Shared control means that two or more people jointly control the document throughout the writing process.

Finally, Posner recognized four collaborative writing strategies: *single writer*, *scribe*, *separate writers* and *joint writing*. The single-writer strategy is similar to the individual writer's strategy. Only one person writes, while other members of the group provide support for activities other than the actual writing. The scribe strategy involves the group discussing the work, with one person in charge of recording the writing which evolves. The separate-writer strategy entails dividing the document into sections, so the writers can work on different portions of the document in parallel. In the joint-writing strategy, two or more writers compose the text together. Any of these strategies may be augmented by the assistance of one or more consultants, who may make stylistic recommendations, structural suggestions, editorial comments and/or contribute domain-specific expertise.

Since there are many methods that collaborative writers use to accomplish their tasks, any collaborative writing system would ideally support a range of writing practices, allowing users to select their preferred strategy for each collaborative project undertaken. Reference to Posner's taxonomy illustrates specific kinds of assistance that become necessary when collaborative writers work together in certain ways. Furthermore, Posner noted that during the writing of a document, the roles, control method, and/or writing strategy may change. Therefore, it is also important that computer software allows a group the flexibility to redefine their mode of writing at any time.

2.4 Support for the collaborative writer

Flower and Hayes (1980) have identified various types of constraints that writers must contend with in their attempts to produce a written document. Word processors and other types of computer writing technology, such as spelling checkers and idea planners, have primarily been developed to lessen the effects these constraints have on writers. The better the support that technology can supply writers with, by allowing them to off-load difficult or error-prone cognitive burdens, the more writers' minds are freed to concentrate on the writing process. Ultimately, this will lead to both better writing and more satisfied writers. When writers collaborate, some of these constraints are distributed (e.g., background knowledge), but at the same time new constraints are added (e.g., need for consensus). In addition, certain problems associated with writing become more difficult to handle because of the collaborative situation. There are also new practices that people adopt when they write together. Developers of collaborative writing software could help collaborative writers by providing support for these new practices and problems. I will discuss ten requirements for a collaborative writing system that would allow writers to choose their preferred practices, and provide assistance for problems associated with collaborative writing: communication, synchronous writing, roles, collaborator identity, annotation, version control, software and hardware compatibility, a global perspective, format consistency, and style consistency.

2.4.1 Communication

The most obvious way in which collaborative writing differs from individual writing is in the need for members of a group to communicate with one another. Optimally, collaborative writing involves face-to-face meetings supplemented by a range of communication channels, spoken and written, synchronous and asynchronous, so that collaborative writers can choose the most suitable channel for particular tasks within the writing process (Kraut, Galegher, Fish & Chalfonte, 1992; Sharples et al., 1993). Since the ultimate goal of collaborative writing software is to break down the physical distance between potential collaborative writers, it should not only support the collaborative task, but also offer appropriate means of communication for accomplishing the task. Ideally, this goal involves identifying and providing support for communicating in ways not currently easy, or even possible, in addition to facilitating customary forms of communication (Pea, 1994; Scardamalia & Bereiter, 1994).

Computer support for collaborative writing is crucial for writers who are separated by distance, yet wish to use either the scribe or joint-writer strategy of writing. Both of these strategies require the participants to write together, in contrast to the single-writer and separate-writers strategies. However, all potential and actual collaborative writers could benefit from a range of communication tools. Kraut, Egido and Galegher's (1990) study of the effects of proximity on group work found that maintaining social cohesion through frequent, high-quality, low-cost communication was an essential factor in both initiating and sustaining collaborative relationships. They point out the need for communication tools which facilitate both planned and unplanned, synchronous and asynchronous contacts among collaborators, be it of writing or any other group activity. By high-quality, they mean that any information which needs to be communicated can be transferred quickly and accurately. By low-cost, they mean the low behavioural cost to the users; the communication should not require a planned effort to use it. For planned interaction, collaborators should be able to choose the most appropriate means of communication. Any restrictions on avenues of communication will limit the collaboration. Given the rich referential field of face-to-face interaction (e.g., facial expressions, gestures to physical objects, external representations, etc.), a highly interactive multimedia environment is imperative (Pea, 1994).

2.4.2 Synchronous writing

Collaborative writing systems should allow both synchronous and asynchronous writing. Posner (1991) recommends synchronous writing support for jointly authored documents because her interviews revealed that collaborators did sometimes work on a section of a document simultaneously, despite lack of support for such a strategy. Synchronous writing capabilities also support asynchronous, distributed writing. If writers share control of the document, there may be times when their writing occurs simultaneously although they do not intend this, particularly if there are many writers, or if a deadline is near. Thus, software which does not support synchronous writing may restrict asynchronous writing by preventing access to a document already in use by a collaborator.

2.4.3 Roles

Posner (1991) identified four different roles which members of the collaborative writing team might play (see section 2.3, above). To support these roles, she suggests that writing software should make the roles explicit. For example, writers would have unrestricted access to the document, consultants would have read-only access and reviewers would have read and comment access. Depending on how much leeway the writers allowed them, editors might have full access, or be restricted to commenting. Such clearly defined roles would protect the document from revision by unwanted sources. Sharples et al. (1993) raise the concern, however, that explicit roles may be too rigid, since roles may be assigned prematurely, before the collaborative writing process has really begun, or roles may change as the writing develops. Flexibility in reassigning roles would avoid these potential problems.

2.4.4 Collaborator identity

Preservation within the text of the identity of the writer(s) of each piece would provide collaborators with important information, particularly when using relay or shared document control. Knowing who wrote what may assist writers in resolving questions or disputes that may arise with regard to content.

2.4.5 Annotation

Consultants, editors and reviewers, who work independently of the writer, would be better supported if they could insert editing marks on the electronic document and have access to an on-line comment section for same-page annotation, particularly if they are not permitted to alter the text in any permanent way. Representational editing tools would allow them to perform the traditional type of editing and annotation done on proofs, but would eliminate the need for transferring a hard copy back and forth, and ease integration of the accepted changes into the document. The tools would allow writers to see the original text and suggested revisions closely linked, thus reducing the ambiguity of the suggested changes. Such tools would also offer an additional mode of communication for all of the writers involved in the document production.

2.4.6 Version control

As writers make changes to the document without consulting their collaborators, the difficulty of merging these revisions and/or recovering deleted material may quickly escalate into a serious problem. Version control would help avoid wasted time and lost text engendered by distributed writing. Two collaborative writing practices for which version control is essential are the joint-writing strategy and the relay-control method (Posner, 1991).

2.4.7 Software and hardware compatibility

Incompatible software or hardware may have a deleterious effect on collaborative writers' practices. Rimmershaw's (1992) interviews of collaborative writers revealed that incompatible word processors affected decisions which were made about document control, writing strategies, and writing roles. Writers are sometimes forced to write collaboratively in ways they would prefer not to. The development of distributed computer-supported collaborative writing systems would eliminate incompatibility problems, as all users would have access to the same document.

2.4.8 A global perspective

Although word processors have facilitated writing in many ways, there are some drawbacks to the technology. One problem that is gaining attention is the difficulty of achieving a global perspective of the text (Severinson Eklundh, 1992). Particularly when producing long documents, writers often require access to overviews to guide their writing. In a longitudinal survey of seventy writers, which Severinson Eklundh and a colleague conducted (cited in Severinson Eklundh, 1992), approximately three-quarters of the respondents said that they had difficulty getting a global view of the text when using a word processor. Although it can also be difficult to gain an overview when using paper copy, there are two factors which exacerbate this problem on the word processor (Severinson Eklundh, 1992). First, the standard word processor generally allows a restricted view of the text—not even a full page can be viewed at one time, let alone several. Second, due to the scrolling, the user has no spatial information on which to rely when navigating through the text. These factors affect the ongoing evaluation of the text as a whole. Although writers using a word processor tend to revise more than when writing on paper, this increase is due mainly to additional local, low-level revision, whereas they tend to postpone global revision until later in the writing process. This leads to less coherent text, especially as the text grows longer. These problems are compounded when the text is a group effort—particularly when the joint writing strategy is used, and the document control method is relay or shared. Extra working time and frustration may be avoided by providing collaborative writers with access to a global perspective of the joint document as it develops. Severinson Eklundh (1992) makes several suggestions as to how designers of computer-based writing systems may help users access a global view of the evolving text: provide more than one type of overview of the document; make overviews active, thus allowing users to move text and move around in text using the overview; allow writers to move easily from one representation to another.

2.4.9 Format consistency

Inconsistency of format might not confuse meaning, but it may distract readers, thereby interfering with comprehension. It might also diminish the credibility of a document, its author and its publisher; inconsistency of format suggests that the document was produced in haste or with a lack of care (Farkas, 1985). Maintaining consistent spelling, punctuation, font style, and layout (particularly in figures and tables) throughout a document can be time-consuming and requires extra attention. English is notorious for its multiple spellings, such as *centre/center* and *yogurt/yoghurt/yoghourt*. There are also some punctuation rules which have alternative forms (e.g., *Charles'/Charles's*). As well, within a document of any length, there may be font, face, or size changes to indicate headings, subheadings, etc. Although many people have preferred spellings and punctuation, and font regularities are relatively easy to check by observation, within-document consistency control is useful—particularly if the document is large, or if the writer is a poor speller who spells erratically throughout the document, causing the spelling checker to recommend first one, then an alternative spelling. When the document becomes a joint effort, especially when using independent or relay document control, or the separate-writing strategy, such consistencies will become even more difficult to maintain. Within-document consistency software could be designed to flag variations in spelling, punctuation etc. conventions to ensure the document is consistent throughout.

2.4.10 Stylistic consistency

In addition to bringing their preferred styles of working to the collaborative writing process, individuals also bring their own styles of writing, which they have developed during their previous writing experiences. By writing *style*, I mean a writer's linguistic choices that are characteristic of the individual, a group, a genre, a historical time period, a communicative goal, a register, and/or a rhetorical stance. These choices may be conscious (e.g., consideration for the subject matter, the audience, the purpose, etc.) or unconscious (e.g., dialect). Despite the influence of genres and group styles, writers generally develop individual styles within the group norms. Indeed, many writers cultivate distinctive styles, as can easily be seen by comparing two authors—even, contemporaries from the same country with similar educational backgrounds writing in the same domain. Ede and Lunsford (1990) found 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). Attempts to merge styles may result in time-consuming revision, frustration and interpersonal conflict, and they are not always successful.

Aside from the fact that the writers may be required to produce stylistically consistent documents, why is consistency important? There are several reasons, two of which are analogous with the problems that format inconsistency may create. First, a variety of writing styles within one document may be distracting to readers. Second, as with format inconsistencies, style shifts may lead readers to believe that the document was written in a careless or hurried manner (Farkas, 1985). Finally, shifts in style are a cognitive burden to readers, since they force readers to change their expectations (Enkvist, 1964). Reading comprehension may be impaired by the additional load, particularly if the reader is unfamiliar with the subject matter, or is not a fluent reader.

One way to avoid the problem of merging styles would be to impose a particular style on the writing before the collaboration begins. However, there are three reasons why it is difficult for collaborative writers to produce a stylistically consistent document from the beginning, unless the single-writer strategy is used.[5] First, many people are loath to abandon a preferred writing style. It is not simply obstinacy on a writer's part to cling to a personal writing style, however; one of the five strategies writers use to reduce the cognitive burden of writing 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. Second, many writers do not want to impose stylistic restrictions on others. Writing together often involves conflicts about content and procedure, which can be time-consuming and stressful to resolve. Allowing each collaborator to write in his or her own writing style avoids an additional source of potential conflict among the group members. Third, people tend to have difficulty describing style, so even in cases when they want to impose a specific style, they may not be able to adequately provide their writing partners with the information needed to write in that style.

Once a document is written, there are two reasons why integrating different writing styles is potentially difficult for collaborative writers, even if the document is given to one person at the end to merge the various parts (centralized control). First, many people are poor at consciously recognizing inconsistent style. They might be dissatisfied with the document, yet not know why. Second, even when people recognize that the document does not have a single style, they are often unable to articulate the specific stylistic inconsistencies they have noticed.

For these reasons, unless centralized control is maintained throughout the document production and the single-writer strategy of collaborative writing is used, writing a document with a single voice may be very difficult. Collaborative writing systems that could help writers merge their styles would be especially useful for writers of documents which require a consistent style not only within a document, but from document to document (e.g., documents that are part of a series; documents produced by a large corporation).

2.5 Existing collaborative writing systems

Many of the existing collaborative writing systems provide support for only one or two specific aspects of collaborative writing. For example, Grove is a joint outliner (Posner, 1991); CRUISER provides visual and audio channels that allow informal interactions between potential collaborators (Duin, 1991). However, some current systems do furnish collaborative writers with more comprehensive support (e.g., Aspects). Nevertheless, since collaborative writing involves a wide range of practices, such systems do not necessarily support the same aspects of collaborative writing. Even when the practices that are facilitated and the problems that are alleviated are the same, they are often supported in different ways, to different extents. Therefore, it is difficult to make direct comparisons between systems. To present an overview of which of the above requirements are being addressed, and how, in current computer-supported collaborative writing systems, I will describe, in some detail, three current systems that exemplify various approaches to designing collaborative writing software, focusing on the type of assistance provided for the issues identified above.

2.5.1 Aspects

Aspects is a commercial collaborative editor developed by Group Technologies Inc. (now called GroupLogic) that supports basic word processing. It also has drawing and painting tools, and graphics can be added to text documents. It is available for both Macintosh, and IBM and compatible systems. Users can choose to work off-line, over a network or point-to-point via a modem or serial cable, allowing both synchronous and asynchronous editing. Users create conferences to jointly edit documents that were created either with Aspects's writing application, or another word processor. Up to sixteen people can join each conference. The creator of the conference can define

[5] Even if a document is written by one person, there is the potential for stylistic inconsistency, particularly if the document is large and is written over a long period of time.

others’ roles and access rights to any documents in the conference in two ways. First, they can control who joins each conference, or allow free access. Second, they can specify whether others are allowed to edit, and whether a turn-taking protocol is enforced, allowing only one editor to revise the document at any one time. Communication among users is facilitated in two ways: each member of the conference can select a unique telepointer to use to indicate to their collaborators any part of the document they need to refer to, and a “chat box” is available for users to exchange notes about their work. The document is continually updated to ensure that all users have the same version. Simultaneous editing of the same section can cause inconsistencies, but users are warned when this problem occurs and the divergent copies are saved and renamed. Collaborators can choose to share their view of the document with some or all users, or they can unlink the view, allowing independent work. To find out who else is currently editing, collaborators can use the Who is Viewing command. Aspects does not provide any support for format consistency, a global perspective or style merging. Collaborators are only identified when using communication channels; no information is maintained to indicate who wrote which parts.

2.5.2 MILO

MILO is a system developed at the Dundee Institute of Technology, implemented in the X-Windows system (Jones, 1993). It supports distributed asynchronous writing of structured documents that may contain both text and graphics. Exchange of documents is both hardware and system independent. The document is stored as one or more notes, each of which contains a text field, a text editor, and a graphics editor. The notes can be structured before, during, or after writing using an idea processor, which allows both a linear and a tree-like global view of the document. Notes that function as annotations can also be created. A history of the document, including who created each note and when, is continually updated. There is destructive semi-automated version merging to keep documents up-to-date. Co-authors can communicate through email. Additional tools include a search facility and a spelling checker. Future work on the integration of annotations, and provision of two additional kinds of text structure view is planned. MILO does not provide means for: enforcing format consistency; specifying users’ roles; synchronous writing; or ensuring stylistic consistency.

2.5.3 SASSE

SASSE (Synchronous Asynchronous Structured Shared Editor) is a prototype collaborative writing system that was developed at the University of Toronto (Baecker et al., 1993). It supports both synchronous and asynchronous editing of documents from workstations linked over local or wide-area networks. Import and export facilities allow the use of other file formats. Text is colour-coded by author, allowing writers to easily find out who wrote what, although this information is lost if the file is exported. The names of all collaborators, their colours, and whether they are currently active can be displayed when desired. An active outline editor allows users to see and manipulate the structure of the text. A gestalt view of the entire document provides a second type of global document view for users. The gestalt view also includes information about where collaborators are working in the text, as do colour-coded scrollbars. For more detailed information about a collaborator’s activities, there is an observation view that lets writers see exactly what a selected collaborator is doing. To prevent writers from simultaneously working on the same section of a document, there is a locking mechanism, indicated by a padlock icon, that only allows one person to change selected text and text which has just been typed. There is a simple version-control mechanism that shows which parts of the document were changed, and by whom. A telepointer allows collaborators viewing the same part of the document the ability to select or refer to particular text without locking it. There is an annotation mechanism for exchanging notes about the document. SASSE also supports brainstorming. SASSE does not, however, support different roles for collaborators, format consistency or style merging.

These brief descriptions indicate the limitations of most of the current collaborative writing software in meeting writers' needs. They also demonstrate the difficulty of comparing existing collaborative writing systems to one another, since the goals of the systems vary widely. Support for the identified issues is being provided to some extent in at least one of these collaborative writing systems, except for format consistency and style consistency. In fact, there are no references to the problem of maintaining consistency throughout a document in the literature concerning collaborative writing systems.

2.6 Other computational approaches to consistency

2.6.1 Consistency of format

Consistency of format is not an issue that has been addressed in the computer-supported collaborative writing literature. However, the difficulty of maintaining a consistent format in long, complex, multi-authored documents has been recognized in the field of technical writing, particularly by large, multinational corporations (Schreurs & Adriaens, 1992). This concern with ensuring document consistency has led to the development of *in-house style* guidelines: a set of predefined rules governing a wide variety of document constraints on issues ranging from layout to style. In addition to enforcing consistency, another aim of in-house style rules is to create documents that are easily understood (by both native and non-native speakers) and easily translated (by humans or machines). Since the manuals required to outline all the necessary rules are long and complex, software has been developed to help technical writers with the difficult task of conforming to the desired standards (e.g., Hoard, Wojcik & Holzhauser, 1992; Schreurs & Adriaens, 1992; Dale & Douglas, 1992).

Companies have tended to protect their in-house style guidelines, so little of this work is available for public use (Schreurs & Adriaens, 1992), which may be one reason that this type of software does not seem to have had an impact on collaborative writing systems. Although the focus of the in-house style software is to impose a particular company's rules, the method used to maintain a consistent format could be adapted for collaborative writing software, allowing collaborators to impose their own rules to ensure internal consistency. Indeed, Dale and Douglas (1992) found that the mechanisms underlying the system they designed to apply in-house style rules had more general applicability.

The maintenance of format consistency appears to be the type of problem that computers are well able to handle, since many features of format are well defined and straightforward to detect. A component to check format consistency could be easily incorporated into a collaborative writing system, allowing writers to run their document through format-checking software, much as they use a spelling checker, when they are in the final stages of writing. Fairly dumb systems could be designed to accept user input specifying preferred conventions, which would work well for issues such as punctuation and typeface consistency. Higher level text aspects that require consistency would require a more intelligent system, such as Dale and Douglas's (1992) "language sensitive" copy editor. By incorporating some knowledge about natural language, they have been able to produce text tools that are more robust than current natural language systems, but smarter than current commercial tools, which rely on simple string searches.

2.6.2 Consistency of writing style

Consistency of writing style has also not been addressed in the computer-supported collaborative writing literature. Unlike format consistency, however, the maintenance of stylistic consistency is much more difficult to handle computationally. In fact, there has been little study, even informally let alone computationally, of style in the qualitative, quotidian sense that I use it here.[6] Style is

[6] In the next chapter, I will review the study of style.

difficult to study because it operates at many linguistic levels, and is intertwined with the realization of the propositional content of a text. However, there are also specific reasons why the scope of software designed to help writers with their style has been limited.

First, much of the research on style is subjective, and defies computational formulation (Ryan, DiMarco & Hirst, 1992). For example, although with the appropriate reading ability and background knowledge readers could recognize irony in text, how does one formally describe what it means for a text to be ironic? Second, certain aspects of style rely on a large amount of world knowledge (e.g., pun), that is presently computationally unmanageable. Third, many features of style that are well defined cannot be detected by computers of today (e.g., inversion). Stylistic features appear in a wide range of linguistic levels, many of which require sophisticated language processing to be recognized. Until robust natural language processing is a reality, only certain types of stylistic features will be computationally tractable. Finally, we do not know how to compute relevant indices of stylistic evaluation, even for simple aspects of style, such as sentence length (Sanford & Moxey, 1989).

Despite these constraints, there are three related areas of software development that claim to have tackled stylistic issues: style checkers, stylistic instruction, and in-house style checkers. There has also been some investigation of style in the natural language processing literature. I will briefly summarize some recent work done in these areas, first providing an overview of the current state of commercial style checkers, then discussing some prototypes, including research done by computational linguists, to illustrate the difficulties involved in designing software to analyze stylistic aspects of language.

Style checkers The broad usage of the word “style” is reflected to some extent in the multiple meanings of *style checker* or *analyzer*. “Style checker” is a term often used interchangeably with *writer’s aid*, thus denoting any tool that aids in the writing process, including such tools as spelling checkers. It is also an alternative name for *grammar checker*, conflating the enforcement of both syntactic rules, the violation of which is simply wrong (e.g., subject-verb agreement), and prescriptive stylistic rules, which should be interpreted as guidelines for writers to follow at their own discretion (e.g., avoid the use of passives). In my discussion of current style checkers, I will focus on the detection of features that would be considered stylistic according to my definition of style (see section 2.4.10).

Style checkers are designed to recognize undesirable stylistic features and help writers eliminate them from the text. They are primarily intended for use by business people (Bolt, 1993), although there is at least one style checker that allows the user to choose from five standard sets of stylistic rules (Lancashire, 1991). Despite some claims to the contrary, there is no actual stylistic analysis done by currently available style checkers. Rather than embodying a semantic theory of style, these programs use simple heuristics that involve little or no linguistic processing to impose principles of “good writing style” as prescribed in style guides, such as Strunk and White’s (1959) (Dale & Douglas, 1992). Instead of analyzing a complete sentence, only parts of sentences are analyzed, relying mainly on part-of-speech information stored in the checker’s lexicon. The lexicon is also used to identify jargon, weak modifiers, colloquialisms etc., (Bolt, 1993). A simple string search is used to detect the use of many “forbidden” features, such as nominalizations or hackneyed phrases. Feedback to the writer may be given in the form of summary statistics (e.g., average sentence length), advice (e.g., *use the active, rather than the passive voice*) or an alternative form (e.g., consider *proceed* instead of *precede*). The best known, as well as one of the oldest, of these systems is the **UNIX Writer’s Workbench**. Since these systems have no real knowledge of syntax, semantics, or pragmatics, many errors are missed (e.g., systems cannot detect the subject-verb agreement error in: **The dog groomed by the children of my friends are…*, Dale & Douglas, 1992). Also, correct forms are often flagged as possible errors. For example, many systems erroneously detect the passive voice in a sentence such as: *They were tired.* Even worse, errors that are purportedly governed by the rules in the programs are not always detected (e.g., *He writing a book is surprising.*). After reviewing six well-known commercial style checkers, Bolt (1993) concludes that although the performances of these programs vary, all of them perform so poorly that they are of questionable value to writers, especially those who are writing in a second language, since they may not have the skills to separate false flags from true errors or to detect errors which the programs miss.

Stylistic instruction A subset of style checkers are those that are designed to improve users' understanding of writing style, as well as to improve their written style (e.g., McGowan, 1992; Payette & Hirst, 1992). Stylistic instruction involves teaching writers the principles and conventions of written language beyond good grammar. Violations of stylistic rules are identified, explanations for the rules are given, and improvements are suggested. Ideally, instruction involves informing the user not only of the correct application of stylistic features, but also of the effects that these devices tend to have on the reader. Although regular style checkers often provide explanations for the rules they espouse, stylistic instruction software provides more detailed stylistic information. Since currently available stylistic instruction programs do not differ significantly from regular style checkers, two prototype stylistic instruction programs will be presented as examples of how style checkers could be augmented to help writers improve their style.

Since these style checkers are intended primarily for students, particularly those who are learning a second language, designers of stylistic instruction software should be aware of two important issues, if they are to provide effective instruction. First, the explanations that are given to the users should be appropriate to their needs; otherwise, users might be overwhelmed or frustrated by unsuitable feedback. An example of how this objective may be accomplished is given in McGowan's (1992) work on a project called **McRuskin**. This style checker focuses on providing explanations of stylistic rules (only one of which has actually been implemented), that vary according to the user's stated level of knowledge about the rule (low, average, or high), and the context of the document (audience, purpose, subject, use, and author's ability). In addition, there is a seven-layer discourse module that allows the user to seek further clarification if the explanation is not understood.

Second, the software should ideally help users develop a flexible understanding of stylistic rules, so that they also learn how and when to apply the rules in different writing situations. Payette and Hirst's (1992) prototype, **STASEL**, has, in addition to a syntactic style analyzer, a goal-directed style analyzer, which enables it to judge sentences according to the apparent stylistic goals of the writer. Rules for the stylistic goals are implemented according to the analysis of syntactic correlates. If the sentence is judged to have met the goal of clarity (the only one which has been incorporated), the user is informed, and the stylistic rule that was used in the analysis is provided. If not, stylistic problems are pointed out, and remedial feedback is given. The structural analysis is presented along with the diagnosis, allowing the user to connect the system's comments with the actual sentence elements.

Most currently available stylistic instruction software, however, has neither a user model nor a text model. Therefore, instruction is generally restricted to a single explanation of each normative feature of style. Also, the software is as limited as current style checkers in terms of the level of stylistic analysis that is performed.

In-house style checkers Another subset of style checkers is in-house style checkers. As well as handling format consistency (mentioned above), they often also purport to deal with stylistic consistency. An in-house writing style is usually developed by defining a *controlled* or *simplified* grammar, which is a subset of a grammar of some natural language. Current prototypes are promising because, due to the imposed syntactic restrictions, more robust parsing is possible, thus allowing the analysis of more complex syntactic structures than is possible in general style checkers (e.g., Hoard et al., 1992; Schreurs & Adriaens, 1992). However, although using predefined rules makes many aspects of writing consistency easier to control (e.g., standard verb tenses), there is a potential cost involved: the restrictive nature of controlled grammars may result in writing that does not sound natural (Hoard et al., 1992). There is also the cost of developing a controlled grammar that allows adequate expressivity—a time-consuming and open-ended task.

One of the Writer's Workbench programs, **prose**, handles the issue of in-house stylistic consistency in a different way. As well as performing stylistic analysis, **prose** allows a writer to compare a document (ideally, at least 2000 words long) to one of three pre-defined standards developed from selected technical and training documents. The document is checked for readability, variation of sentence type, average sentence length, the use of passives and the use of nominalizations. If the measures are more than one standard deviation from the mean of the standard to which the document is being compared, the writer is alerted to this fact. Users can also define their own standard, using the **mkstand** program. Statistics for the five variables are derived from one to seventy-five documents (ideally, twenty or more) supplied by the user, each of which must be at least 1900 words or 90 sentences long. New documents can then be compared to this standard (MacDonald, 1983). This approach is easier and less time-consuming than developing a controlled grammar, but there are several drawbacks. First, a body of already-consistent texts is needed to bootstrap the style standard. Second, fewer stylistic variables (out of an already-limited number) can be controlled. Finally, Gringrich (1983) found that even experienced writers were often unsure of how to apply the advice given by **prose** to their writing.

Natural language processing Computational linguists have explored many aspects of language, including style (e.g., Hovy, 1990; DiMarco & Hirst, 1993). Work on generation and understanding of style is concerned with the choices writers make and how these choices affect (or are intended to affect) readers. Ensuring that these effects are not lost in machine translation is also a prime motivation for research in this area.

The implementations summarized above focus mainly on syntactic stylistic choice, which has also been explored in natural language processing. DiMarco and Hirst (1993) have developed a stylistic grammar that codifies syntactic stylistic knowledge. They began by classifying commonly used stylistic terms into one of three abstract properties associated with sentence style: balance, dominance and position. Next, they defined abstract elements of style that are related to these properties. These stylistic elements were then correlated with syntactic elements. Finally, they defined formal rules to link the abstract elements with specific stylistic goals, such as clarity. Building on this grammar, they have developed a stylistic parser that analyzes text according to these rules, correlating observed syntactic patterns with the writer's possible stylistic goals on three dimensions: clarity/obscurity, staticness/dynamism and concreteness/abstraction. This work demonstrates a way to create a formal representation for the vague notion of style, which may eventually be used to create style checkers that use more than simplistic heuristics.

Aside from this research, however, style has not received much attention from computational linguists (DiMarco & Hirst, 1993). The work that has been done remains largely exploratory, and is therefore not currently robust nor comprehensive enough for use in commercial products.

2.7 Conclusion

Despite my criticism of existing software for its failure to provide comprehensive support for collaborative writing, cognitive research has not yet produced a detailed-enough model to enable the development of such support. Therefore, it is difficult to create a collaborative writing environment from first principles. Consequently, I take an empirical approach to investigating a component of collaborative writing that has been identified in the literature, but has not yet been addressed in software design: stylistic consistency. Creating a consistent style from more than one distinct writing style is a cognitively difficult task that may well be supported by a tool to off-load some of the cognitive burden. I do not propose that such a tool should stand alone, however, but rather that it be offered as only one of a suite of tools within a collaborative writing environment.

How to design such a tool remains a question. It is clear from the descriptions in section 2.6.2 that the current computational handling of style is limited, and will remain so for some time. However, there is another area of computational research which has proven useful in stylistic investigation: *stylostatistics*.

3. Stylistics

3.1 The study of style

3.1.1 History

The study of style has its roots in the ancient study of rhetoric. One of the seven liberal arts, rhetoric was formally codified in the fifth century B.C. in Sicily, although it had been practiced for hundreds of years before this time (Lanham, 1991). Classical rhetoric consists of five parts, the third of which is *lexis* (style). The first two parts of rhetoric, *heuresis* (invention) and *taxis* (arrangement), are concerned with finding and organizing the topic of discourse; the last two, *mneme* (memory) and *hypokrisis* (delivery), are concerned with the oral presentation of the topic, since originally, rhetoric was associated with oratory. Rhetoric was eventually applied to written discourse, particularly after the invention of the printing press (Corbett, 1971). Due to the subsequent shift in emphasis of rhetorical study to text, and the association of invention and arrangement with logic, the study of rhetoric now focuses on style (Lanham, 1991).

As I said in section 2.4.10, style involves a writer's linguistic choices. Although style has been much discussed, however, none of the major rhetoricians actually attempted to define it, which illustrates the vagueness of the term since its inception. Rather, discourse on the classification of style (generally into one of three categories: low, middle, or high), and debates about the functional versus the decorative nature of style, Asianism (highly ornamental style) versus Atticism (brief, witty style), written versus spoken style, and economy versus copia of words were popular. Discussions about how to achieve the translation of thoughts into words centred around choice of words, the composition of words into phrases and clauses, and the use of figures of speech (Lanham, 1991; Corbett, 1971).

3.1.2 Models of style

Despite the words spilled on the topic, after 2000 years of study, style remains an imprecise, multiply defined, and not well understood entity. The theory and practice underlying the study of style has been adversely affected by the variety of different emphases given to the conception of style, due to the resulting number of often-contradictory theories[7] (Quirk, 1969). These various theories have arisen in part because of the overlapping interests of different disciplines in the study of style. Of these theories, two are prominent and influential in the current literature.[8] The first of these has an underlying evaluative model of style. Value judgments are made about writing style according to prescriptive stylistic guidelines. The emphasis is on teaching and imposing norms of style, which tends to discourage the development of an individual style. This understanding of style originated in Greek times (Milic, 1967), and no doubt owes its popularity to many pedagogues over the centuries who, faced with the difficult task of imparting knowledge about style to their students, have opted for the simplest, and most measurable, approach. The evaluative model is espoused in style and rhetoric manuals designed to help students learn to apply stylistic norms in their writing. Since such manuals form their basis, the evaluative model also underlies commercial style checkers. The other, contradictory, theory views style as an individuating factor. Style is considered to be the distinctive expression of a writer, a group, a society, or a combination thereof (Cluett, 1990). Although this model has its roots with Greek rhetoricians (Milic, 1967), it was strongly influenced and popularized during the Romantic period due to the Romantic preoccupation with individuality (Milic, 1991). The aim of this model is to identify or understand writers or groups of writers by studying their writing style (Milic, 1967). Work done in *stylistics* (see below) embodies this second model of style.

[7] For a discussion of various conceptions of style from ancient to modern times, see Enkvist (1964) or Milic (1967).

[8] These two models seem to be the most useful ones for developing software to help collaborative writers merge their styles, since they both embody the belief that one can discover linguistic features of style, unlike, for example, the Crocean view that style and content cannot be separated, or the Senecan view that morality is closely linked with writing style (Milic, 1967).

There are two reasons that I have chosen to investigate the latter rather than the former model of style to find a way in which collaborative writers might be helped to accomplish the task of merging their various writing styles. First, imposing a single normative style on writers would be inappropriate for many genres (and sub-genres) of writing, since there are different stylistic norms associated with different types of writing. In addition, texts are written for different purposes, which will also influence the stylistic choices their authors make. The alternative is to use the evaluative model to develop a variety of normative styles designed to reflect the possible types of writing that the collaborators could be doing (e.g., technical writing, academic writing, etc.). However, such a task is overwhelming, given the variety of possible styles even within one such category (compare, for example, academic writing in computer science with that in literary studies). The difficulty with this approach, as Lanham (1974) points out, is that, "successful prose styles vary as widely as the earth" (p. 17). Second, even if the development of a variety of norms turns out to be an easier task than it appears, I am interested in revealing stylistic differences to help collaborative writers produce more stylistically consistent writing, rather than imposing yet another style on their writing.

Therefore, I will focus on work done in *stylistics*, an area of language study in which the dominant understanding of style is as an individuating factor.

3.2 Stylistics

Although how we express ourselves is restricted by many aspects of language, language use does involve a significant amount of choice. Writers (and speakers) can select from among a variety of words and other linguistic features the ones needed to get their message across to their audience. Linguistic choices are made on the basis of both personal characteristics of the writer (e.g., dialect), and contextual constraints (e.g., medium). The study of linguistic choices is called *stylistics*. In its most general usage, stylistics is: "The study of any situationally distinctive use of language, and of the choices made by individuals and social groups in their use of language" (Crystal, 1992, p. 371). The aim of stylistics is to identify stylistically significant features of language, or *style markers*, and the functions they fulfill (Crystal & Davy, 1969). Stylistics has its origins in linguistics, but is used widely in literary studies.

3.2.1 Stylostatistics

One area of stylistic investigation that has received increasing attention in recent years is *stylostatistics*, also referred to as *stylometry* or *statistical stylistics*. Stylostatistics is concerned with describing a writer's style quantitatively. Rather than relying on a scholar's subjective, and often vague, responses, the goal of stylometry is to find information about style from countable features of text (Potter, 1991). The foundation of the statistical theory of stylistics is that style is a probabilistic concept, and therefore stylistic tendencies can be revealed within the variability of actual writing samples (Dolezel, 1969).

Most of the work done in stylostatistics involves research on the statistical structure of literary texts (Crystal, 1991), although biblical studies (e.g., Radday & Shore, 1985), as well as more general linguistic investigations have also been undertaken (e.g., Biber, 1988). Computers are increasingly used to compile and analyze the information in such studies, which has led to the adoption of the appellation *computational stylistics*. In fact, one of the reasons that stylostatistics has gained in popularity over the last few decades is that computers have taken much of the drudgery out of compiling information from text, and have made statistical analysis less time-consuming and less demanding. Since automated word-tagging became possible in the mid-1980's, it has become even easier to conduct certain types of stylometric investigations (Potter, 1991).

3.2.2 History

The collection of quantitative data from literary works has an old tradition. Alexandrian scholars, who standardized the Homeric text, compiled lists of singly-occurring and unique words from the text. Biblical scholars counted words and verses of each book in the Masoretic text of the Bible to determine its middle word and middle letter (Milic, 1967). The modern idea of statistical stylistics has been around for about 150 years. The first known reference to this idea comes from Augustus de Morgan, who was interested in the authenticity of disputed texts ascribed to St. Paul. He wrote in 1851, "I should expect to find that one man writing on two different subjects agrees more nearly with himself than two different men writing on the same subject. Some of these days spurious writings will be detected by this test." (cited in Bailey, 1969, p. 217). De Morgan suggested that word length might prove to be a distinguishing trait of a writer (Bailey, 1969). The first person to actually test this hypothesis was a geophysicist named Mendenhall (Kenny, 1982). Since that time, researchers in stylometry have studied diverse linguistic features in their search for style markers.[9]

Although there has been over a century of research in statistical stylistics, and the last thirty years have seen a marked increase in such research, along with the technological advances that have facilitated data collection and analysis, there has not been a great deal of progress made in statistical stylistics. There are many problems that are associated with stylometric research that have impeded its development.

3.2.3 Problems in stylostatistical research

The problems affecting stylostatistical occur at all stages of research, from theory to evaluation. I will briefly summarize them to provide a cautionary note, before summarizing research relevant to my investigation.

Theory The first of the problems associated with statistical stylistics is the paucity of literary theory underlying stylistic investigations (Potter, 1991). However, since my use of stylometric techniques is not literary, I will simply refer the interested reader to van Peer (1989) and Potter (1991).

Amount of research The second problem in stylostatistics is the small cohort of researchers, and the resulting lack of research in this area (Potter, 1991). There are a number of biases in literary studies that have discouraged or impeded the pursuit of this type of research by academics. First, Bailey (1969) claims that the explicitness of the assumptions that are necessary to any statistical investigation are not highly valued in literary disciplines, and therefore, tend to be looked upon with suspicion. Second, literary studies encourage the examination of details in text, whereas statistical analysis is better suited to the investigation of broader tendencies (Bailey, 1969). Third, the empirical rather than theoretical emphasis in quantitative studies diverges from modern literary theories (Potter, 1991). Finally, the understanding of statistics is not promoted among students in the humanities. This deficiency has limited both access to such research, and the number of academics who are capable of conducting statistical studies (Bailey, 1969; Potter, 1991).

Knowledge of related work Stylostatistical researchers often have little knowledge of research related to their own work. This problem is illustrated by the fact that investigations are often repeated due to ignorance, rather than to replicate, improve, or verify earlier work (Potter, 1991). The main cause of this ignorance appears to be the lack of communication among academics conducting stylometric research. There are several reasons for this lack of communication. First, stylostatistical studies are conducted by scholars in a variety of disciplines. Since many academic fields do not encourage cross-disciplinary work, some researchers remain unaware of related studies done in other disciplines. Second, work done in certain areas may not be easily accessible to

[9] See Milic (1967) Chapter II "The Problem of Style", Bailey (1969), or Kenny (1982) Chapter 1 "The Statistical Study of Literary Style" for more detailed histories of modern stylostatistics.

academics in others. Most literary critics for example, even those who have taken courses in statistics, lack the necessary statistical background to understand some of the more complex work done by mathematicians and statisticians (Potter, 1991). Finally, the range of styles and variety of languages that have been analyzed is another barrier to accessibility. Without a native-like grasp of the language under investigation, readers might not be able to benefit from the study's insights (Dolezel & Bailey, 1969).

Quality The quality of some of the research is questionable, mainly because the lack of statistical background of most researchers in this area has had an adverse effect on the use of statistics in their studies.[10] First, for the most part, only simple statistics are used, which has limited the scope of the investigations. Second, sampling procedures are at times suspect, which could have resulted in skewed results. Finally, inflated claims are often made by researchers who do not fully understand how to interpret statistics (Smith, 1987).

Replication Deliberate replication or improvement of studies happens rarely (Potter, 1991). Replication is difficult, if not impossible, primarily because of the lack of standards for linguistic analysis. Researchers often invent their own codes, but do not define them explicitly enough for other researchers to use. The use of automatic part-of-speech tagging allows replicable results, but taggers vary in the level of linguistic detail,[11] as well as in how some tags are defined.

Another contributing factor to this problem is that the methodology in these studies is not always described clearly enough for researchers to accurately replicate it. Since there are no explicit standards, and adherence to the ones which have been gradually evolving is inconsistent, assumptions about methodology cannot be made (Potter, 1991).

Evaluation Finally, there has been little subsequent critical evaluation of stylostatistical work (Smith, 1987; Potter, 1991). This deficit can be partially attributed to some of the already-mentioned problems. Since there are few researchers in this area, and many of them lack the necessary breadth of knowledge, the number of people who can fully understand and critique these studies is limited. In addition, the lack of standards for linguistic analysis and the diversity of studies under the stylostatistical umbrella make the comparison of studies complicated even for researchers who have developed the needed expertise. Moreover, being faced with new codes and methods in study after study makes even understanding the research difficult and time-consuming, which discourages critiques of even single studies.

There is some exemplary work in stylostatistics (e.g., the classic Mosteller and Wallace investigation of the disputed Federalist papers, 1964), but few studies meet these standards. Therefore, caution must used in interpreting results of stylometric studies. The above-mentioned problems are serious and many changes must be made in the field before they are alleviated. However, traditional methods of stylistic investigation are also rife with problems, as they rely primarily on academics' subjective responses. Stylometry at least provides an additional, and more concrete, method of investigation, which is particularly important when intuitive judgements are in opposition to one another (Crystal & Davy, 1969). Furthermore, recent retrospective reviews (e.g., Milic, 1991; Potter, 1991), critical analyses of previous work (e.g., Smith, 1987), and articles that demonstrate greater attention to related studies when conducting new research (e.g., Irizarry, 1990; Laan, 1995) suggest a trend towards more critical awareness and self-evaluation among stylostatistical researchers.

[10] The present author is not immune from such criticism.

[11] For example, the Penn Treebank has 47 tags whereas the London-Lund Corpus of Spoken English has 197 (Marcus et al., 1993).

3.3 Stylostatistical research

Statistical stylistics has been used in three main areas of investigation: the identification of characteristics of authors' writing styles, 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. There is a tension that exists among these three areas, each vying for more explanatory power (Cluett, 1976).

Since collaborative writers are writing at the same (historical) time, and are working on the same text,[12] it is only the idiosyncrasies of individual authors that might cause stylistic inconsistencies in a collaboratively written text. Therefore, I am limiting my discussion of statistical stylistics to work that investigates style markers of individuals. The underlying theory of this research is *basic* or *individualist* theory, which holds that a writer's style is individual, and that the mature style is stable over time. The strong view of basic theory does not admit that writers can significantly alter their mature style, nor that they can develop more than one style (Milic, 1991). Despite my focus on individuality, I do not mean to imply that the genre of a text and the era in which it was written do not affect writing style. I believe that each of these factors contributes to the overall style of any piece of writing. This view is perhaps most clearly stated by Cluett (1990): "Despite constraints of time, place, tenor, and genre, the style of any given writer *tends* to remain distinctive in crucial and identifiable respects. Though a writer may be girt round with the bonds of the language as given, yet will the writer's own identity work against those bonds at a number of conscious and unconscious points" (p. 18). The implication is not that writers are constrained to one mode of expression, however. A good writer is able draw on a variety of styles, in order to select the one that is most appropriate to the situation, but this is not to say that these various styles will differ radically from one another. Rather, there are certain characteristics that are preserved, marking the range of styles as unique to that writer (Corbett, 1971). Winter (1969) uses the analogy of identifying a writer's style, much as one identifies a speaker's dialect, by looking at its characteristic isoglosses.

Milic (1991) divides statistical stylistic studies concerned with the style of an author or group of authors into the following categories: *authorship attribution* (authenticity); *chronology of an author's writings*;[13] *character definition; imagery, theme* (content analysis); and *lexis* or *syntax* (aside from those studies that use lexis or syntax to investigate one of the other categories). I will focus on authorship attribution (also called *author fingerprinting*[14]) studies.

3.4 Authorial style

I have focused on authorship attribution because the aim of such research is to identify significant differences between the writing styles of different authors, and to discover 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. The difficulty faced by collaborative writers when trying to merge writing styles may be aided by the identification of the writers' stylistic differences, or style markers. Therefore, the designs of author attribution studies provide a starting point for a stylometric study of collaborative writing and style merging.

[12] Of course, one text could *deliberately* contain more than one genre or sub-genre, but in such a case, style merging of the different genres would simply not be done.

[13] Chronological studies of a single author seem contradictory to the individualist theory (section 3.3) since they assume that an author's unconscious stylistic features develop over time. However, the two ideas are compatible if it is the case that certain features develop, whereas others are stable (Laan, 1995).

[14] The fingerprint analogy should be used with care. It must be kept in mind that, despite the mathematical approach, there is no definitive authorship test. Even when stylometry advances enough to be highly reliable, the possibility of imitation will not allow the certainty that accompanies fingerprint identification (Kenny, 1982).

I will first discuss why there is a need for authorship attribution, then describe some specific studies.

3.4.1 Authorship attribution

Authorship attribution studies are mainly concerned with settling cases of disputed authorship. There are several reasons why the authorship of a text may be in question.

First, authors have not always had ownership of their writing; rather, these rights fell to the owners or publishers of the text. In ancient Greece, it was common practice to write speeches for other people to use in court cases. The person who paid for the speech to be written gained proprietary rights of the text. In England, up until the eighteenth century, publishers' rights, but not authors', were protected. Therefore, authors' names have in many cases been lost or deliberately left out of texts, since authorship was not at the time of writing deemed important (Morton, 1978).

Second, famous authors' names have been applied to the works of others, sometimes out of deliberate deception, sometimes out of hopefulness that one was in possession of an important text, sometimes because it was accepted practice to "borrow" famous names to ensure being read (e.g., in ancient Greece). Similarly, pieces of writing done in institutions of learning have sometimes been published under the name of the school's founder, despite the fact that the founder may have only supervised the work, or simply permitted it to be carried out in the school. In most such cases, the prime motive for attributing the authorship of the work to a famous person was a desire to profit from the credibility of the association (Morton, 1978).

Third, one of the most common methods of stylistic instruction since Greek times has been to require students to imitate the style of well-known writers. Although modern imitations are unlikely to fool present scholars, ancient exercises written by gifted students in a dead language might mislead them, since today's scholars cannot have the same appreciation as a native speaker for the nuances of that language (Morton, 1978).

Fourth, some writers have deliberately published anonymously, or under one or more pseudonyms. There are a variety of reasons for doing so: to avoid personal criticism due to the views being espoused; to conceal an already-established reputation in a different occupation or writing genre; because it was common practice to publish anonymously at certain times, in certain places, in certain kinds of publications (e.g., in American newspapers of the eighteenth century, Mosteller & Wallace, 1964).

3.4.2 Forensic linguistics

Another area in which stylometric techniques for authorship attribution have been applied is in forensic linguistics. Forensic linguistics is "the use of linguistic techniques to investigate crimes in which language data forms part of the evidence" (Crystal, 1992, p. 142). Most of the work that has been done in forensic linguistics has been concerned with issues that were of little relevance to the present investigation. Indeed, a large part of forensic linguistics, *forensic phonetics*, is concerned mainly with voice identification. Therefore, I will briefly discuss stylometric applications, and problems associated with using stylometry in court, but will not include any of them in the summary of authorship attribution studies (below). The reader who is interested in forensic linguistics is directed to the references mentioned in the following paragraph.

Stylometric techniques have been used to try to distinguish genuine from fabricated police confessions (e.g., Morton, 1978), and to determine the authorship of anonymous letters (e.g., extortion or ransom letters, Perret, 1986), and wills and other legal documents (e.g., Miron, 1990). Although stylometric evidence has been presented in court in a number of countries, this application is controversial, for several reasons. First, there is often not enough writing, particularly commensurable texts, to allow reliable tests (Kenny, 1982). Second, the credibility attributed to a forensic linguistics expert witness by the jurors may far exceed the reliability of the test (Kenny, 1982). Third, it is difficult for an expert witness to convey findings to the court in such a way that non-experts can fully understand (Morton, 1978). Finally, even when the best conditions exist, stylometric tests do not provide conclusive evidence of authorship. When the matter is a literary question, the lack of conclusiveness is unfortunate, but not critical; in a court case, however, the conclusions being drawn will affect people's lives.[15]

3.4.3 Authorship attribution studies

Some authorship attribution studies explore texts that were or are believed to have been written by more than one author. Such studies are most relevant to my investigation because they involve trying to find stylistic differences within one, possibly collaboratively written, document. Therefore, I will briefly summarize several of these studies to provide more information about authorship studies, and to describe some of the potential strategies for finding significant stylistic differences within a single text.

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 variously attributed to Marlowe, Bacon (Morton, 1978), Wilkins, Rowley, Heywood, and Chapman (Smith, 1988). Morton's (1978) 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., ratio of *not* to *no*) 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 (1988), however, believed that Morton's study was deficient in several respects, and that the tests he used lacked the necessary sensitivity to resolve the identification of authorship from among playwrights of the same period. 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 (1978) also analyzed *Sanditon*, a novel that Jane Austen did not have time to complete before her death. Using a summary of *Sanditon* that Austen had written, "Another Lady" finished the book for publication. A great admirer of Austen's writing, 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, culling them from *Sense and Sensibility*, *Emma* 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.

[15] This issue is becoming more crucial, as stylometric techniques are being used more and more in actual court cases. One well-publicized controversy concerns the validity of Morton and Michaelson's QSUM authorship test. It has been used in England to settle court cases involving alleged confessions, and university complaints involving alleged plagiarism and the writing of defamatory pamphlets (Morgan, 1991), despite the fact that critics of the test claim that it is not valid for authorship attribution (e.g., Hilton & Holmes, 1993). The debate continues to rage.

Irizarry's (1991) computer analysis of *Infortunios de Alonso Ramírez* (IAR) attempted to discover whether the novel was collaboratively written, or has 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. The analysis of type-token ratio (both lemmatized and unlemmatized), hapax legomena (words that occur only once in a text), sentence length, syntactical differences in sentence beginnings, and various expressions and constructions (e.g., superlative adjectives) 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*. Rather than asking who actually wrote this poem, or whether the first and second parts were written by same person, the question he posed is 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 fifty-nine 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 concludes that these differences reflect a lack of structural unity in the poem.

3.4.4 Work-internal studies

A slightly different type of study that is also relevant to the present research is the *work-internal study*. In such studies, a single work that is known to have been written by a single author is examined primarily to catalogue his or her writing style. Such a detailed description can then be used as an authority in studies such as those on authorship attribution listed above (Laan, 1995). When stylistic inconsistencies are found within a work, they are examined to discover what kinds of text features are associated with the differences. For example, in Laan's (1995) analysis of metre in Euripedes's *Orestes,* she discovered that significantly high incidences of vowel elision were associated with dramatic intensity, and, similarly, low incidences of elision were associated with non-excited passages. She concludes that stylometrists need to discover text aspects that are reliably associated with certain stylistic features, so that they can be taken into account and not allowed to bias studies which rely on writer's individual styles, such as authorship attribution studies.

Conclusion 3.4.5

These studies, both those that study stylistic variations of a single author and those that seek to attribute authorship of sections within a text, had the common goal of detecting differences within a single text. They therefore suggest techniques that might be appropriate for finding stylistic differences among collaborative writers.

4. Method

To achieve the ultimate goal of helping collaborating writers ensure consistency of style throughout a document 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 how stylistic problems can be corrected.

A catalogue of undesirable stylistic inconsistencies awaits further research. We cannot 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. 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.

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 I turn my attention.

4.1 The detection of stylistic inconsistency

I decided to adapt stylostatistical techniques to the problem of detecting stylistic inconsistencies. Although any purely quantitative method seems, a priori, to be inherently inappropriate for a goal that emphasizes automatic qualitative analysis, quantitative methods are not without advantages. First, they are relatively well understood, and are easy to implement, fast to run and very robust, compared to grammars of style. Second, some qualitative measures of style can be easily correlated with quantitative measures—for example, 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. I concluded that it would be worthwhile to explore quantitative methods to determine whether useful correlations between the results and qualitative aspects of style would be found.

My starting point was the authorship attribution research that was reviewed in section 3.4.3. There are clear similarities between the problem of author identification and that of finding stylistic inconsistencies in a document. In each case, one is 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 characterize 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 collaborative writers, the entire document might be only a few thousand words, and the area of analysis could be as small as a paragraph or less.

4.1.1 My questions

My main question, then, was how well I could adapt methods that are used for identifying authors to identifying stylistic inconsistencies. To obtain data for my study, I devised a task in which subjects would write a text of several hundred words in two parts (see section 4.3.3). The assumption is that each writer’s second part will be stylistically more like their own first part than anyone else’s. By pairing each first part with second parts by other writers, I would be able to construct for analysis a set of “collaborative” documents, with possible stylistic inconsistencies, that were controlled for content. In addition, I would be able to compare each first part with each second part, using various stylistic tests, to find out if I could match them up correctly.

A second question I 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, my personal writing experiences suggested that this might not be the case. I have had the frustrating experience of adding to my own previously written work and finding it difficult to continue the writing in a consistent style. I 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 I could investigate whether the two parts were less consistent for these subjects than for those who wrote both parts on the same day.

Thirdly, I was interested in whether people adapt their style to the other author’s when they add to a previously written document. Specifically, I 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 (relay), rather than partitioning the document (independent), might create fewer stylistic inconsistencies. To investigate this question, I decided to have some subjects write only the second half of a text, after having read 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.

4.2 Stylistic Tests

A stylostatistical investigation of authorship depends primarily on two factors: the selection of stylistic features to test, and deciding which statistical test is most appropriate (Smith, 1987). I will now discuss how I decided upon which stylistic features to analyze. I will discuss my selection of statistical tests in section 4.6.

I began this work by surveying studies in stylostatistics, and compiling a list of stylistic tests that have been used in author attribution studies (see Table 1). Most of the work done in this area has focused on word frequencies and word class tagging (Potter, 1991).

Unanalyzed text

- Frequent words (at least 3 per thousand)
- Register of words used (formal, slang, technical, etc.)
- Sentence length
- Word length

Tagged text

- Distribution of nominal forms (e.g., gerunds)
- Distribution of verb forms (tense, aspect, etc.)
- Distribution of word classes (parts of speech)
- Distribution of word class patterns (e.g., *determiner* + *noun* + *verb*)
- Frequency of word parallelism
- Richness of vocabulary
- Type / token ratio

Parsed text

- Distribution of case frames
- Distribution of direction of branching
- Distribution of genitive forms (of and ’s)
- Distribution of phrase structures
- Frequency of clause types
- Frequency of imperative, interrogative, and declarative sentences
- Frequency of passive voice
- Frequency of syntactic parallelism
- Frequency of topicalization
- Ratio of main to subordinate clauses

Interpreted text

- Degree of alternative word use (preference for synonyms)
- Frequency of deixis
- Frequency of hedges and markers of uncertainty
- Frequency of negation
- Frequency of semantic parallelism

Table 1: Some of the tests that have been proposed for use in author identification, organized by degree of linguistic analysis required. I have not included tests that are used to identify a particular genre (e.g., ratio of *thus* to *therefore*), or a particular subject (e.g., topic words), nor tests that do not seem likely to be of any applicability in collaborative writing (e.g., metre).

Rather than analyzing only one or two variables, as large a pool of stylistic features as possible was investigated, within the confines of this study. There are several reasons for this decision. First, we don’t know exactly what we are looking for. Systematically exploring many variables is more likely to yield results than investigating only one or two, as seemingly improbable but significant features may be discovered (Mosteller & Wallace, 1964). Moreover, it is doubtful that one feature is enough to distinguish among writers (Mosteller & Wallace, 1964). Rather, writers probably each have a set of characteristic stylistic features. These sets might overlap to different extents and in different ways, depending on which writers are being compared. Since texts tend to have more features in common than not, and the distinguishing features will be at least partially dependent on the particular texts being compared, finding the distinguishing features may be difficult (Crystal & Davy, 1969). Finally, since stylistic features are not independent of content, authorial attitude, and rhetorical stance, using a variety of tests may reduce the effect of this dependence (Dixon & Mannion, 1993).

Ideally, the stylistic features investigated are ones that: are subject to unambiguous quantification (Mosteller & Wallace, 1964); are stable; and are not significantly affected by subject matter (Milic, 1967).

Grammatical aspects of style seem to be the best candidates for such an investigation because they fulfill the above criteria. Part-of-speech assignment is relatively unambiguous, and part-of-speech taggers—programs that assign the appropriate part of speech to each word in an input text—are fast and accurate. Further, although writers often carefully select the particular word they want (witness the popularity of thesauri), they do not tend to consciously select the part of speech they want to use. In fact, syntactic preferences tend to remain static in the adult writer (Cluett, 1990). The difficulty 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. Finally, whereas vocabulary is highly dependent on subject matter, syntax is much less variable across different domains (Milic, 1967). Therefore, although the majority of the author attribution studies reviewed in section 3.4.3 primarily investigated word choice, and indeed, most stylometric studies investigate lexical items (Laan, 1995), I chose to focus on syntax.

Stylistic tests were chosen according to three criteria. First, they should been used successfully in previous work on stylistic analysis; since I am interested in a new application of stylostatistics, I wanted to try existing tests rather than inventing new ones. 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 perform using unprocessed text or tagged text, rather than parsed text, since a part-of-speech tagger is currently the only fast, robust automated analyzer available for unrestricted text.

The following tests remained after those that did not meet the above criteria were eliminated:

- sentence length: mean and standard deviation
- word length distribution
- percentage of two- and three-letter words
- part-of-speech distribution (including punctuation)
- relative proportion of common parts-of-speech in:
  - sentence-initial position
  - sentence-final position

4.3 Experiment

4.3.1 Pilot study

Prior to conducting the experiment, I carried out 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 English speakers. Due to the poor writing quality of some of these samples, I decided to impose restrictions on whom I would accept as subjects in the actual experiment.

4.3.2 Subjects

Subjects were solicited by electronic bulletin board and by poster (see Appendix A). They were paid either $15.00 or $25.00 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 (N=20) were mainly graduate students from various departments at the University of Toronto. Native speakers of English were selected in an attempt to reduce the probability of syntactic errors, which could confound the stylistic analysis. Graduate students or people with a graduate degree were specified to ensure that subjects had had enough experience in writing to have developed a personal writing style.[16]

Subjects were given an optional questionnaire to collect information 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, organizational behaviour, psychology, sociology (4), zoology (2).

4.3.3 Procedure

The basic writing task consisted of watching a 25-minute episode of the television program *The Twilight Zone*, entitled “Kick the Can”,[17] and summarizing it in approximately 500 words.

There were groups of two to five subjects in each session. Subjects were given pen and paper when they arrived. Once everyone was assembled, they were given written viewing instructions according to which condition was being run (see Appendix A). Subjects were instructed not to talk to each other during the experiment, but were allowed to approach the experimenter with any questions. Next, subjects 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 (see Appendix A).

- In condition A, subjects (N=9) wrote summaries of what they had seen, 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.

[16] Cluett (1990) claims that, “In many writers…style is largely dependent on reflex developed before age 25” (p. 154).

[17] This episode was used with permission from CBS Incorporated.

- 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. After reading part one, they watched the second half of the video. They were asked to complete the description of the first half after that.

All subjects wrote a summary of the second half following the viewing of it. The viewing and writing instructions for the second half were the same as those in the first half. Although subjects were instructed to write about 500 words in total, writing samples ranged from 476 to 1177 words (see Appendix C for selected writing samples).

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 running the experiment for several weeks, my 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 generally poor quality of writing, I decided to discontinue the solicitation of subjects until further investigation had been carried out.

There are several possible reasons for the generally low quality of writing. First, there was a lack of incentive for subjects to spend extra time on their writing, since they were paid a flat rate. Second, revision of the document was not emphasized in the writing instructions. Perhaps more encouragement would have led to better end products. Third, since there were at least two subjects present at every showing, some subjects might have felt pressured to complete their writing quickly once they noticed that others had already finished.

4.4 Tagging the data

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.[18]

4.4.1 The taggers

Two part-of-speech taggers were used to tag the writing samples: POST (Weischedel, Meteer, Schwartz, Ramshaw & Palmucci, 1993) and the Brill tagger (Brill, 1994). Both taggers use the same set of syntactic categories—the University of Pennsylvania tagset (see Appendix B)—but because they use different methods, they might not always assign the same tag to any given word. These taggers represent two different approaches to the problem of part-of-speech tagging: POST uses a probability model, whereas Brill's tagger uses a learning paradigm called transformation-based error-driven learning. Despite the 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 my data are higher than their reported error rates.

POST POST (part-of-speech tagger) is one component of Weischedel et al.'s (1993) natural language system (PLUM) for extracting data from text. Weischedel et al. have used a probabilistic model for their tagger, which is one area of natural language processing in which the statistical approach has done particularly well (Brill, 1992).

[18] I will use the word *tag* metonymously to refer both to the part of speech of a word and to the tag that is used to label it.

Weischedel et al. use a probability model that is commonly used in part-of-speech taggers: a hidden Markov model. A hidden Markov model (HMM) is a doubly stochastic process, but one of the processes is not directly observable (hence the descriptor *hidden*). Each hidden state of the model is associated with a set of output probability distributions (Huang, Ariki & Jack, 1990). The tagger develops a probabilistic tagging model using training data (preferably from the corpus that is to be tagged) to determine the most likely part of speech for each word. Ideally, for each word in a sentence, all possible tag sequences would be considered, the probability of each tag given all the previous tags evaluated, and the most likely tag chosen. In practice, Weischedel et al. use a first-order HMM that has two simplifying assumptions to lower the number of estimated probabilities: independence and locality. The independence assumption states that a word's conditional probability depends only on the current tag's probability. The locality, or Markov independence, assumption states that local context, generally the one or two preceding tags (bigram and trigram models respectively), provides sufficient information to choose the current tag.

Weischedel et al. developed both a bigram and a trigram model, using the Penn Treebank to train the tagger. The Penn Treebank is a 4-million-word corpus that includes newspaper articles, transcribed dialogues and radio shows. Probabilities for a tag in the trigram model were estimated by counting the number of times a given tag followed a two-tag sequence divided by the number of times the two-tag sequence occurred followed by any tag. The conditional probability of a word was also estimated by dividing the number of times a word appeared with a given tag by the number of times the word occurred. To account for previously unseen tag sequences, Weischedel et al. used the simplest of several estimation techniques called *padding*. The formula for the estimation technique they used is: $p(t' | t_1 t_2) = k/m - 1/jm$, where k is the number of times the tag sequence $t' t_1 t_2$ appears in the corpus, m is the number of times $t_1 t_2$ appears in the corpus, and j is the number of tags.

Varying the size of their training set between 1 million and 64 000 words resulted in error rates of 3–4 % for known words, which is about the same rate of discrepancy they found among the human taggers working on their project.

Unknown words, however, are more difficult to tag accurately. Using context alone to tag them resulted in a 51.6% error rate. To reduce the high error rate, Weischedel et al. used information from orthographic endings, hyphenation, and capitalization. This information was not learned from the training data, but specified in advance. If a word has any of these orthographic features, the probability that the word will have a tag associated with the particular feature is adjusted. These probabilities were estimated from the training data. Adding the probability model of orthographic features decreased the error rate for unknown words to 15%.

When tested on another corpus consisting of news, interviews, and speeches about terrorism, POST's error rate was over 8%; when retrained on the new corpus, the error rate dropped to just over 5.5%. About 8.5% of the words in the test were unknown.

POST can also be run in an alternative mode, in which a set of the most likely tags and their probabilities is returned for each word, rather than a single tag. The single-tag mode was used for this work, since the slight increase in accuracy was not significant for the application. For details about the tag-set mode, see Weischedel et al. (1993).

The Brill tagger Brill (1994) has built a rule-based tagger that uses a transformation-based error-driven learning approach. This learning paradigm involves first passing unannotated text through an initial-state annotator, which may range in complexity from naive to sophisticated. Next, these results are compared to a correctly annotated version of the text. Transformations are derived and learned from this comparison. When these transformations are applied to the output of the initial-state annotator, the results better resemble the correctly annotated text.

In Brill's initial-state tagger, words are assigned their most likely tag on the basis of estimates from a training corpus. Then, a greedy search is applied, which adds only the transformation with the highest score at each iteration of learning to the transformation list. A threshold for error reduction is prespecified to terminate learning. After the transformations have been learned, they are applied, in order, to the output of the initial-state tagger. Two types of contextual transformations templates are used to learn transformations: tag templates (e.g., change tag *a* to tag *b* when the preceding word is tagged *z*) and lexical templates (e.g., change tag *a* to tag *b* when the preceding word is *w*).

Brill compared his tagger's performance to that of Weischedel et al.'s tagger, with the following results. When trained on 600 000 words of the Treebank corpus, then tested on a different set of 150 000 known words from the same corpus, the error rate was 2.8%, which was slightly better than Weischedel et al.'s (see above). Tagging without lexical transformations raised the error rate to 3.1%, which is comparable to POST's error rate.

Brill also built a transformation-based learner to improve the tagging accuracy of unknown words. First, the initial-state tagger labels capitalized words as proper nouns, and tags all other unknown words as common nouns. Next, transformation templates are used to learn rules for more accurately guessing the most likely tag for unknown words—for example, changing the guess of the most-likely tag of a word from tag *a* to tag *b* if character *z* appears in that word.

To test the accuracy of tagging text containing unknown words, Brill used the first 950 000 words of the Treebank corpus to train his tagger, and the next 150 000 to test it. From 84 simple rules, 267 contextual rules and 148 rules for tagging unknown words were learned. The overall tagging accuracy was 96.5%, with an error rate of 15% on the unknown words, which is comparable to Weischedel et al.'s (1993) results.

With a minor modification to the transformations, Brill's tagger can also return multiple tags for each word. Instead of using templates of the form *change tag a to tag b*, templates can be written as *add tag a to tag b*, resulting in a list of alternative tags. Training involves finding the transformations which maximize the increase in accuracy, but add as few extra tags as possible. The single-tag mode was used for this work.

4.4.2 Performance of taggers

Prior to tagging text with the Brill tagger, sentence punctuation (e.g., periods) must be separated from words (with a space), contracted and nominal possessive forms (e.g., *can't*; *father's*) must be separated into two parts (e.g., *ca n't*; *father 's*), and each sentence must be put on a separate line. POST, however, requires no text preprocessing.

Each full sample (parts 1 and 2 combined) was tagged separately. For each sample, the Brill tagger took approximately twenty minutes of real time on an SGI processor with a SPECint92 and SPECfp92 rating of 20. The POST tagger was faster, tagging each file in approximately ten minutes. See Appendix D for tagged writing samples.

4.5 Cleaning up the tagged data

4.5.1 Tag manipulation

The Penn Treebank conflates subordinate conjunctions with prepositions, and *to* as an infinitive marker with *to* as a preposition, in order to make the tagset more parsimonious. Since the distinctions can be recovered from lexical or syntactic information, they can be eliminated from the tagset, but remain recoverable (Marcus, Santorini & Marcinkiewicz, 1993). As I wanted to know whether prepositions, infinitive markers and subordinate conjunctions were stylistically distinctive tags, the data were "massaged" to recover counts for these tags. Therefore, by performing a simple lexical look-up, all instances of subordinate conjunctions[19] were retagged using a new tag, SB. I also separated *to* as an infinitive marker from *to* as a preposition. *To* was identified as an infinitive marker whenever it preceded a base form verb (or the sequence adverb, base form verb, i.e., a split infinitive). All other instances of *to* were retagged as IN (preposition), thus leaving only infinitive markers tagged as TO (*to*).[20]

4.5.2 Tagger errors

Several of the tagged files were checked by hand to detect common errors that might interfere with the results, and to compare the types of errors made by the two taggers. Since Brill's tagger always tagged the word *can* as a modal, and this word was used often as a noun because of its importance as both an object and a symbol in the story, *can*'s tag was changed to noun, where appropriate. Each instance in the POST-tagged text in which the noun *can* was erroneously tagged as a modal was also corrected, to ensure an equitable comparison.

POST did not tag any exclamation marks (!). Rather, all twenty instances of exclamation marks in the data were treated as the final character of the previous word. These words were generally tagged incorrectly, since they were not recognized as known words (e.g., *her!* was tagged as [CD]—cardinal number). Since periods and question marks were tagged appropriately, this error must be a bug in the tagger. I decided to re-tag them as sentence-final punctuation, since I thought that they would adversely affect too many measures, and it is trivial to recognize and correctly tag exclamation marks. However, the words preceding them were not corrected, since it is not obvious how they would have been tagged by the tagger.

Since the ultimate goal of this investigation is to develop a robust tool based on current tagger capabilities, no other errors were corrected.

The Brill tagger never assigned the tag RP (particle) in these data, although Brill's tagger does contain rules that deal with particles, and RP is one of the tags in the lexicon for *up* (Brill, personal communication), a word that was commonly used as a particle in these samples (e.g., *He picks up the can…*). Although POST assigned the tag RP, only about a third of the words it tagged as such were really particles, and it also mistagged about a quarter of the actual particles as other parts of speech.

POST often mistagged abbreviations (e.g., *e.g.*), mistaking the final period of the abbreviation as the end of the sentence. Since the text was preprocessed for the Brill tagger, there was no confusion between abbreviations and sentence-final punctuation in that data.

POST assigned the tag RB (adverb) to *cannot* , except once when it assigned VBD (past tense verb); the Brill tagger tagged *cannot* as a modal every time. *Cannot* was used eight times in the data.

[19] The subordinate conjunctions that were retagged: *after, although, as, because, before, ere, for, hence, if, inasmuch, lest, nevertheless, otherwise, provided, save, since, so, than, though, till, unless, until, what, when, whence, whenever, where, whereas, whereat, wherever, wherefore, whether, while*. Certain subordinate conjunctions (*else, other, rather, how, still, such*) were never tagged IN, so I did not include them in the list. Also, I consider *that* to be a relative pronoun, but since the tagger assigned the tag IN to *that,* it was retagged as a subordinate conjunction, rather than leaving it with the preposition count.

[20] Some information was lost due to intervening punctuation (e.g., *to "act"*), and at least one instance of a two-word adverb between the infinitive and the infinitive marker (e.g., *to at least try*). These constructions were simply treated as noisy data.

Different types of mistakes were more commonly made by one tagger than the other, but neither tagger stood out as significantly better. Words ending in *-ed* when used as adjectives were usually tagged as such by the Brill tagger, whereas POST usually tagged them incorrectly as past participles (e.g., *The boys were very* *displeased.*). On the other hand, when the auxiliary verb did not immediately precede the base form of the verb (e.g., *Will Mr. Cox* *find* *Charles et al?*). POST was more likely to correctly tag the verb as the base form, whereas Brill's tagger incorrectly tagged it as the inflected form. An instance in which both taggers made the same error, was in tagging the verb *mutters* as a plural noun in the sentence: *Ben* *mutters* *that he won't find them there….* In a few cases where there was a discrepancy in how a word was tagged, the "correct" tag was not clear. For example, the Brill tagger labelled *fellow* as a noun and POST labelled it an adjective in the phrase *fellow resident.* A case could be made for either interpretation, illustrating that word class assignment is not wholly unambiguous.

4.5.3 Special problems

One difficulty common to every investigation is how to handle special problems, such as quotations by other authors (Mosteller & Wallace, 1964). Since there is no standard practice, such questions have been settled in the past according to the judgement of the investigators. Although there are now gradually evolving standards in stylometry, many researchers are unaware of them, or choose to ignore them (Potter, 1991). Kenny (1982), however, points out that what is more important than which decisions are made is that, once made, they should be consistently adhered to because inconsistency in such matters, particularly when the decisions made are not clearly stated (a common oversight in journal articles) makes comparison, replication, or improvement difficult or impossible. Below is a description of special problems I encountered in the data, how they were handled, and why I chose to handle them in the manner described.

What is a word? There is no unequivocal measure of the number of words in a written sample. For example, the Writer's Workbench program **style**, the Unix program **wc**, and the Brill and POST taggers treat hyphens, possessives, and contractions differently. Since there is no standard definition of *word* used in the stylometric literature, researchers justify the definition of *word* used in their investigations, if the issue is discussed at all, on the basis of their own intuitions, preferences, and the writing under consideration (see for example Burrows, 1987).

Since I was using the taggers to collect information on parts of speech, any decisions regarding word counts were made according to how the taggers assigned word and punctuation tags. Since punctuation tags were assigned only to non-words (e.g., quotation marks), they were excluded from the word counts. Contracted forms were counted as two words (e.g., *ca* and *n't*), a name and title were treated as two words (e.g., *Mr.* and *Whitley*), and the possessive marker *'s* was considered to be a word.

Numbers were also treated as the taggers tagged them. Thus, a numeral was treated as one word, as was a written number. There were no occurrences of multi-word numbers (e.g., *one million*), but they would have been treated as separate words by the taggers.

What is a letter? The taggers handle word punctuation differently from sentence punctuation. Word punctuation is considered to be part of the word (e.g., *Dr.*), whereas sentence punctuation is tagged (e.g., a period at the end of a sentence is tagged [.]). Therefore, for word length, punctuation that was treated as part of the word was counted in the letter-count. Thus, *'s* was regarded as a two-letter word.

Whose style? Quotations are a source of noise in stylistic investigation, since they introduce the style of another writer into the text (Mosteller & Wallace, 1964). Consequently, they are often eliminated from stylistic comparisons. However, in these samples, quotations were rarely used, and (except those which may have been copied verbatim from the video) were written from memory, so they were probably influenced by the writer's personal style. I therefore decided to treat quotations no differently than the rest of the text.

Dialogue is another source of potential noise. Dialogue from the video could have been copied verbatim, but even the invented dialogue of characters may diverge from a writer's normal voice (see for example, Burrows's (1987) investigation of the various speech patterns of Jane Austen's characters). Dialogue also generates confusion about sentence and utterance boundaries (Cluett, 1990). POST automatically established such boundaries, whereas the text for the Brill tagger was pre-processed, and therefore the boundaries are established by the user (in this case, me). In some instances, these judgements diverged, which added to the noise in the data (e.g., I analyzed: *"I said we could talk about it." his son explains weakly.* as one sentence, whereas POST analyzed the underlined part as a second sentence). Since dialogue was seldom used, however, it was not excluded from the analysis.

4.5.4 Conclusion

Imperfect taggers and the absence of standards for determining how to handle special problems are the realities of the current stylostatistical investigation. Added to that are ambiguous word classifications, writing errors, illegible script, and typographical errors. Whether or not the data are robust enough not to be seriously affected by the mistakes and discrepancies will in part be answered in the next chapter.

4.6 Statistical analysis

Once the stylistic data were collected, the information was analyzed 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.

4.6.1 Comparison of means

Most current style checkers calculate the average sentence length of a document. To take this measure a bit further, I performed the comparison of means test to compare differences between the sentence means of the samples. Since the majority of samples were less than 30 sentences long, I used the *t*-test, rather than calculating a *z*-score (Kenny, 1982). Therefore, the formula used was:

$$t = \frac{x - \mu}{s / \sqrt{n}}$$

4.6.2 Chi-square

After reviewing the statistical stylistic literature, I chose, for several reasons, to use the chi-square test for homogeneity to do the rest of the analysis. Most importantly, the chi-square test for homogeneity is appropriate for testing the heterogeneity, or the likelihood that samples were drawn from the same population, among a number of different samples (Brainerd, 1974). Moreover, the chi-square test is commonly used in stylometric investigations, and using previously-tested techniques ensures a more reliable outcome (Smith, 1987). Also, I was interested in a new application of existing techniques, rather than in developing a new method of analysis. Finally, I chose the chi-square test because it is a simple test, since, like many researchers in stylostatistics, I am not a statistical expert.

The formula for the chi-square test is:

$$X^2 = \Sigma\ (O - E)^2 / E$$

where *O* stands for observed value, and *E* stands for expected value. 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. The expected value is calculated from the observed values of both writing samples being compared, since the null hypothesis assumes that both halves were written by the same person. Because the total sample is larger than either of its parts, this method provides a better estimate of the features being investigated than simply using the part-of-speech counts from each part one as the expected value (Kenny, 1982).

Categories Before I compared the frequency of the part-of-speech tags in each pair of writing samples—either the tags in the complete text, the sentence-initial tags, or the sentence-final tags—I had to find out which tags occurred frequently enough to be included in the analysis. Since the chi-square test is not reliable when the observed frequency is 0, tags which did not appear in at least 85%[21] of the first part writing samples or 85% of the second parts were, where appropriate, collapsed with other tag counts into a meaningful group (e.g., cardinal numbers were combined with nouns; see Appendix B for all the categories). If a tag did not occur often enough, and there was no valid criterion for combining it with another tag or group of tags (e.g., existential *there*), the tag was eliminated from the analysis. I decided to eliminate them, rather than creating an other category of unrelated tags, because I felt that the analysis of such a category would not provide any useful information. In the small number of cases for which the expected frequencies could not be reliably calculated (i.e., the observed frequency was 0), they were simply not calculated.

In the comparisons of tag occurrences over the entire sample, word tags that were either collapsed with other categories or eliminated, for the most part, occurred so infrequently that there was little loss of potential information. Moreover, most of them were parts of speech that tend to occur infrequently in most text (e.g., foreign words (FW)). Unfortunately, though, the majority of the punctuation tags were thus eliminated. After the combinations and eliminations were completed, there were 25 categories[22] (22 word tags and 3 punctuation tags) remaining out of the 48 tags (34 word tags and 14 punctuation tags) in the tagset.

In the comparisons of typical sentence-initial and sentence-final tags, however, more information was lost. More tags had to be combined, substantially lowering the number of categories and more tags had to be eliminated from the analysis. Since the beginnings and endings of sentences are highly conventional structures, unlike the middle of a sentence, which is relatively free (Morton, 1978), it is not surprising that certain tags almost never occur (e.g., a determiner in sentence-final position). However, some of the categories that were eliminated occurred fairly frequently and seemed to be important ones to analyze. For example, although over half of the writing samples had at least one instance of a verb at the beginning of a sentence, there were too many samples that did not, thus preventing a meaningful comparison of verbs occurring in sentence-initial position. Therefore, the occurrences of common constructions such as questions (e.g., *Did I tell you...?*), imperatives (e.g., *Take me with you...*) and sentences with initial verbal phrases (e.g., *Using a blue filter,...*) were not analyzed. The most likely reason for this problem is that the writing samples were too short. I believe that longer writing samples would contain a greater variety of sentence-

[21] I chose this number as a cut-off point simply by looking at the data, noticing how many comparisons would be lost due to expected frequencies of zero, and deciding whether there would be enough comparisons left to warrant the analysis.

[22] There were only 24 categories for the Brill tagger since particles (RP) were never tagged.

initial and final tags, since as a document gets longer, writers are more likely to use constructions that are not their most frequent ones. This hypothesis is supported by the fact that it was primarily the shorter samples that lacked instances of sentence-initial and sentence-final tags that did occur in a relatively large number of the longer samples.

After eliminating and combining tags, there were four categories in the analysis of sentence-final parts of speech—modifiers (adjectives and adverbs), nouns (all types), pronouns (both personal and possessive), and verbs (all tenses) and particles—and five categories in the analysis of sentence-initial parts of speech—determiners, modifiers, nouns, pronouns, and conjunctions (both subordinate and coordinate).

Although infrequently used tags had to be left out of the analysis, I do not mean to imply that infrequent tags, or a writer's omission of a tag, are not stylistically significant. On the contrary, a tag that never occurs in one sample, but occurs often in another might noticeably affect the overall stylistic consistency. Indeed, stylistically significant features display different types and different degrees of distinctiveness, and those that are most significant either occur frequently, or are uncommon features of most texts (Crystal & Davy, 1969). Significant features that occur frequently are the only type that are found by my approach; the investigation of infrequent features requires a different method.

Method of analysis for the chi-square 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 homogenous the samples are, the more likely it is that they were written by the same person.

Qualitative, rather than quantitative conclusions, can be drawn from such rankings. The values of chi-square were first interpreted qualitatively, following Smith (1987), who points out that if the data cannot be shown to be normal, and the writing samples cannot be considered to be random samples, interpreting chi-square in relation to probabilities is not valid, and therefore, using levels of significance to analyze results is misleading.

Although the data do not meet the necessary criteria, I did also analyze the chi-squares in relation to their levels of significance for several reasons. First, I am interested in finding ways for documents to be checked for stylistic consistency without comparisons to other documents. Rankings will only be useful to writers who are producing documents that should conform to a specific style of document (e.g., in-house style). Writers who do not have documents for comparison cannot use rankings to help them achieve stylistic consistency. Second, an examination of the rankings revealed that the values of the chi-squares from one ranking to the next were sometimes extremely close (e.g., 21.683 and 22.704, for the first- and second-ranked values from a 2x25 contingency table), whereas at other times the increase was relatively large (e.g., 24.661 and 43.212, also for the first- and second-ranked values from a 2x25 contingency table). So, a part two could be ranked relatively low, but still be nearly as homogenous with a given part one as the top-ranked part two, or, conversely, it could be ranked relatively high, but not be at all similar to the given part one. This information is not easily perceived when examining the rankings. Finally, the ranking alone does not give any clues as to what is contributing to the stylistic differences.

However, on its own, the level of significance does not provide information as to the source of the difference either (unless of course it is a 2x2 contingency table, in which case the absolute value of the difference between the expected and the observed values is the same for both categories). That is, the difference is spread over all the categories; therefore, particularly when the chi-square value is large and there are many categories, it is not clear whether it is one or just a few categories that are responsible for the differences, or whether the differences are distributed relatively equally among all of them. Therefore, I 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.[23]

[23] I 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).

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. <u>He</u> 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 <u>Charles</u>.*).

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 <u>ø</u> 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 <u>them</u>. <u>He</u> 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 <u>he</u> 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, <u>we</u> 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 <u>and</u> of making mistakes <u>and</u> refuses to go along with Ben's view that they are "old men" <u>and</u> need rest <u>and</u> cannot act impulsively <u>and</u> 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 <u>at</u> an idyllic old-age home.*; *<u>Inside</u> the home we are introduced to some <u>of</u> its residents sitting quietly, not communicating <u>with</u> 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 <u>down</u>* [RP] *the stairs...*), and in tagging other parts of speech as prepositions (e.g., *As he walked <u>outside</u>* [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 (*One 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 <u>counting</u>.*

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* <u>bed.</u>) 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 <u>(whether psychiatrist or general practitioner is somewhat unclear)</u>…*).

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.

6. Future work

In this thesis, I have analyzed the problem of deleterious inconsistencies of style in collaborative writing, and laid out an approach to research on the topic. My work was intended to be exploratory. I have described an experiment aimed at collecting data for the research, and some of the limitations and problems that arose. I have shown that some stylometric tests can match up different parts of one writer's text fairly well. Moreover, some of these tests flag inconsistencies that are likely to occur when different sections of a document are written by different people. For example, inconsistencies in verb tense, percentage of modifiers, level of vocabulary, type of nominal preference, placement of adverbials, and use of coordinate conjunctions were revealed by the stylostatistical tests. Part-of-speech distribution provided the most information about stylistic inconsistencies, but the relationship between some of the inconsistencies detected and how this translates into advice for the user requires considerable linguistic analysis which has yet to be done. The most promising tests at this time are ones that have a straightforward relationship between the inconsistency they reveal and the interpretation of this information. Tests that reveal inconsistencies in verb tense and word-length distribution are two that can be understood by writers without the need for an intermediary.

Many questions have been raised by this investigation, and much research remains to be done before we will know whether adapting stylostatistical techniques to the problem of stylistic inconsistency will result in an effective computer aid for writers. In this chapter, I will discuss interesting questions that arose during the course of this work and suggest some approaches that might begin to answer them.

6.1 Correlation with human perception

I was first interested in finding tests which were computationally tractable, but it is obviously important to find out what people notice when reading different styles, since subjective judgements of the stylistic qualities of a text are not always borne out by empirical tests (Bailey, 1969), and likely the opposite is also true. As anyone who has used an automatic writer's aid can attest, it is annoying and sometimes detrimental to the editing process to be presented with false positives, particularly when the system is slow and one is trying to meet an imminent deadline. False negatives are equally bothersome, especially since copyediting accuracy can actually drop if there are few errors—which may well be the case after using a good writer's aid. How, then, can both types of errors be reduced to optimize a writing tool that uses stylostatistical techniques? Unfortunately, unlike spelling and grammar, there is no agreed upon set of rules for stylistic consistency. Instead, we need to find out what people attend to when they judge a document to be stylistically consistent or inconsistent.

6.1.1 How can we ensure that stylometric tests identify only those documents that people perceive to be stylistically inconsistent?

In other words, how can we avoid false positives? In my investigation, most comparisons that were flagged as inconsistent were also perceived by me as inconsistent. However, some of the comparisons that were flagged showed no obvious inconsistency upon examination. Although there are several possible explanations for such results (see section 5.2.3), they might simply have been false positives. To find out whether there is a high correlation between stylostatistical tests and human judgements, experiments are needed in which people are presented with texts that have been flagged as stylistically inconsistent and asked for their judgements as to whether or not the alleged inconsistencies are present. Since many people are poor at consciously recognizing inconsistent style, experts, such as editors, would be the best subjects for such an investigation.

6.1.2 If people perceive a document as inconsistent, do stylostatistical tests identify it as inconsistent?

The flip side of the previous question is whether stylometric tests flag texts that people consider to be inconsistent. To investigate this question, researchers would first need a set of texts that had been judged to be stylistically inconsistent by a large sample of readers. Then, analysis of what aspects affected the perception of differences needs to be done (either by expert subjects or the researchers). Finally, the stylistic tests would be performed on the texts to find out whether they flagged the same inconsistencies that people perceived, and if they differed, what might account for the discrepancy.

6.1.3 Are previously-defined significance levels the best measures?

Related to both of the above questions is the problem of a cut-off point. Do standard significance levels identify stylistic inconsistencies in texts that people also perceive to be stylistically inconsistent? If not, what criteria can be used to uncover the presence of perceptible inconsistencies? One approach to answering these questions is to give different groups of readers texts that had been found to be stylistically inconsistent at different "significance" levels to find out which level correlates best with readers' perceptions. Since the level may not be the same for all stylostatistical tests, investigations using a variety of tests is necessary.

6.1.4 Do salient stylistic factors need to be more consistent?

Certain genres and text types are strongly associated with certain stylistic factors. Indeed, good parodists exploit the salient text characteristics of the style they are imitating, as do less scrupulous impostors. Because of their saliency, the consistency of such stylistic factors might need to be more consistent than other stylistic factors. One way of investigating such a question would be to compare more than one imitation with the real author's work. After agreement is reached about which imitation is better, stylostatistical analysis could then be done to find out whether salient aspects of style are closer to the original in the best imitation. Such comparisons may provide information about what clues people attend to in their perception of style and style similarity.

6.2 Generalizability

There are a number of variables in my experiment that can be altered. Doing so might change the results considerably, or might have no significant effect on the outcome. To find out whether the results of my investigation are generalizable to other situations, there are several questions that must be answered.

6.2.1 Does the text type affect the reliability of stylostatistical tests?

Writing style is influenced by factors other than personal style, such as genre and purpose. In my experiment, subjects wrote a retelling after viewing a television episode. Comparisons of the samples suggest that they are, in some respects, relatively homogenous, which might be due to this writing task. Bereiter and Scardamalia's (1987) cognitive research on writing suggests that writing style tends to show less variation when people are simply reporting what they know (*knowledge telling*) rather than intentionally reworking knowledge as they write (*knowledge transforming*) (Scardamalia, personal communication). Since the writing task in this research involved straight retelling and limited time was given to complete the task, which allowed little chance for revision, it is likely that most of the writing samples were examples of knowledge telling. Comparisons of texts resulting from tasks that encourage knowledge transforming might reveal a wider variety of, or more distinctive, stylistic differences. To find out whether the stylostatistical tests would be useful in uncovering inconsistencies not seen in the type of samples I used, but that might occur in other types of texts, experiments involving polished writing samples of various genres and purposes need to be done.

6.2.2 To what extent does sample size affect the sensitivity of stylostatistical tests?

The corpora traditionally used in stylostatistical investigations are huge. Moreover, many researchers in this area claim that the stylistic properties of a text can only be adequately determined when a great deal of text is available (Bailey, 1969). Cluett (1976) suggests a 700-word minimum, but points out that the appropriate sample size depends on which aspects of the text are under investigation; if the sentence is being analyzed, for example, the minimum would have to be measured in sentences, rather than words. This observation is consistent with my finding that most tests that had considerably fewer data points (e.g., comparisons of sentence-initial and final tags) were noticeably less reliable than tests that had many. Of course, in any study, larger sample sizes contribute to more reliable results because they are less affected by small errors and outliers, but huge sample sizes are not practical for my desired application. Although the problem of inconsistency usually emerges over longer stretches of text, stylistic inconsistency can occur within a sentence. Therefore, stylostatistical tests would ideally be of use even in the analysis of very short texts. Research on the effects of sample size on the reliability of stylostatistical tests would provide information about the optimal minimal sample size for the aspect of text under investigation, and perhaps suggest ways to adapt the tests when the text is shorter than is ideal.

6.2.3 Is my linguistic analysis valid for other writing situations?

In my analysis of the inconsistencies that had been flagged, I found correlations between the quantitative inconsistencies that were detected and the qualitative inconsistencies that I observed. In some cases, the correlations were relatively straightforward (e.g., tense inconsistency), and the interpretation of the inconsistency is likely to be the same in other writing situations. However, some of the inconsistencies (e.g., rate of pronominalization) did not have a single cause, but rather resulted from one or more of several factors that were not always immediately obvious. Further, the fact that the mapping of inconsistencies is not one-to-one suggests that the patterns of differences associated with certain part-of-speech inconsistencies will not necessarily be the same in other cases. Finally, my investigation could not be exhaustive because of the limitations of the samples. Before people can be given stylistic feedback from stylostatistical tests, aside from the information that an inconsistency in the use of a specific tag has been detected, a great deal of linguistic analysis of the patterns revealed in many writing contexts is needed to determine which are valid.

6.3 Applicability of tests

6.3.1 Which tests are most useful?

My work did not include an exhaustive list of currently possible automatic stylostatistical tests, so further research that investigates other stylostatistical tests is needed. Since some of the tests I used revealed more useful information (e.g., overall tags) than did others (e.g., two- and three-letter words), it is likely that gradations of effectiveness would be revealed when the various tests are compared. Effectiveness would have to be measured on a number of counts, such as: how much information is revealed; how consistent the results are; and how well the results correlate with human perceptions.

6.3.2 Which inconsistencies matter?

As mentioned earlier, stylistic consistency is not always necessary or desirable. Given information about stylistic inconsistencies, users will have to decide whether each inconsistency is detrimental in the context of the writing, and if so, how to fix it. However, there may be inconsistencies that are generally undesirable, and other that are generally innocuous. Studies in which subjects read texts with a variety of stylistic inconsistencies, then are asked for their judgements as to which are noticeably taxing might identify inconsistencies that do not tend to burden readers and others that interfere with the reading process. This information might provide a focus for future software development.

6.4 Benefits to users

6.4.1 Does the information about where the inconsistencies are located in the text improve people's speed or ability to merge different writing styles?

Methods and terms for explaining stylistic problems to users and helping them with improvements are not yet available. However, simply presenting information about the presence of stylistic inconsistencies might help people merge inconsistent writing styles, first, by making style an object of conversation, thus engendering discussion of style, and second, by providing a focus for such discussion (i.e., which inconsistencies have been detected). To find out whether such feedback is useful, experiments which involve a writing task to make two pieces of writing more consistent would provide some answers. One group of subjects would be given only the writing, while another group of subjects would be provided with additional information about stylistic inconsistency. Differences in length of time spent merging, perceived ease of the merging and the quality of the finished product can be investigated.

6.4.2 What kind of stylistic advice is helpful to people?

Even if simply providing information about where inconsistencies are located in a text is useful to people, a good style tool needs to explain each stylistic problem and suggest to users how the problems can be corrected in terms that users can understand. This is no easy undertaking. Sanford and Moxey (1989) point to the lack of clear psychological evidence as to whether stylistic instructions given to users can be followed, in what way they are followed, and what side effects may result. For example, in an experiment involving the use of **Writer's Workbench**, several people complained about not being sure how to apply the computer's stylistic advice (Gingrich, 1983). Once clear correlations between quantitative measures and qualitative aspects of style had been established, this information could be used in studies to investigate whether people can use stylistic advice to improve the consistency of their documents. Comparisons between subjects that are given the advice and those that are only told that certain inconsistencies are present could be made to find out whether the stylistic advice contributes to the quality, pace and perceived ease of style merging. To further compound the difficult task of providing appropriate stylistic advice, developers will also have to take into account that users will not all have the same background knowledge about or understanding of style. Therefore, explanations that might be appropriate for experts, such as academics who have studied style and are writing in their native language, will not be comprehensible to novices, such as high school students who are attempting to write in their second language. This diversity suggests that there is a need for a user model to avoid presenting people with explanations that are too complex or too simplistic for their understanding.

6.5 Technological advances

6.5.1 What improvements in existing technology might provide better results?

Existing tools, such as taggers, although robust, are not as accurate as they could be. Parsers remain too slow, fragile, and inaccurate to be used for unrestricted text. However, some promising improvements are now under investigation. Taggers are currently being developed as front ends to parsers, allowing parsers to work at the tag level, thus increasing their accuracy (Charniak, 1994). The resulting tool would allow a wider investigation of stylistic inconsistencies, since many important stylistic variables which cannot currently be computed from a text, such as distribution of phrase structures, would be subject to automatic analysis. Additional stylistic information would potentially lower the number of problems missed by a stylistic tool.

6.6 Collaborative writing strategies

6.6.1 Do certain methods of collaboration result in more consistent text?

Although not directly related to my main question, one of the questions that initially interested me was whether certain collaborative writing strategies produce more stylistically consistent text. I was interested in whether collaborative writers who pass the document from writer to writer (*relay*), rather than partitioning the document (*independent*), might create fewer stylistic inconsistencies. Due to the poor response rate, I did not have enough subjects to contrast the relay method of collaborative writing with the independent method, which was simulated by comparing two parts of the same writing task written separately by different writers. Assigning collaborative strategies to subjects in a collaborative writing experiment to study the effect of different collaborative writing situations on writing consistency would be an interesting experiment that might provide some useful information on the pros and cons of the various methods.

6.7 Pre-defined styles

6.7.1 Can stylostatistical information be incorporated into automatic style mergers?

Perhaps in the farther future, computer style aids will actually be capable of performing the style merging independently. However, even if the computer itself performs the style merging, there will have to be user input, namely the goals the writers are trying to meet. The problem of style merging would thus become similar to the work involved in style generation, in which goals are specified, the information on the topic is available, audience characteristics are provided, and the software must generate an appropriate text. Perhaps in the nearer future, tools that recognize a “canned” style will provide feedback to writers who are writing in heavily genre-influenced areas.

6.8 Conclusion

Despite the many unresolved issues related to this work, the application of stylostatistical techniques to the problem of automatically detecting stylistic inconsistency appears promising. Given the current state of most writing support software, the development of an automatic style merger seems very far off. However, as editing devices evolve in the direction of Dale and Douglas’s (1992) language sensitivity, text base retrieval becomes more automated, and understanding of stylistics progresses, it seems possible that the writing tools of today may eventually be developed into a sophisticated literary assistant, which will further extend the definition of collaborative writing.

References

Baecker, R. M., Nastos, D., Posner, I. R., & Mawby, K. L. (1993). The user-centred iterative design of collaborative writing software. *Human Factors in Computing Systems Interchi '93 Conference Proceedings*, 399–405.

Bailey, R. W. (1969). Statistics and style: A historical survey. In L. Dolezel & R. W. Bailey (Eds.), *Statistics and style* (pp. 217–236). New York, NY: American Elsevier Publishing Company Inc.

Batschelet, M., Karis, W. M., & Trzyna, T. (1991). Selected, annotated bibliography on collaborative writing. In M. M. Lay & W. M. Karis (Eds.), *Collaborative writing in industry: Investigations in theory and practice* (pp. 263–274). Amityville, NY: Baywood Publishing Company, Inc.

Beck, E. E. (1992). *Methodology for studying the dynamics of co-authoring for the design of CSCW writing systems* (Collaborative Writing Research Group paper no. 2). Brighton, England: University of Sussex, School of Cognitive and Computing Sciences.

Beck, E. E. (1993). A survey of experiences of collaborative writing. In M. Sharples (Ed.), *Computer supported collaborative writing* (pp. 87–112), London: Springer-Verlag.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Biber, D. (1988). *Variation across speech and writing.* Cambridge, England: Cambridge University Press.

Bolt, P. (1993). Grammar checking programs for learners of English as a foreign language. In M. Yazdani (Ed.), *Multilingual multimedia*, (pp. 140–197). Oxford: Intellect Books.

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

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, 152–155.

Brill, E. (1994). A report of recent progress in transformation-based error-driven learning. *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI),* 722–727.

Brown, A. L. & Campione, J. C. (1990). Communities of learning and thinking, or A context by any other name. *Human Development*, *21*, 108–126.

Burrows, J. F. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method.* Oxford: Clarendon Press.

Charniak, E. (1994, January). *Taggers for parsers*. Paper presented at the meeting of the Computational Linguistics Research Group, Department of Computer Science, University of Toronto, Toronto, ON.

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

Cluett, R. (1990). *Canadian literary prose: A preliminary stylistic atlas.* Toronto, ON: ECW Press.

Cochran-Smith, M. (1991). Word processing and writing in elementary classrooms: A critical review of related literature. *Review of Educational Research, 61* (1), 107–155.

Corbett, E. P. J. (1971). *Classical rhetoric for the modern student* (2nd ed.). New York: Oxford University Press.

Couture, B., & Rymer, J. (1991). Discourse interaction between writer and supervisor: A primary collaboration in workplace writing. In M. M. Lay & W. M. Karis (Eds.), *Collaborative writing in industry: Investigations in theory and practice* (pp. 87–108). Amityville, NY: Baywood Publishing Company, Inc.

Crystal, D. (1991). *A dictionary of linguistics and phonetics* (3rd ed.). Oxford: Basil Blackwell, Ltd.

Crystal, D. (1992). *An encyclopedic dictionary of language and languages.* Oxford: Blackwell Publishers.

Crystal, D., & Davy, D. (1969). *Investigating English style*. London: Longmans, Green & Co. Ltd.

Dale, R., & Douglas, S. (1992). *Intelligent text processing through natural language sensitivity.* Unpublished manuscript, University of Edinburgh, Centre for Cognitive Science, Edinburgh.

DiMarco, C., & Hirst, G. (1993). A computational theory of goal-directed style in syntax. *Computational Linguistics, 19* (2), 451–499.

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

Dolezel, L. (1969). A framework for the statistical analysis of style. In L. Dolezel & R. W. Bailey (Eds.), *Statistics and style* (pp. 10–25). New York, NY: American Elsevier Publishing Company Inc.

Dolezel, L., & Bailey, R. W. (1969). Preface. In L. Dolezel & R. W. Bailey (Eds.) *Statistics and style* (pp. vii–viii). New York, NY: American Elsevier Publishing Company Inc.

Dorner, J. (1992). Authors and information technology: New challenges in publishing. In M. Sharples (Ed.), *Computers and writing: Issues and implementations* (pp. 5–14). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Duin, A. H. (1991). Computer-supported collaborative writing: The workplace and the writing classroom. *Journal of Business and Technical Communication, 5* (2), 123–150.

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

Enkvist, N. E. (1964). On defining style: An essay in applied linguistics. In J. Spencer (Ed.), *Linguistics and style* (pp. 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, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31–50). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Forman, J. (1992). Introduction. In J. Forman (Ed.), *New visions of collaborative writing* (pp. xi–xxii). Portsmouth, NH: Boynton/Cook Publishers.

Gingrich, P. S. (1983). The Unix Writer's Workbench software: Results of a field study. *The Bell System Technical Journal, 62* (6), 1909–1921.

Harris, J. (1994). Toward a working definition of collaborative writing. In J. S. Leonard, C. E. Wharton, R. M. Davis, & J. Harris (Eds.), *Author-ity and textuality: Current views of collaborative writing* (pp. 77–84). West Cornwall, CT: Locust Hill Press.

Hartley, J. (1991). Psychology, writing and computers: A review of recent research. *Visible Language, 25* (4), 339–375.

Hayes, J. R., & Flower, L. S. (1987). On the structure of the writing process. *Topics in Language Disorders, September,* 19–30.

Hilton, M. L., & Holmes, D. I. (1993). An assessment of cumulative sum charts for authorship attribution. *Literary & Linguistic Computing, 8* (2), 73–80.

Hoard, J. E., Wojcik, R., & Holzhauser, K. (1992). An automated grammar and style checker for writers of Simplified English. In P. O. Holt & N. Williams (Eds.), *Computers and writing: State of the art* (pp. 278–296). Oxford, England: Kluwer Academic Publishers.

Holt, P. O. (1992). Preface. In P. O. Holt & N. Williams (Eds.), *Computers and writing: State of the art* (pp. vii–xi). Oxford, England: Kluwer Academic Publishers.

Hovy, E. H., (1990). Pragmatics and natural language generation. *Artificial Intelligence, 43,* 153–197.

Huang, X. D., Ariki, Y., & Jack, M. A. (1990). *Hidden Markov models for speech recognition.* Edinburgh: Edinburgh University Press.

Irizarry, E. (1990). Stylistic analysis of a corpus of twentieth century Spanish narrative. *Computers and the Humanities, 24,* 265–274.

Irizarry, E. (1991). One writer, two authors: Resolving the polemic of Latin America's first published novel. *Literary & Linguistic Computing, 6* (3), 175–179.

Jones, S. (1993). MILO: A computer-based tool for (co)-authoring structured documents. In M. Sharples (Ed.), *Computer supported collaborative writing* (pp. 185–202). London: Springer-Verlag.

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

Kraut, R. E., Egido, C., & Galegher, J. (1990). Patterns of contact and communication in scientific research collaboration. In J. Galegher, R. E. Kraut, & C. Egido (Eds.), *Intellectual teamwork: Social and technological foundations of cooperative work* (pp. 149–171). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Ltd.

Kraut, R. E., Galegher, J., Fish, R., & Chalfonte, B. (1992). Task requirements and media choice in collaborative writing. *Human-Computer Interaction*, *7* (4), 375–407.

Laan, N. M. (1995). Stylometry and method. The case of Euripedes. *Literary & Linguistic Computing, 10* (4), 271–278.

Lancashire, I. (1991). *The humanities computing yearbook, 1989–1990: A comprehensive guide to software and other resources.* Oxford: Oxford University Press.

Lanham, R. A. (1974). *Style: An anti-textbook.* New Haven, CT: Yale University Press.

Lanham, R. A. (1991). *A handlist of rhetorical terms* (2nd ed.). Berkeley, CA: University of California Press.

Lunsford, A., & Ede, L. (1986). Why write…together: A research update. *Rhetoric Review, 5* (1), 71–81.

Macdonald, N. H., (1983). The Unix Writer's Workbench software: Rationale and design. *The Bell System Technical Journal, 62* (6), 1891–1908.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19* (2), 313–330.

McColly, W. B. (1987). Style and structure in the Middle English poem *Cleanness*. *Computers and the Humanities, 21*, 169–176.

McGowan, S. (1992). Ruskin to McRuskin—Degrees of interaction. In P. O. Holt & N. Williams (Eds.), *Computers and writing: State of the art* (pp. 297–318). Oxford, England: Kluwer Academic Publishers.

Milic, L. T. (1967). *A quantitative approach to the style of Jonathan Swift.* The Hague: Mouton & Co., Publishers.

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

Miron, M. S., (1990). Pyscholinguistics in the courtroom. In R. W. Rieber & W. A. Stewart, (Eds.), The language scientist as expert in the legal setting: Issues in forensic linguistics. *Annals of the New York Academy of Sciences, 606,* 55–64.

Morgan, B. (1991, August 9). Authorship test used to detect faked evidence. *The Times Higher Educational Supplement*, p. 1.

Morton, A. Q. (1978). *Literary detection: How to prove authorship and fraud in literature and documents.* Bath, England: Bowker Publishing Company.

Mosteller, F. & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist.* Reading, MS: Addison-Wesley Publishing Company, Inc. (republished with an additional chapter as *Applied Bayesian and classical inference: The case of the Federalist papers.* Second edition of *Inference and disputed authorship: The Federalist.* New York: Springer-Verlag., 1984).

Neuwirth, C. M., Kaufer, D. S., Chandhok, R., & Morris, J. H. (1994). Computer support for distributed collaborative writing: Defining parameters of interaction. In R. Furuta & C. Neuwirth (Eds.), *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (pp. 145–152). Chapel Hill, NC: ACM Press.

Newman, J., & Newman, R. (1992). Three modes of collaborative writing. In P. O. Holt & N. Williams (Eds.), *Computers and writing: State of the art* (pp. 20–28). Oxford, England: Kluwer Academic Publishers.

Oliver, L. J. (1985). The case against computerized analysis of student writings. *Journal of Technical Writing and Communication, 15* (4), 309–322.

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

Pea, R. D. (1994). Seeing what we build together: Distributed multimedia learning environments for transformative communications. *The Journal of the Learning Sciences. 3* (3), 285–299.

Perret, U. (1986). Linguistique judiciare soutenue par l'ordinateur. In E. Brunet (Ed.), *Methodes quantitatives et informatiques dans l'étude des textes: Vol. 2* (pp. 699–704). Geneva: Editions Slatkine.

Plowman, L. (1993). Tracing the evolution of a co-authored text. *Language and Communication, 13* (3), 149–161.

Posner, I. R. (1991). *A study of collaborative writing.* Unpublished master's thesis, University of Toronto, Toronto, ON.

Potter, R. G. (1991). Statistical analysis of literature: A retrospective on *Computers and the Humanities,* 1966–1990. *Computers and the Humanities, 25,* 401–429.

Quirk, R. (1969). Forward. In D. Crystal & D. Davy. *Investigating English style* (pp. v–vi). London: Longmans, Green & Co. Ltd.

Radday, Y. T., & Shore, H. (1985). *Genesis: An authorship study.* Rome: Biblical Institute Press.

Rimmershaw, R. (1992). Collaborative writing practices and writing support technologies. In M. Sharples (Ed.), *Computers and writing: Issues and implementations* (pp. 15–28). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rose, S., (1994). Toward a revision decision model of collaboration. In J. S. Leonard, C. E. Wharton, R. M. Davis, & J. Harris (Eds.), *Author-ity and textuality: Current views of collaborative writing* (pp. 85–100). West Cornwall, CT: Locust Hill Press.

Ryan, M., DiMarco, C., & Hirst, G., (1992). *Focus shifts as indicators of style in paragraphs* (Research Report CS-92-35). Waterloo, ON: University of Waterloo, Department of Computer Science.

Salomon, G. (1993). No distribution without individuals' cognition: A dynamic interactional view. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations.* (pp. 111–138). Cambridge: Cambridge University Press.

Sanford, A. J., & Moxey, L. M. (1989). Language understanding and the cognitive ergonomics of style. In P. O. Holt & N. Williams (Eds.), *Computers and writing: Models and tools* (pp. 38–49). Oxford, England: Intellect Ltd.

Scardamalia, M. & Bereiter, C. (1994). Computer support of knowledge-building communities. *The Journal of the Learning Sciences. 3* (3), 265–283.

Schreurs, D., & Adriaens, G. (1992). Controlled English (CE): From COGRAM to ALCOGRAM. In P. O. Holt & N. Williams (Eds.), *Computers and writing: State of the art* (pp. 206–221). Oxford, England: Kluwer Academic Publishers.

Severinson Eklundh, K. (1992). Problems in achieving a global perspective of the text in computer-based writing. In M. Sharples (Ed.), *Computers and writing: Issues and implementations* (pp. 73–84). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sharples, M. (1991). The development of a cognitive model for computer support of collaborative writing. *Journal of Computer Assisted Learning, 7,* 203–204.

Sharples, M. (1992). Introduction. In M. Sharples (Ed.), *Computers and writing: Issues and implementations* (pp. 1–4). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sharples, M. (1993). Introduction. In M. Sharples (Ed.), *Computer supported collaborative writing* (pp. 1–7). London: Springer-Verlag.

Sharples, M., & Pemberton, L. (1992). Representing writing: External representations and the writing process. In P. O. Holt & N. Williams (Eds.), *Computers and writing: State of the art* (pp. 319–336). Oxford, England: Kluwer Academic Publishers.

Sharples, M., Plowman, L., & Goodlet, J. (1993). *The development of a cognitive model for computer support of collaborative writing: End of project report* (Collaborative Writing Research Group paper no. 9). Brighton, England: University of Sussex, School of Cognitive and Computing Sciences.

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.

Strunk, W., Jr., & White, E. B. (1959). *The elements of style.* New York: Macmillan.

Van Peer, W. (1989). Quantitative studies of literature: A critique and an outlook. *Computers and the Humanities, 23,* 301–307.

Van Pelt, W., & Gillam, A. (1991). Peer collaboration and the computer-assisted classroom: Bridging the gap between academia and the workplace. In M. M. Lay & W. M. Karis (Eds.), *Collaborative writing in industry: Investigations in theory and practice* (pp. 170–205). Amityville, NY: Baywood Publishing Company, Inc.

Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., & Palmucci, J. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, *19* (2), 359–382.

Winter, W. (1969). Styles as dialects. In L. Dolezel & R. W. Bailey (Eds.), *Statistics and style* (pp. 3–9). New York, NY: American Elsevier Publishing Company Inc.

Appendix A

Subject Solicitation Notice

Subjects Wanted

Subjects whose first language is English are required for a writing experiment. Subjects should currently be enrolled in a graduate program and/or hold a graduate degree.

The experimental task involves watching a 25-minute video and writing about it for up to one hour. Participants will be paid $10.00/hour, with a guaranteed minimum of $15.00.

Subjects will be randomly assigned to one of several experimental conditions. This may involve a second one-hour writing task at a later date, for which remuneration will be paid.

If you are interested, please phone Ms. Dublin at 978-6383, Monday to Friday, 9:30-1:00 or 2:00-4:30, or contact Tom by e-mail at bellman@dgp.utoronto.ca.

Person in charge of research: Professor Marilyn Mantei, Department of Computer Science, University of Toronto.

Instructions to Subjects

Viewing Instructions

- In condition A:

> You are about to watch an approximately 25 minute video in two parts. At the end of each part, you will be asked to write about it. To facilitate remembering, you may take notes as you watch.

- In condition B, the first week:

> You are about to watch the first half of a 25-minute video. You will watch the second half next week. After viewing the first half, you will be asked to write about it. To facilitate remembering, you may take notes as you watch.

- In condition B, the second week:

> You are about to watch the second half of the video you began watching last week. You will again be asked to write about it. To facilitate remembering, you may take notes as you watch.

- In condition C:

> As you watch the second half of the video, you may take notes to facilitate remembering. At the end of the film you will be shown a description of the first half of the video. You will be required to write about the second half, thus completing the description.

Writing instructions

In each condition, the following was included in the instructions:

> Retell the story as you saw it. You need not restrict yourself to a plot description. For example, as you describe the action that took place, you may discuss the theme of the story or what the motivations of the various characters are. Please refrain, however, from discussing things not directly related to the story.
>
> Thank you for your help in this study.

• In condition A, the last instruction was:

> Please write approximately 250 words for each of the two parts, in full sentences.

• In condition B, the last instruction both weeks was:

> Please write approximately 250 words in full sentences.

The second week, the following leader was included:

> As you did with the first part last week, please write about the second half of the video.

• In condition C, the last instruction was the same as in condition B, and the following leader was included:

> 1) Read the description of the first half of the video.
> 2) Complete the description by writing about the second half of the video.

Appendix B: Tagsets

Penn Treebank Part-of-Speech Tagset (Marcus et al., 1993)

1.	CC	coordinating conjunction
2.	CD	cardinal number
3.	DT	determiner
4.	EX	existential there
5.	FW	foreign word
6.	IN	preposition or subordinating conjunction
7.	JJ	adjective
8.	JJR	comparative adjective
9.	JJS	superlative adjective
10.	LS	list item marker
11.	MD	modal
12.	NN	singular or mass noun
13.	NNS	plural noun
14.	NNP	singular proper noun (NP in current Penn tagset)
15.	NNPS	plural proper noun
16.	PDT	pre-determiner
17.	POS	possessive ending
18.	PRP	personal pronoun (PP in current Penn tagset)
19.	PRP$	possessive pronoun (PP$ in current Penn tagset)
20.	RB	adverb
21.	RBR	comparative adverb
22.	RBS	superlative adverb
23.	RP	particle
24.	SYM	symbol (mathematical or scientific)
25.	TO	to
26.	UH	interjection
27.	VB	verb, base form
28.	VBD	verb, past tense
29.	VBG	verb, present participle or gerund
30.	VBN	verb, past participle
31.	VBP	verb, present (not 3rd person singular)
32.	VBZ	verb, present, 3rd person singular
33.	WDT	wh-determiner
34.	WP	wh-pronoun
35.	WP$	possessive wh-pronoun
36.	WRB	wh-adverb
37.	#	pound sign
38.	$	dollar sign
39.	.	sentence-final punctuation
40.	,	comma
41.	:	colon, semi-colon
42.	(	left bracket character
43.	)	right bracket character
44.	"	straight double quote
45.	‘	left open single quote
46.	“	left open double quote
47.	’	right close single quote
48.	”	right close double quote
45.	‘	left open single quote
46.	“	left open double quote

Part-of-Speech Tagset Used for Tags Overall

1. coordinating conjunction
2. nouns: singular or mass noun, and cardinal number
3. determiners: determiner and pre-determiner
4. preposition
5. adjectives: adjective, comparative adjective, and superlative adjective
6. modal
7. plural noun
8. proper nouns: singular proper noun and plural proper noun
9. possessive ending
10. personal pronoun
11. possessive pronoun
12. adverbs: adverb, comparative adverb, and superlative adverb
13. particle
14. infinitive marker
15. verb, base form
16. verb, past tense
17. verb, present participle or gerund
18. verb, past participle
19. verb, present (not 3rd person singular)
20. verb, present, 3rd person singular
21. wh-words: wh-determiner, wh-pronoun, possessive wh-pronoun, and wh-adverb
22. subordinating conjunction
23. comma
24. sentence-final punctuation
25. other punctuation

Part-of-Speech Tagset Used for Sentence-Initial Tags

1. determiners: determiner and pre-determiner
2. modifiers: adjective, comparative adjective, superlative adjective, adverb, comparative adverb, and superlative adverb
3. nouns: singular or mass noun, plural noun, singular proper noun, plural proper noun, and cardinal number
4. pronouns: personal pronoun and possessive pronoun
5. conjunctions: coordinating conjunction and subordinating conjunction

Part-of-Speech Tagset Used for Sentence-Final Tags

1. modifiers: adjective, comparative adjective, superlative adjective, adverb, comparative adverb, and superlative adverb
2. nouns: singular or mass noun, plural noun, singular proper noun, plural proper noun, and cardinal number
3. pronouns: personal pronoun and possessive pronoun
4. verbs and particles: verb, base form; verb, past tense; verb, present participle or gerund; verb, past participle; verb, present; modal; particle

Appendix C: Writing samples

Due to space considerations, selected samples are included here. There are two samples from condition A, two from condition B, and one from condition C. All of the writing samples are available from the author upon request.

Samples will be referred to by condition (A, B or C), sample number for that condition (1 or 2), and part (i or ii). For example, part two of the first sample in condition B is referred to as B 1 ii.

In the tables that follow the writing samples, comparisons that showed significant differences at the .05 level are indicated by an asterisk (*). Those that showed significant differences at the .001 level are indicated by two asterisks. Impossible comparisons are indicated by N/A.

Condition A

In condition A, subjects wrote summaries of what they had seen, then watched the second half of the episode immediately after the completion of writing.

Sample A 1 i

The following part one is characterized by few three-letter words, many seven- and eight-letter words, few adverbs, many adjectives, few pronouns, adverbials placed sentence-initially, long sentences, and a section of commentary.

Charles, an elderly resident of a senior citizens' home, feels that his age is getting the better of him. Recently, he has been told by his son that he cannot live with him and his family because he has a small house and a new child on the way. Dejected, Charles returns to the home only to interrupt a game of Kick-the-can being played by some local children. After a brief appearance by Rod Serling, we are shown a scene that includes Charles and his roommate. The two discuss their experiences of getting old and how each interprets the significance of the phenomenon. The roommate believes that aging, both physically and socially, is inevitable and that one should live out one's "golden years" in dignity and solitude.

Charles, on the other hand, takes a different point of view. He believes that age is directly related to one's state of mind. More concretely, he claims that when children stop playing games they eventually start to get old. He sees a direct causal relationship between youth and playing Kick-the-can. Consequently, in order to regain his youth, he begins to act in childish ways, playing silly pranks and performing immature actions such as dashing through a sprinkler. Charles's roommate becomes concerned and speaks with the director of the retirement home regarding Charles's behaviour. The director appears concerned and tells the roommate that he will pay special attention to Charles lest he hurt himself. After the sprinkler incident, Charles is ordered under strict observation, an action that his roommate fears will kill Charles.

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.

Sample A 1 ii

The following part two is characterized by few three-letter words, many eight-letter words, and few adverbs.

Ben, Charles's roommate, tries to convince him that unless he begins to act his age, he will be put under strict observation. Charles accepts Ben's advice at first, but then decides that Kick-the-can is just too important to pass up. That night, he wakes the other men and women and tries to convince them the importance of taking a new perspective in relation to their lives. He tries to get them to remember how fun playing Kick-the-can was when they were children. His success as a recruiter is confirmed when the other residents agree that a game of Kick-the-can would be a good idea. Nostalgic feelings of mischief and fun are elicited by the crew as they remember how carefree they felt playing Kick-the-can as children. Unfortunately, Charles is less successful with Ben. Ben tells Charles that the realities of the physical body make playing a children's game absolutely out of the question. Charles, however, does not accept this logic. He confronts Ben, as he did the others, with the concept of magic. Magic, he believes, is something that is found with all people but especially with children. He tries to convince Ben however that magic can still exist as you age, if you only recognize. It was magic, claims Charles when he first kissed the woman that would become his wife. It was magic when his son was born. These romantic notions, however, fail to sway Ben. He remains determined in his opinions.

Thus, Charles and the others decide that they will have a go at Kick-the-can with or without Ben. After setting off a firecracker in order to distract the duty nurse, Charles and his crew sneak outside. Ben hears the noise and guesses what they are up to. In a vain attempt to stop the others, Ben wakes Mr. Cox, the superintendent of the institution. Both men dash outside in an attempt to stop the others from doing something they might regret later. However, as they get outside expecting to see a bunch of elderly people, they are confronted with a group of running, yelling, laughing children. Mr. Cox goes off to find the residents. Ben, however, finally comes to the realization that Charles's scheme worked—that Kick-the-can actually transformed (sic) him into a child.

Sample A 2 i

The following part one is characterized by few three- and four-letter words, many nine-letter words, many modifiers (particularly in sentence-initial position), and many prepositions.

The story is about the North American geriatric industry. It depicts a particular hospital-residence, and the story about one resident, who, after being disappointed at not being taken away, obsessively tries to pretend at childhood.

Everything about the film is depressing (as probably intended). The residence itself is overcrowded, with the residents sitting around waiting to die. Little private space exists in the image (and one is reminded of the geriatric industry's separation of spouses into common gender based rooms, along with the general assumption that old people don't have sex).

Explicitly, the film's protagonist starts off pseudo-sneakily leaving the residence, explaining that his son is to pick him up. This doesn't happen, but rather he finds himself abandoned on the street (under observation of the other inmates). As a group of children nearby are playing, one is led to believe that he seizes on the game of Kick-the-can as an escape (although he never plans this).

Upon returning to the residence, he becomes dotty, apparently believing that acting childish keeps one young. The other option presented is that, as an old person, he has no option but to quietly await death. This opposition is of course standard for television, so perhaps not weird.

Back at the residence, his childish behaviour is interpreted not as a psychological problem (or generic eccentricity) but rather as a sign of senility, which in the industry has grave implications. After a particularly active scene of wading in sprinkler (upsetting bourgeois values), he is put into isolation, with the probable expectation of reducing him to mindless boredom like everyone else.

Table 2: Comparison of sample A 1 i with all part twos in this appendix.

		Tags overall	Sentence-initial tags	Sentence-final tags	2-and 3-letter words	Word length	Sentence length
A 1 ii	Post	45.179 *	5.077	5.072	0.087	9.165	2.28
	Brill	49.677 *	3.628	4.706			
A 2 ii	Post	19.260	6.852	13.072*	0.130	15.108	0.62
	Brill	27.051	6.586	12.860*			
B 1 ii	Post	49.857 *	7.637	4.433	2.751	13.715	3.95*
	Brill	54.984 **	5.012	2.708			
B 2 ii	Post	41.925 *	3.119	1.714	4.342*	16.855	1.12
	Brill	47.472	2.134	2.984			
C 1 ii	Post	49.312 *	2.270	7.821*	0.048	20.217*	1.37
	Brill	50.140 **	1.476	7.532			

Sample A 2 ii

The following part two is characterized by few two-letter words, the use of intransitive verbs, and the placement of prepositional phrases sentence-initially.

In the second half of the film, the magical part associated with the series occurs. The protagonist rather than being isolated (as promised earlier) is left in the common bedroom, where a childhood friend (who'd earlier informed on him) warns him against further misbehaviour.

That night, the protagonist wakes all the other inmates (except his informer friend Ben), and convinces them to play Kick-the-can outside. They agree, sudden spark of life imagery included. He then wakes Ben, asking him to join, but Ben plays the role of depressing voice of false reason and refuses.

After creating a distraction outside to distract the night-nurse/guard, the inmates rush off to play. This distraction also awakens Ben, who immediately arises to awaken the residence's director (who in the film sleeps in his office). The director and Ben agree that great harm would be done by people playing a game, and so run outside to prevent this occurrence.

When the last two leave the building, they encounter the former inmates, who have metamorphosed into children upon playing. The director does not recognize them, and so rushes to the back in hopes of finding his aged dears. This leaves Ben alone, staring bleakly at the vanishing children, properly punished for his refusal to participate.

Table 3: Comparison of sample A 2 i with all part twos in this appendix.

		Tags overall	Sentence-initial tags	Sentence-final tags	2-and 3-letter words	Word length	Sentence length
A 1 ii	Post	70.676 **	8.282	N/A	0.0000851	21.553*	1.74
	Brill	69.435 **	6.816	N/A			
A 2 ii	Post	23.399	4.419	N/A	0.343	19.726*	0.46
	Brill	18.514	4.419	N/A			
B 1 ii	Post	57.031 **	11.072*	4.902	1.636	27.694*	2.99*
	Brill	55.756 **	8.510	5.527			
B 2 ii	Post	42.317 *	6.205	5.990	2.569	23.232*	0.86
	Brill	46.206	4.536	3.736			
C 1 ii	Post	60.912 **	3.723	N/A	0.238	33.776*	1.01
	Brill	56.071 **	2.748	N/A			

Condition B

In condition B, subjects wrote summaries of the first half, but were required to return the following week to complete the experiment.

Sample B 1 i

This part one is characterized by few two- and three-letter words, few modifiers, and long sentences.

The setting is an old people's home, out in the country. An elderly gentleman walks about the rooms, aided by a walking stick.

The house nurse notices one of the residents on the staircase: it's Charles, dressed in a suit and holding a suitcase. The nurse is surprised, not expecting anybody to be all ready for going out. Charles explains that his son is coming to pick him up. He seems pleased that he will finally be leaving the residence to go and live with his son. He shakes hands with other residents, bidding them farewell.

A car arrives beside the house, and Charles happily enters to meet his son. The residents are all watching intently, seemingly trying to imagine what it would feel like for somebody coming to "get them", and being happy for Charles at the same time. Charles's son the tells his father that he just came to talk about them living together, not to "come and get him".

There are children playing, first in the background, then in the foreground as Charles leaves the car. The residents are a little surprised and glad, thinking that he won't be leaving them after all. He watches one child kicking a can as the other children leave him behind. One kick sends the can near to Charles, and he picks it up. This is the turning point in the plot. The boy asks for the can back, but then leaves as Charles ignores him, holding the can with both hands, thinking deeply about his early years.

Rod Serling then comes into the picture, explaining that the man knows he will die here unless he can find a way to escape into the "twilight zone".

Next day, Charles and his friend, Ben, are watching the playful children from their window, and talking about them (and the noise they make—"enough to wake the dead"). Charles looks sad, and tells Ben how his son turned on him; that he has a wife and "kid", and doesn't want his father to live with them. Charles is thinking more, and asks Ben whether he believes in magic, when he stopped believing in magic, and why he no longer believes. Ben is skeptical, wondering what's going through his friend's mind. Then Charles, after this discussion with Ben which helped his idea along, believes he has found the secret of youth, though he doesn't explicitly say this.

It's another day, and Ben, concerned about his good friend, discusses their recent conversation with a doctor who is also a friend.

Back at the residence, Charles's behaviour is now shockingly different. He becomes playful, pushing an empty wheelchair, and making silly faces and noises at the other residents, just like a kid. He then runs through a lawn sprinkler whilst the others watch in horror; this finally gets the attention of the residence's superintendent, who ushers Charles back into the building and promises to put him in a special ward for observation, isolated from his peers. What a shame that this superintendent sees Charles as a threat to the local community, rather than an inspiration.

Table 4: Comparison of sample B 1 i with all part twos in this appendix.

		Tags overall	Sentence-initial tags	Sentence-final tags	2-and 3-letter words	Word length	Sentence length
A 1 ii	Post	43.289 *	4.353	2.381	0.618	15.581	1.78
	Brill	41.956 *	4.051	4.283			
A 2 ii	Post	18.518	7.512	6.736	1.681	19.719*	0.10
	Brill	13.060	8.301	13.477*			
B 1 ii	Post	28.760	5.248	1.324	0.587	5.869	3.54*
	Brill	29.352	4.678	3.124			
B 2 ii	Post	28.530	4.682	1.810	1.227	11.143	0.60
	Brill	39.079 *	5.143	2.525			
C 1 ii	Post	46.902 *	1.833	3.729	1.857	8.270	0.80
	Brill	42.019 *	1.628	3.176			

Sample B 1 ii
This part two is characterized by few adjectives, many prepositions sentence-initially, and short sentences.

Charles is angry about being put into the special ward, and Ben is giving him some company. Charles becomes thoughtful as night time approaches.

Everybody is sleeping, but Charles wakes up, with a plan in mind. He awakens everybody, one by one, all except for Ben. The residents assemble together into another room.

Charles begins to remember how it was like to be youthful, to play Kick-the-can. The others also start to reminisce. Then Charles tells them his secret, the secret of youth. They are all skeptical at first, but Charles manages to persuade them to take a shot at playing the game.

Charles goes back alone to the large bedroom and awakens Ben, asking him to join them. Ben tries to convince Charles to be realistic: they are old, and there is nothing they can do about it. Charles is not convinced, however; he still has hope. He associates the "magic" of playing Kick-the-can with the magic of being in love, of having his son.

Charles leaves Ben in the bedroom and rejoins the rest of the residents. They are watching the nurse from the top of the staircase, thinking how they can distract her. One of them throws a firecracker through a window: this alarms the nurse, who goes away to investigate. The old people quietly rush downstairs and out of the door.

Ben follows after, and alerts Mr. Cox, the superintendent. They both go outside, and can see children playing Kick-the-can on the street. Mr. Cox has no clue as to what is going on. However, Ben realizes what has happened. He feels abandoned. He talks to a kid, calling him "Charles", and begs to join in the game. But the other kids have left, and Charles, seemingly having no memory of Ben or the previous situation, runs away.

Sample B 2 i

This part one is characterized by many two- and eight-letter words, few five-letter words, few adverbs, many adjectives, few pronouns, few common nouns, many commas, and long sentences.

The scene opens on an elegant country home, panning down to a sign which reads: "Sunnyvale Rest". Instantly, the setting is revealed to be a rest home for the elderly.

Inside, people are sitting quietly, as if drugged. The viewer follows one slowly shuffling man through the rooms of this house; he is the only figure in motion. This rest home is a place where people wait to die.

Coming down the stairs is a happier figure who cheerfully announces, "My son is coming to get me!" as he doffs his hat and exits. He is mistaken, though, for on entering his son's car, he learns that he has misunderstood his son's intentions. "I said we could talk about it." his son explains weakly.

As the old man steps from the car, we know all too well the conclusion of the conversation. His son is busy and has no time; he is old and nothing more than baggage.

In the background, children have engaged in a game of Kick-the-can. The old man, broken-hearted, approaches the battered old tin can. There is clearly a sense of loss on his face: lost childhood, lost dreams, lost time. He picks up the can, a dented, useless piece of junk made valuable by the imagination of children.

The young boys do not reclaim their lost prize, seeming to fear the old man, or perhaps the future he represents.

Suddenly, from out of the bushes, Rod Serling appears to introduce this "Twilight Zone" episode. His most telling comment is that Mr. Whitley, the old man, will die in this rest home, unless he can escape "into the Twilight Zone". It is a trademark comment, but the "Zone" represents an imaginary place made real, a second childhood so full of life that it overcomes death.

The scene changes, and Mr. Whitley is watching children play from his upstairs window. He still holds the old tin can.

He begins to talk with his roommate and lifelong friend, circling in on the thoughts that most trouble him. Mr. Whitley refers to Kick-the-can as a "summertime ritual", and states dreamily that the game itself may be what keeps children young. For, as he observes, "the minute they stop, they grow old". He embarks on a monologue of the magic of childhood, and reaches a conclusion which no doubt will serve as the moral of the story: "Maybe the Fountain of Youth isn't a fountain, it's a way of looking at things." All the while he cradles the tin can, a symbol for the old becoming young again; the battered tin can is the centre of a child's game, as much as it is the focus of the story and Mr. Whitley's quest for a second childhood.

His roommate, out of concern, approaches the retirement home director, who says that a close eye will be kept on Mr. Whitley. Here we are exposed to a bitter irony: though treated like small children by the director, the elderly are labelled as "senile" the moment they behave like children.

Meanwhile, Mr. Whitley is busy trying to recapture his joie de vivre, and is attempting to stir up the same in his fellow house mates. He reminds them of their youth, catching polliwogs and running in the sprinkler. He then proceeds to do the latter, putting to work his theory of the Fountain of Youth.

Of course, the director sees this and decides that isolation will soon cure Mr. Whitley of his senility. "But that will kill him!" his roommate exclaims and so the climax is reached, and things seem as bleak as they can be.

Sample B 2 ii

This part two is characterized by many two-letter words, many adjectives, few pronouns, and high use of punctuation other than commas and periods.

The second half begins with Ben (Mr. Whitley's roommate and friend) warning Mr. Whitley (Charlie) to "act like everyone else" or he’ll be put in isolation. Charlie refuses to "sit like a vegetable" but realizes his dilemma. Already, the dichotomy of Charlie's youthful spirit and Ben's weary, aged existence is made clear in this short exchange. Here too, Charlie realizes he must find an escape—he cannot risk isolation, nor can he allow himself to sit and await death. A sideways glance shows him the key to his salvation: the battered old can.

Later, at night, Charlie puts on his coat and wakens the others, gathering them together for a game of Kick-the-can. "Maybe you gotta be a little crazy to make the magic work," he says, to explain away his odd behaviour. After a brief back-and-forth about aging, the elderly become swept up in a wave of nostalgia. "Did I tell you I used to be the fastest runner on my block?" one of them intones twice, in a manner simultaneously child-like and senile.

Mr. Whitley manages to bring the nostalgia to a fever-pitch of excitement, and calls for his friends to "Wake up! This is your last chance!" he tells them. And, finally, almost pleading "I can't play Kick-the-can alone!" It is this final plea which stirs them all to action: the realization that they cannot live in their tired, isolated worlds and that they must, in joining together, celebrate life.

Mr. Whitley returns to his room to encourage Ben to join in, but Ben is still the voice of reason. He is cold, dispassionate, and defeatist. He is old. Charlie, on the other hand, is youthful and insists wonderingly that "the magic must be out there." He recalls the magic of love, of birth, of friendship, and concludes that "maybe Kick-the-can is the greatest magic of all," returning to the symbol of childhood as an expression for his thoughts. Ben is adamant, however, and silently refuses to participate, leaving Charlie to rejoin the others.

They, in the meantime, have encountered an obstacle: the duty-nurse at the bottom of the stairs. They have already begun the journey back to childhood as they huddle together at the top of the stairs, up long past their bedtime. To escape, a string of firecrackers is thrown from the upstairs window to draw the nurse outside. (The firecrackers are no deus ex machina, they were alluded to by Mr. Cox, the director of the home, in his discussion with Ben).

Ben quickly realizes what is happening and rushes to alert Mr. Cox. There is fear in him, plainly, but not fear for his friends, rather, some unrealized fear that Charlie's magic threatens Ben's reality. He cannot accept that and so defends it as best he can—by turning to Mr. Cox.

The two race outside, but see only children engaged in a game of Kick-the-can. Mr. Cox runs off to find where the elderly have gotten to, but Ben already knows.

"Charlie," he says, recognizing a confused lad as his old childhood friend, "take me with you."

But the children run off, afraid of the adult world they have abandoned and Ben—too old to be young again—is left holding the rusty old can, realizing how very old he has become and how alone he now is.

Table 5: Comparison of sample B 2 i with all part twos in this appendix.

		Tags overall	Sentence-initial tags	Sentence-final tags	2-and 3-letter words	Word length	Sentence length
A 1 ii	Post	50.816*	9.734*	1.874	3.202	16.738	0.77
	Brill	51.608**	8.539	1.462			
A 2 ii	Post	15.974	6.717	6.938	4.488*	25.569*	-0.84
	Brill	15.119	6.071	6.836			
B 1 ii	Post	45.852*	12.372*	1.786	0.017	10.567	2.56
	Brill	42.319*	12.736*	2.941			
B 2 ii	Post	20.735	8.466	2.096	0.002	10.544	-0.34
	Brill	27.010	7.873	1.306			
C 1 ii	Post	52.294**	6.740	4.068	5.730*	20.693*	-0.28
	Brill	51.755**	5.540	4.197			

Condition C

In condition C, subjects viewed the first half, but did not write about it. Instead, they were given someone else’s description to read. After reading part one, they watched the second half of the video. They were asked to complete the description of the first half after that.

Sample C 1 ii
This sample is characterized by few two- and three-letter words, many five- and seven-letter words, many modifiers, and long sentences.

The director, after trying to reassure the roommate that it is quite natural to grow senile, decides to isolate Mr. Whitely, supposedly for his own good.

Hearing about his impending isolation strengthens Mr. Whitley's resolve about seeing if it is possible to somehow regain his youth. His observations of the young kids playing Kick-the-can convinces him that the game itself contains magic that empowers kids to be young.

That night Whitley wakes up the residents of the home. They convene in a common area where Mr. Whitley (Charles) tells them of his belief in a simple child's game. Naturally hesitant because they too believe that Mr. Whitley is senile because of his previous stunt, running through the sprinkler. Slowly they begin to reassess their fears about engaging in something in which old people "aren't supposed to". They are naturally concerned, fearing that their bodies won't hold up like they used to. Finally, after Mr. Whitley tells the other old people that this could be the last time they would have in doing something young, they decide to take part in Mr. Whitley's bold venture called Kick-the-can.

Mr. Whitley devises a plan and then goes upstairs to tell Ben (his roommate) about their plans. Ben refuses to play, even though his lifelong friend Mr. Whitley claims that they always did everything together.

As a diversion they needed to light firecrackers to draw the attendant nurse away from the night desk, outside. This was because they had to slip through the lobby to the back door outside. All the residents except Ben made it outside.

Ben decides to wake up the director of the home and tell him about Mr. Whitley's plans. The director and Ben quickly run outside so that they can talk sense to the inspired residents.

Outside though instead of old people playing Kick-the-can or hide and seek there are young children. With disbelief on his face he sees Mr. Whitley, as he once was; a young boy. He recognizes him and realizes that Whitley was right, there is magic, you can't always disbelieve. He calls for the now young Whitley to take him, make him young, almost begging. The boy staring at Ben, almost scared, turns and runs into the forest with the other kids. Ben walks away telling the director not to bother searching for the old.

Appendix D: Tagged writing sample

Due to space considerations, one sample from the appendix of samples (B1) was included here to provide an example of the tagging results from each of the taggers. See Appendix B for the Penn tagset. The complete tagged writing samples are available from the author upon request.

POST-tagged sample

B 1 i

The [DT] setting [NN] is [VBZ] an [DT] old [JJ] people [NNS] 's [POS] home [NN] , [,] out [RB] in [IN] the [DT] country [NN] . [.]
An [DT] elderly [JJ] gentleman [NN] walks [VBZ] about [IN] the [DT] rooms [NNS] , [,] aided [VBN] by [IN] a [DT] walking [VBG] stick [NN] . [.]

The [DT] house [NN] nurse [NN] notices [VBZ] one [CD] of [IN] the [DT] residents [NNS] on [IN] the [DT] staircase [NN] : [;] it [PP] 's [VBZ] Charles [NP] , [,] dressed [VBN] in [IN] a [DT] suit [NN] and [CC] holding [VBG] a [DT] suitcase [NN] . [.]
The [DT] nurse [NN] is [VBZ] surprised [VBN] , [,] not [RB] expecting [VBG] anybody [NN] to [TO] be [VB] all [DT] ready [JJ] for [IN] going [VBG] out [RP] . [.]
Charles [NP] explains [VBZ] that [IN] his [PP$] son [NN] is [VBZ] coming [VBG] to [TO] pick [VB] him [PP] up [RP] . [.]
He [PP] seems [VBZ] pleased [VBN] that [IN] he [PP] will [MD] finally [RB] be [VB] leaving [VBG] the [DT] residence [NN] to [TO] go [VB] and [CC] live [VB] with [IN] his [PP$] son [NN] . [.]
He [PP] shakes [VBZ] hands [NNS] with [IN] other [JJ] residents [NNS] , [,] bidding [VBG] them [PP] farewell [NN] . [.]

A [DT] car [NN] arrives [VBZ] beside [IN] the [DT] house [NN] , [,] and [CC] Charles [NP] happily [RB] enters [VBZ] to [TO] meet [VB] his [PP$] son [NN] . [.]
The [DT] residents [NNS] are [VBP] all [DT] watching [VBG] intently [RB] , [,] seemingly [RB] trying [VBG] to [TO] imagine [VB] what [WP] it [PP] would [MD] feel [VB] like [IN] for [IN] somebody [NN] coming [VBG] to [TO] " ['] get [VB] them [PP] " ['] , [,] and [CC] being [VBG] happy [JJ] for [IN] Charles [NP] at [IN] the [DT] same [JJ] time [NN] . [.]
Charles [NP] 's [POS] son [NN] the [DT] tells [VBZ] his [PP$] father [NN] that [IN] he [PP] just [RB] came [VBD] to [TO] talk [VB] about [IN] them [PP] living [VBG] together [RB] , [,] not [RB] to [TO] " ['] come [VB] and [CC] get [VB] him [PP] " ['] . [.]

There [EX] are [VBP] children [NNS] playing [VBG] , [,] first [RB] in [IN] the [DT] background [NN] , [,] then [RB] in [IN] the [DT] foreground [NN] as [IN] Charles [NP] leaves [VBZ] the [DT] car [NN] . [.]
The [DT] residents [NNS] are [VBP] a [DT] little [JJ] surprised [JJ] and [CC] glad [JJ] , [,] thinking [VBG] that [IN] he [PP] wo [MD] n't [RB] be [VB] leaving [VBG] them [PP] after [IN] all [DT] . [.]
He [PP] watches [VBZ] one [CD] child [NN] kicking [VBG] a [DT] can [NN] as [IN] the [DT] other [JJ] children [NNS] leave [VBP] him [PP] behind [RB] . [.]
One [PP] kick [VBP] sends [VBZ] the [DT] can [NN] near [RB] to [TO] Charles [NP] , [,] and [CC] he [PP] picks [VBZ] it [PP] up [RP] . [.]
This [DT] is [VBZ] the [DT] turning [VBG] point [NN] in [IN] the [DT] plot [NN] . [.]
The [DT] boy [NN] asks [VBZ] for [IN] the [DT] can [NN] back [RB] , [,] but [CC] then [RB] leaves [VBZ] as [IN] Charles [NP] ignores [VBZ] him [PP] , [,] holding [VBG] the [DT] can [NN] with [IN] both [DT] hands [NNS] , [,] thinking [VBG] deeply [RB] about [IN] his [PP$] early [JJ] years [NNS] . [.]

Rod [NP] Serling [NP] then [RB] comes [VBZ] into [IN] the [DT] picture [NN] , [,] explaining [VBG] that [IN] the [DT] man [NN] knows [VBZ] he [PP] will [MD] die [VB] here [RB] unless [IN] he [PP] can [MD] find [VB] a [DT] way [NN] to [TO] escape [VB] into [IN] the [DT] " ['] twilight [NN] zone [NN] " ['] . [.]
Next [JJ] day [NN] , [,] Charles [NP] and [CC] his [PP$] friend [NN] , [,] Ben [NP] , [,] are [VBP] watching [VBG] the [DT] playful [JJ] children [NNS] from [IN] their [PP$] window [NN] , [,] and [CC] talking [VBG] about [IN] them [PP] ([(] and [CC] the [DT] noise [NN] they [PP] make [VBP]) [)] - [;] " ['] enough [RB] to [TO] wake [VB] the [DT] dead [JJ] " [']) [)] . [.]
Charles [NP] looks [VBZ] sad [JJ] , [,] and [CC] tells [VBZ] Ben [NP] how [WRB] his [PP$] son [NN] turned [VBD] on [IN] him [PP] ; [;] that [WDT] he [PP] has [VBZ] a [DT] wife [NN] and [CC] " ['] kid [NN] " ['] , [,] and [CC] does [VBZ] n't [RB] want [VB] his [PP$] father [NN] to [TO] live [VB] with [IN] them [PP] . [.]
Charles [NP] is [VBZ] thinking [VBG] more [JJR] , [,] and [CC] asks [VBZ] Ben [NP] whether [IN] he [PP] believes [VBZ] in [IN] magic [NN] , [,] when [WRB] he [PP] stopped [VBD] believing [VBG] in [IN] magic [NN] , [,] and [CC] why [WRB] he [PP] no [RB] longer [RB] believes [VBZ] . [.]
Ben [NP] is [VBZ] skeptical [JJ] , [,] wondering [VBG] what [WP] 's [VBZ] going [VBG] through [IN] his [PP$] friend [NN] 's [POS] mind [NN] . [.]
Then [RB] Charles [NP] , [,] after [IN] this [DT] discussion [NN] with [IN] Ben [NP] which [WDT] helped [VBD] his [PP$] idea [NN] along [IN] , [,] believes [VBZ] he [PP] has [VBZ] found [VBN] the [DT] secret [NN] of [IN] youth [NN] , [,] though [IN] he [PP] does [VBZ] n't [RB] explicitly [RB] say [VB] this [DT] . [.]

It [PP] 's [VBZ] another [DT] day [NN] , [,] and [CC] Ben [NP] , [,] concerned [VBN] about [IN] his [PP$] good [JJ] friend [NN] , [,] discusses [VBZ] their [PP$] recent [JJ] conversation [NN] with [IN] a [DT] doctor [NN] who [WP] is [VBZ] also [RB] a [DT] friend [NN] . [.]

Back [RB] at [IN] the [DT] residence [NN] , [,] Charles [NP] 's [POS] behaviour [NN] is [VBZ] now [RB] shockingly [RB] different [JJ] . [.]
He [PP] becomes [VBZ] playful [JJ] , [,] pushing [VBG] an [DT] empty [JJ] wheelchair [NN] , [,] and [CC] making [VBG] silly [JJ] faces [NNS] and [CC] noises [NNS] at [IN] the [DT] other [JJ] residents [NNS] , [,] just [RB] like [IN] a [DT] kid [NN] . [.]
He [PP] then [RB] runs [VBZ] through [IN] a [DT] lawn [NN] sprinkler [NN] whilst [IN] the [DT] others [NNS] watch [VBP] in [IN] horror [NN] ; [;] this [DT] finally [RB] gets [VBZ] the [DT] attention [NN] of [IN] the [DT] residence [NN] 's [POS] superintendent [NN] , [,] who [WP] ushers [VBZ] Charles [NP] back [RB] into [IN] the [DT] building [NN] and [CC] promises [VBZ] to [TO] put [VB] him [PP] in [IN] a [DT] special [JJ] ward [NN] for [IN] observation [NN] , [,] isolated [VBN] from [IN] his [PP$] peers [NNS] . [.]
What [WP] a [DT] shame [NN] that [IN] this [DT] superintendent [NN] sees [VBZ] Charles [NP] as [IN] a [DT] threat [NN] to [TO] the [DT] local [JJ] community [NN] , [,] rather [RB] than [IN] an [DT] inspiration [NN] . [.]

B 1 ii
Charles [NP] is [VBZ] angry [JJ] about [IN] being [VBG] put [VBN] into [IN] the [DT] special [JJ] ward [NN] , [,] and [CC] Ben [NP] is [VBZ] giving [VBG] him [PP] some [DT] company [NN] . [.]
Charles [NP] becomes [VBZ] thoughtful [JJ] as [IN] night [NN] time [NN] approaches [NNS] . [.]

Everybody [NN] is [VBZ] sleeping [VBG] , [,] but [CC] Charles [NP] wakes [VBZ] up [RB] , [,] with [IN] a [DT] plan [NN] in [IN] mind [NN] . [.]
He [PP] awakens [VBZ] everybody [NN] , [,] one [CD] by [IN] one [CD] , [,] all [DT] except [IN] for [IN] Ben [NP] . [.]

The [DT] residents [NNS] assemble [VBP] together [RB] into [IN] another [DT] room [NN] . [.]

Charles [NP] begins [VBZ] to [TO] remember [VB] how [WRB] it [PP] was [VBD] like [IN] to [TO] be [VB] youthful [JJ] , [,] to [TO] play [VB] Kick-the-can [NP] . [.]
The [DT] others [NNS] also [RB] start [VB] to [TO] reminisce [VB] . [.]
Then [RB] Charles [NP] tells [VBZ] them [PP] his [PP$] secret [NN] , [,] the [DT] secret [NN] of [IN] youth [NN] . [.]
They [PP] are [VBP] all [DT] skeptical [JJ] at [IN] first [JJ] , [,] but [CC] Charles [NP] manages [VBZ] to [TO] persuade [VB] them [PP] to [TO] take [VB] a [DT] shot [NN] at [IN] playing [VBG] the [DT] game [NN] . [.]

Charles [NP] goes [VBZ] back [RB] alone [RB] to [TO] the [DT] large [JJ] bedroom [NN] and [CC] awakens [VBZ] Ben [NP] , [,] asking [VBG] him [PP] to [TO] join [VB] them [PP] . [.]
Ben [NP] tries [VBZ] to [TO] convince [VB] Charles [NP] to [TO] be [VB] realistic [JJ] : [;] they [PP] are [VBP] old [JJ] , [,] and [CC] there [EX] is [VBZ] nothing [NN] they [PP] can [MD] do [VB] about [IN] it [PP] . [.]
Charles [NP] is [VBZ] not [RB] convinced [VBN] , [,] however [RB] ; [;] he [PP] still [RB] has [VBZ] hope [NN] . [.]
He [PP] associates [VBZ] the [DT] " ['] magic [JJ] " ['] of [IN] playing [VBG] Kick-the-can [NP] with [IN] the [DT] magic [NN] of [IN] being [VBG] in [IN] love [NN] , [,] of [IN] having [VBG] his [PP$] son [NN] . [.]

Charles [NP] leaves [VBZ] Ben [NP] in [IN] the [DT] bedroom [NN] and [CC] rejoins [VBZ] the [DT] rest [NN] of [IN] the [DT] residents [NNS] . [.]
They [PP] are [VBP] watching [VBG] the [DT] nurse [NN] from [IN] the [DT] top [NN] of [IN] the [DT] staircase [NN] , [,] thinking [VBG] how [WRB] they [PP] can [MD] distract [VB] her [PP] . [.]
One [PP] of [IN] them [PP] throws [VBZ] a [DT] firecracker [NN] through [IN] a [DT] window [NN] : [;] this [DT] alarms [NNS] the [DT] nurse [NN] , [,] who [WP] goes [VBZ] away [RB] to [TO] investigate [VB] . [.]
The [DT] old [JJ] people [NNS] quietly [RB] rush [VB] downstairs [NN] and [CC] out [RB] of [IN] the [DT] door [NN] . [.]

Ben [NP] follows [VBZ] after [RB] , [,] and [CC] alerts [VBZ] Mr. [NP] Cox [NP] , [,] the [DT] superintendent [NN] . [.]
They [PP] both [CC] go [VB] outside [JJ] , [,] and [CC] can [MD] see [VB] children [NNS] playing [VBG] Kick-the-can [NP] on [IN] the [DT] street [NN] . [.]
Mr. [NP] Cox [NP] has [VBZ] no [DT] clue [NN] as [IN] to [TO] what [WP] is [VBZ] going [VBG] on [RP] . [.]
However [RB] , [,] Ben [NP] realizes [VBZ] what [WP] has [VBZ] happened [VBN] . [.]
He [PP] feels [VBZ] abandoned [VBN] . [.]
He [PP] talks [VBZ] to [TO] a [DT] kid [NN] , [,] calling [VBG] him [PP] " ['] Charles [NP] " ['] , [,] and [CC] begs [VBZ] to [TO] join [VB] in [IN] the [DT] game [NN] . [.]
But [CC] the [DT] other [JJ] kids [NNS] have [VBP] left [VBN] , [,] and [CC] Charles [NP] , [,] seemingly [RB] having [VBG] no [DT] memory [NN] of [IN] Ben [NP] or [CC] the [DT] previous [JJ] situation [NN] , [,] runs [VBZ] away [RB] . [.]

Brill-tagged sample

B 1 i

The [DT] setting [VBG] is [VBZ] an [DT] old [JJ] people [NNS] 's [POS] home [NN] ,[,] out [IN] in [IN] the [DT] country [NN] . [.]

An [DT] elderly [JJ] gentleman [NN] walks [VBZ] about [IN] the [DT] rooms [NNS] , [,] aided [VBN] by [IN] a [DT] walking [VBG] stick [NN] . [.]

The [DT] house [NN] nurse [NN] notices [VBZ] one [CD] of [IN] the [DT] residents [NNS] on [IN] the [DT] staircase [NN] : [:] it [PRP] 's [VBZ] Charles [NNP] , [,] dressed [VBN] in [IN] a [DT] suit [NN] and [CC] holding [VBG] a [DT] suitcase [NN] . [.]
The [DT] nurse [NN] is [VBZ] surprised [VBN] , [,] not [RB] expecting [VBG] anybody [NN] to [TO] be [VB] all [DT] ready [JJ] for [IN] going [VBG] out [RB] . [.]
Charles [NNP] explains [VBZ] that [IN] his [PRP$] son [NN] is [VBZ] coming [VBG] to [TO] pick [VB] him [PRP] up [RB] . [.]
He [PRP] seems [VBZ] pleased [VBN] that [IN] he [PRP] will [MD] finally [RB] be [VB] leaving [VBG] the [DT] residence [NN] to [TO] go [VB] and [CC] live [VB] with [IN] his [PRP$] son [NN] . [.]
He [PRP] shakes [VBZ] hands [NNS] with [IN] other [JJ] residents [NNS] , [,] bidding [NN] them [PRP] farewell [NN] . [.]

A [DT] car [NN] arrives [VBZ] beside [IN] the [DT] house [NN] , [,] and [CC] Charles [NNP] happily [RB] enters [VBZ] to [TO] meet [VB] his [PRP$] son [NN] . [.]
The [DT] residents [NNS] are [VBP] all [DT] watching [VBG] intently [RB] , [,] seemingly [RB] trying [VBG] to [TO] imagine [VB] what [WP] it [PRP] would [MD] feel [VB] like [IN] for [IN] somebody [NN] coming [VBG] to [TO] " ["] get [VB] them [PRP] " ["] , [,] and [CC] being [VBG] happy [JJ] for [IN] Charles [NNP] at [IN] the [DT] same [JJ] time [NN] . [.]
Charles [NNP] 's [POS] son [NN] the [DT] tells [VBZ] his [PRP$] father [NN] that [IN] he [PRP] just [RB] came [VBD] to [TO] talk [VB] about [IN] them [PRP] living [VBG] together [RB] , [,] not [RB] to [TO] " ["] come [VB] and [CC] get [VB] him [PRP] " ["] . [.]

There [EX] are [VBP] children [NNS] playing [VBG] , [,] first [JJ] in [IN] the [DT] background [NN] , [,] then [RB] in [IN] the [DT] foreground [NN] as [IN] Charles [NNP] leaves [VBZ] the [DT] car [NN] . [.]
The [DT] residents [NNS] are [VBP] a [DT] little [JJ] surprised [VBN] and [CC] glad [JJ] , [,] thinking [VBG] that [IN] he [PRP] wo [MD] n't [RB] be [VB] leaving [VBG] them [PRP] after [IN] all [DT] . [.]
He [PRP] watches [VBZ] one [CD] child [NN] kicking [VBG] a [DT] can [NN] as [IN] the [DT] other [JJ] children [NNS] leave [VBP] him [PRP] behind [RB] . [.]
One [CD] kick [NN] sends [VBZ] the [DT] can [NN] near [VB] to [TO] Charles [NNP] , [,] and [CC] he [PRP] picks [VBZ] it [PRP] up [RB] . [.]
This [DT] is [VBZ] the [DT] turning [VBG] point [NN] in [IN] the [DT] plot [NN] . [.]
The [DT] boy [NN] asks [VBZ] for [IN] the [DT] can [NN] back [RB] , [,] but [CC] then [RB] leaves [VBZ] as [IN] Charles [NNP] ignores [VBZ] him [PRP] , [,] holding [VBG] the [DT] can [NN] with [IN] both [DT] hands [NNS] , [,] thinking [VBG] deeply [RB] about [IN] his [PRP$] early [JJ] years [NNS] . [.]

Rod [NNP] Serling [NNP] then [RB] comes [VBZ] into [IN] the [DT] picture [NN] , [,] explaining [VBG] that [IN] the [DT] man [NN] knows [VBZ] he [PRP] will [MD] die [VB] here [RB] unless [IN] he [PRP] can [MD] find [VB] a [DT] way [NN] to [TO] escape [VB] into [IN] the [DT] " ["] twilight [NN] zone [NN] " ["] . [.]

Next [JJ] day [NN] , [,] Charles [NNP] and [CC] his [PRP$] friend [NN] , [,] Ben [NNP] , [,] are [VBP] watching [VBG] the [DT] playful [JJ] children [NNS] from [IN] their [PRP$] window [NN] , [,] and [CC] talking [VBG] about [IN] them [PRP] ([(] and [CC] the [DT] noise [NN] they [PRP] make [VBP]) [SYM] - [:] " ["] enough [RB] to [TO] wake [VB] the [DT] dead [JJ] " ["] . [.]

Charles [NNP] looks [VBZ] sad, [CD] and [CC] tells [VBZ] Ben [NNP] how [WRB] his [PRP$] son [NN] turned [VBD] on [IN] him [PRP] ; [:] that [IN] he [PRP] has [VBZ] a [DT] wife [NN] and [CC] " ["] kid [NN] " ["] , [,] and [CC] does [VBZ] n't [RB] want [VB] his [PRP$] father [NN] to [TO] live [VB] with [IN] them [PRP] . [.]
Charles [NNP] is [VBZ] thinking [VBG] more [JJR] , [,] and [CC] asks [VBZ] Ben [NNP] whether [IN] he [PRP] believes [VBZ] in [IN] magic [NN] , [,] when [WRB] he [PRP] stopped [VBD] believing [VBG] in [IN] magic [NN] , [,] and [CC] why [WRB] he [PRP] no [RB] longer [RB] believes [VBZ] . [.]
Ben [NNP] is [VBZ] skeptical [JJ] , [,] wondering [VBG] what [WP] 's [POS] going [VBG] through [IN] his [PRP$] friend [NN] 's [POS] mind [NN] . [.]
Then [RB] Charles [NNP] , [,] after [IN] this [DT] discussion [NN] with [IN] Ben [NNP] which [WDT] helped [VBD] his [PRP$] idea [NN] along [RB] , [,] believes [VBZ] he [PRP] has [VBZ] found [VBN] the [DT] secret [NN] of [IN] youth [NN] , [,] though [IN] he [PRP] does [VBZ] n't [RB] explicitly [RB] say [VB] this [DT] . [.]

It [PRP] 's [VBZ] another [DT] day [NN] , [,] and [CC] Ben [NNP] , [,] concerned [VBN] about [IN] his [PRP$] good [JJ] friend [NN] , [,] discusses [VBZ] their [PRP$] recent [JJ] conversation [NN] with [IN] a [DT] doctor [NN] who [WP] is [VBZ] also [RB] a [DT] friend [NN] . [.]

Back [RB] at [IN] the [DT] residence [NN] , [,] Charles [NNP] 's [POS] behaviour [NN] is [VBZ] now [RB] shockingly [RB] different [JJ] . [.]
He [PRP] becomes [VBZ] playful [JJ] , [,] pushing [VBG] an [DT] empty [JJ] wheelchair [NN] , [,] making [VBG] silly [JJ] faces [NNS] and [CC] noises [NNS] at [IN] the [DT] other [JJ] residents [NNS] , [,] just [RB] like [IN] a [DT] kid [NN] . [.]
He [PRP] then [RB] runs [VBZ] through [IN] a [DT] lawn [NN] sprinkler [NN] whilst [NN] the [DT] others [NNS] watch [NN] in [IN] horror [NN] ; [:] this [DT] finally [RB] gets [VBZ] the [DT] attention [NN] of [IN] the [DT] residence [NN] 's [POS] superintendent [NN] , [,] who [WP] ushers [VBZ] Charles [NNP] back [RB] into [IN] the [DT] building [NN] and [CC] promises [VBZ] to [TO] put [VB] him [PRP] in [IN] a [DT] special [JJ] ward [NN] for [IN] observation, [CD] isolated [VBD] from [IN] his [PRP$] peers [NNS] . [.]
What [WP] a [DT] shame [NN] that [IN] this [DT] superintendent [NN] sees [VBZ] Charles [NNP] as [IN] a [DT] threat [NN] to [TO] the [DT] local [JJ] community [NN] , [,] rather [RB] than [IN] an [DT] inspiration [NN] . [.]

B 1 ii
Charles [NNP] is [VBZ] angry [JJ] about [IN] being [VBG] put [VB] into [IN] the [DT] special [JJ] ward [NN] , [,] and [CC] Ben [NNP] is [VBZ] giving [VBG] him [PRP] some [DT] company [NN] . [.]
Charles [NNP] becomes [VBZ] thoughtful [JJ] as [IN] night [NN] time [NN] approaches [NNS] . [.]

Everybody [NN] is [VBZ] sleeping [VBG] , [,] but [CC] Charles [NNP] wakes [VBZ] up [RB] , [,] with [IN] a [DT] plan [NN] in [IN] mind [NN] . [.]
He [PRP] awakens [VBZ] everybody [NN] , [,] one [CD] by [IN] one [CD] , [,] all [DT] except [IN] for [IN] Ben [NNP] . [.]
The [DT] residents [NNS] assemble [VBP] together [RB] into [IN] another [DT] room [NN] . [.]

Charles [NNP] begins [VBZ] to [TO] remember [VB] how [WRB] it [PRP] was [VBD] like [IN] to [TO] be [VB] youthful [JJ] , [,] to [TO] play [VB] Kick-the-can [NNP] . [.]
The [DT] others [NNS] also [RB] start [VB] to [TO] reminisce [NN] . [.]
Then [RB] Charles [NNP] tells [VBZ] them [PRP] his [PRP$] secret [JJ] , [,] the [DT] secret [NN] of [IN] youth [NN] . [.]

They [PRP] are [VBP] all [DT] skeptical [JJ] at [IN] first [JJ] , [,] but [CC] Charles [NNP] manages [VBZ] to [TO] persuade [VB] them [PRP] to [TO] take [VB] a [DT] shot [NN] at [IN] playing [VBG] the [DT] game [NN] . [.]

Charles [NNP] goes [VBZ] back [RB] alone [RB] to [TO] the [DT] large [JJ] bedroom [NN] and [CC] awakens [VBZ] Ben [NNP] , [,] asking [VBG] him [PRP] to [TO] join [VB] them [PRP] . [.]
Ben [NNP] tries [VBZ] to [TO] convince [VB] Charles [NNP] to [TO] be [VB] realistic [JJ] : [:] they [PRP] are [VBP] old [JJ] , [,] and [CC] there [EX] is [VBZ] nothing [NN] they [PRP] can [MD] do [VB] about [IN] it [PRP] . [.]
Charles [NNP] is [VBZ] not [RB] convinced [VBN] , [,] however [RB] ; [:] he [PRP] still [RB] has [VBZ] hope [NN] . [.]
He [PRP] associates [VBZ] the [DT] " ["] magic [NN] " ["] of [IN] playing [VBG] Kick-the-can [NNP] with [IN] the [DT] magic [NN] of [IN] being [VBG] in [IN] love [NN] , [,] of [IN] having [VBG] his [PRP$] son [NN] . [.]

Charles [NNP] leaves [VBZ] Ben [NNP] in [IN] the [DT] bedroom [NN] and [CC] rejoins [VBZ] the [DT] rest [NN] of [IN] the [DT] residents [NNS] . [.]
They [PRP] are [VBP] watching [VBG] the [DT] nurse [NN] from [IN] the [DT] top [NN] of [IN] the [DT] staircase [NN] , [,] thinking [VBG] how [WRB] they [PRP] can [MD] distract [VB] her [PRP$] . [.]
One [CD] of [IN] them [PRP] throws [VBZ] a [DT] firecracker [NN] through [IN] a [DT] window [NN] : [:] this [DT] alarms [NNS] the [DT] nurse [NN] , [,] who [WP] goes [VBZ] away [RB] to [TO] investigate [VB] . [.]
The [DT] old [JJ] people [NNS] quietly [RB] rush [VBP] downstairs [NN] and [CC] out [IN] of [IN] the [DT] door [NN] . [.]

Ben [NNP] follows [VBZ] after [RB] , [,] and [CC] alerts [VBZ] Mr. [NNP] Cox [NNP] , [,] the [DT] superintendent [NN] . [.]
They [PRP] both [DT] go [NN] outside [RB] , [,] and [CC] can [MD] see [VB] children [NNS] playing [VBG] Kick-the-can [NNP] on [IN] the [DT] street [NN] . [.]
Mr. [NNP] Cox [NNP] has [VBZ] no [DT] clue [NN] as [RB] to [TO] what [WP] is [VBZ] going [VBG] on [RB] . [.]
However [RB] , [,] Ben [NNP] realizes [VBZ] what [WP] has [VBZ] happened [VBN] . [.]
He [PRP] feels [VBZ] abandoned [VBN] . [.]
He [PRP] talks [VBZ] to [TO] a [DT] kid [NN] , [,] calling [VBG] him [PRP] " ["] Charles [NNP] " ["] , [,] and [CC] begs [VBZ] to [TO] join [VB] in [IN] the [DT] game [NN] . [.]
But [CC] the [DT] other [JJ] kids [NNS] have [VBP] left [VBN] , [,] and [CC] Charles [NNP] , [,] seemingly [RB] having [VBG] no [DT] memory [NN] of [IN] Ben [NNP] or [CC] the [DT] previous [JJ] situation [NN] , [,] runs [VBZ] away [RB] . [.]