Ensuring Stylistic Congruity in Collaboratively Written Text: Requirements Analysis and Design Issues

by

Melanie A. Baljko

Department of Computer Science
University of Toronto
Toronto, Canada
May 1997

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Ensuring Stylistic Congruity in Collaboratively Written Text: Requirements Analysis and Design Issues

Melanie A. Baljko
M.Sc Thesis, 1997
Department of Computer Science
University of Toronto

Abstract

Often, texts that have been written collaboratively do not "speak with a single voice." Eliminating stylistic incongruity, a difficult undertaking for both collaborative and singular writers, is the desired function of a software tool. This thesis describes the first cycle of an iterative software development process towards meeting this goal. The user requirements are analyzed with respect to a model that synthesizes established research, and then the requirements are taxonomized. Then, a framework for performing computational stylistic assessments is developed for later tool design. An experiment designed to measure the subjectivity in stylistic assessment — a relevant issue for making deterministic, computational stylistic assessments — was performed; the results indicate that future stylistic assessment tools must account for different patterns of assessment. Several design directions motivated by these results are suggested.

Acknowledgments

After writing so much about collaborative writing and style, I am happy to take a step back and write about the assistance and support that I received throughout the research and writing process.

First, I would like to thank my supervisor, Graeme Hirst, for his many insightful observations, his detailed comments and reviews, his availability for discussions, and his tireless pursuit of transcendent prose (those many multiple-hour review sessions are now in the back of my mind every time I set out to write). I am also very grateful to Marilyn Mantei, my second reader, for sharing her expertise in experimental design, for posing intriguing questions, and for reading my thesis and making many insightful comments.

I am grateful to Kevin Schlueter for our many discussions and for helping me develop the analysis technique for my experimental data. I am also thankful for having such a splendid office-mate, Daniel Marcu, who not only is up for late-night theoretical questions and discussions (which helped me develop and clarify my ideas), but also for kicking the butt in general and making our office a joyous enclave of eminbicity™ — and where eminbicity is, Alex Budanitsky soon follows. In addition to Alex, I would like to thank all the members of the computational linguistics group for sharing their views and resources with me.

I am grateful for the financial support that I have received from the University of Toronto, the Natural Sciences and Engineering Research Council of Canada, and my parents. I also want to thank Professor Chrysanne DiMarco at the University of Waterloo for giving me the encouragement and the support to start graduate studies.

Most importantly, I would like to thank my friends and family. This thesis could not exist if not for their support and patience. I am so fortunate to have the *najbolje* parents; I would like to say thank you for all the unwavering support for as long as I have known. I also want to thank my sister and brother, Christine and Jeffrey, for their patience; this thesis has caused them all kinds of indirect torment and they tolerated it gracefully (get ready, there is more to come...). And, to save the best for last, I want to thank Bilaki. He volunteered to teach me bridge and has proved to be an excellent *partneraki*.

Contents

Chapter 1

Introduction

1.1 Stylistic Congruity in Text

A requisite for readable, polished text is that the style of its segments must harmonize, or be *stylistically congruous*. While harmony doesn't imply sameness, which causes a text to be lack-lustre, it does ensure that the segments cohere together to form a unified text. Without this, a text doesn't *speak with a single voice*; something is *out of place*; it *doesn't sound right*, to use some common metaphors. A disorderly set of styles within a text reflects poorly on the professionalism of the authors of the text (Farkas, 1985) and detracts from the effectiveness of the text's communication to its readers. So while stylistic congruity is not the only requisite for a high-quality text, it is crucial and, furthermore, it is a requisite that writers find especially troublesome to fulfill.

Example 1, shown in figure 1.1 below, is taken from a LaTeXreference manual and is stylistically congruous. However, Example 2, a modification of Example 1, doesn't speak with a single voice. As later discussions will show, aspects of the author's communication, such as formality and intended interpersonal distance, are conveyed in the style of a text and shape its stylistic quality. This stylistic inconsistency in interpersonal distance between the paragraphs in Example 2 results in stylistic incongruity. For example, the first paragraph conveys a respective directness in the communication with short sentences, with a neutral explanation of a third party's limitations. In contrast, the second, longer-winded paragraph is patronizing and makes directs reference to the reader's assumed lack of design abilities.

Although writing containing stylistic incongruities such as these could have been pro-

Example 1	***Example 2***
The use of a consistent layout throughout a document helps the reader understand the various visual clues associated with a given component. It allows the document to be reused for producing online documentation, or eases the automatic extraction of information via predefined keywords. One should bear in mind the fact that typography is a creative skill, requiring a level of experience and craftsmanship that is rarely found in the untrained layman. Therefore, the development of a new style is better left to specialist designers, and casual users should restrict themselves mostly to *small* and *consistent* modifications to an already existing style. Extreme care should be taken not to upset the subtle visual balance between the various document elements.	A consistent layout should be used throughout a document because it helps the reader. They need to understand the various visual clues associated with a given component. By sticking to a consistent layout, the document can also be reused for producing online documentation. If not, at least information can be automatically extracted via predefined keywords. Since craftsmanship in typography requires creative skill and experience, specialist designers are better suited than laymen to design new styles since others simply cannot ensure that the visual balance between the document elements is not perturbed. Provided the changes made are minor and consistent, the uninitiated can dabble by modifying existing styles.

Figure 1.1: Examples of stylistic congruity and stylistic incongruity.

duced by a single author, it commonly arises as the product of collaborative writing activity. The stylistic incongruities occur especially between text segments written by different authors — for instance, in the paragraphs of Example 2. As we will see in section 2.2, achieving stylistic congruity is the most troublesome goal for collaborative writers to achieve. Collaborative writing, a set of practices rather than a specific activity, has the frequent feature that the planned text is partitioned into segments, with group members or sub-groups each writing a segment. In these cases, it is possible that, while each segment is an exemplary piece of writing, they do not cohere into a document that speaks with a single voice (Glover and Hirst, 1995). More likely, each segment is not an exemplary piece of writing and also contains stylistic incongruities. This feature of collaborative writing is not the only one that provides occasion for stylistic incongruities to arise. The other sub-activities of collaborative writing — the planning, transcribing, reviewing and revising — also shape the final stylistic form of a text, and therefore, play a role in achieving stylistic congruity.

1.2 Objective of this Thesis

The ultimate objective motivating this thesis is to help writers produce stylistically congruous texts, by means of a computational tool or suite of tools. While collaborative writers are the main focus, singular writers are considered as well. Since the eventual, envisioned solution to the problem is software, we use a software design methodology in place of a more general investigation or analysis.

Using an iterative software design strategy gives this thesis an orientation to focus on the underlying issues relevant to the goal of producing software. This goal is preferable to the premature development of a particular computational technique for analyzing stylistic incongruity. The components of the best software solution are tailored to the needs of the users. It is these needs, rather than development of computational techniques, that should drive the software design. There are many interesting algorithmic and computational approaches that can be used to gather information about the style of a given segment of text, but not all of them are useful to writers during the text composition process. The most basic building block of good software design is a good abstraction of the problem, a separation of *what* the software should do from *how* the software should accomplish this. The *what* corresponds to an analysis of the requirements of the problem. If the software meets all the requirements set forth by the analysis, but does not turn out to be a good solution, then this is the fault of the analysis rather than the software design. The *how* corresponds to the design and implementation plan for software that can achieve the requirements. The design is produced to satisfy the software requirements and might be straightforward if the requirements are stated clearly enough.

It is not within the scope of this thesis to design and fully implement a full-scale solution; this is a job for a software design team. Rather, the contribution of this thesis is the development of an abstraction of the problem and the development of a set of software requirements, derived through analysis of both the underlying problem and the work habits of the target user. In general, the design process consists of iterations of software requirements and design refinements; the work presented in this thesis constitutes the first iteration. This means that, although the requirements are not stated algorithmically, they form a basis for future software development. Additionally, the preliminary design of an essential software component, stylistic analysis, is given as well.

Substantial amounts of work from related research fields, as listed in Figure 1.2, were used in the requirement analysis. These related fields cannot be summarized in their entirety due to space constraints[1]; rather, the relevant work has been identified and included.

---

[1] Overviews found in the following theses (each provides coverage of the relevant research areas to various degrees): (Green, 1992) and (Mah, 1991) for stylistics, (Mawby, 1991), (Mitchell, 1996) and (Posner, 1991)

Writing: Models of the writing process, the sub-activities of composition, how the style of a text is created.

Reading: What makes a text incongruous, agreement among readers' interpretations.

Software Applications for writers: The design of current style checkers.

Computer-supported collaborative work (CSCW): User requirements; models of group work, negotiation, and communication.

Collaborative writing (CW): Descriptions and models of collaborative writing.

Software applications for CSCW: Existing tools and environments, integration into current collaborative writing environments.

Rhetoric: How one can write to accomplish a particular goal.

Stylistic prescriptivism: How one *should* write.

Communication theory: Models of written communication, how style conveys information, theories of style.

Computational linguistics: Computational models of style, computational stylistic assessment.

Literary stylistic analysis: Stylostatistics, analysis of style in literature, authorship attribution, detection of plagiarism, non-computational descriptions of literary style.

Figure 1.2: Research areas relevant to the objective of the thesis.

1.3 Organization of Thesis

Figure 1.3 provides an overview of the content of this thesis. The first three chapters analyze requirements of the software. In chapter 2, a taxonomy of support strategies is developed. In chapter 3, the requirements for making stylistic assessments are analyzed. Chapter 4 contains a description and discussion of an exploratory study designed to assess the role of subjectivity in stylistic assessment, an issue with broad consequences for computational stylistic assessment.

Several global issues were considered throughout the analysis:

Feasible Implementation The computational requirements of the tool, in terms of time and space, should not prevent its implementation on commonly available software and hardware platforms. Even a partial solution would be useful. Additionally, implementing

for collaborative writing, and (Glover, 1996) for both.

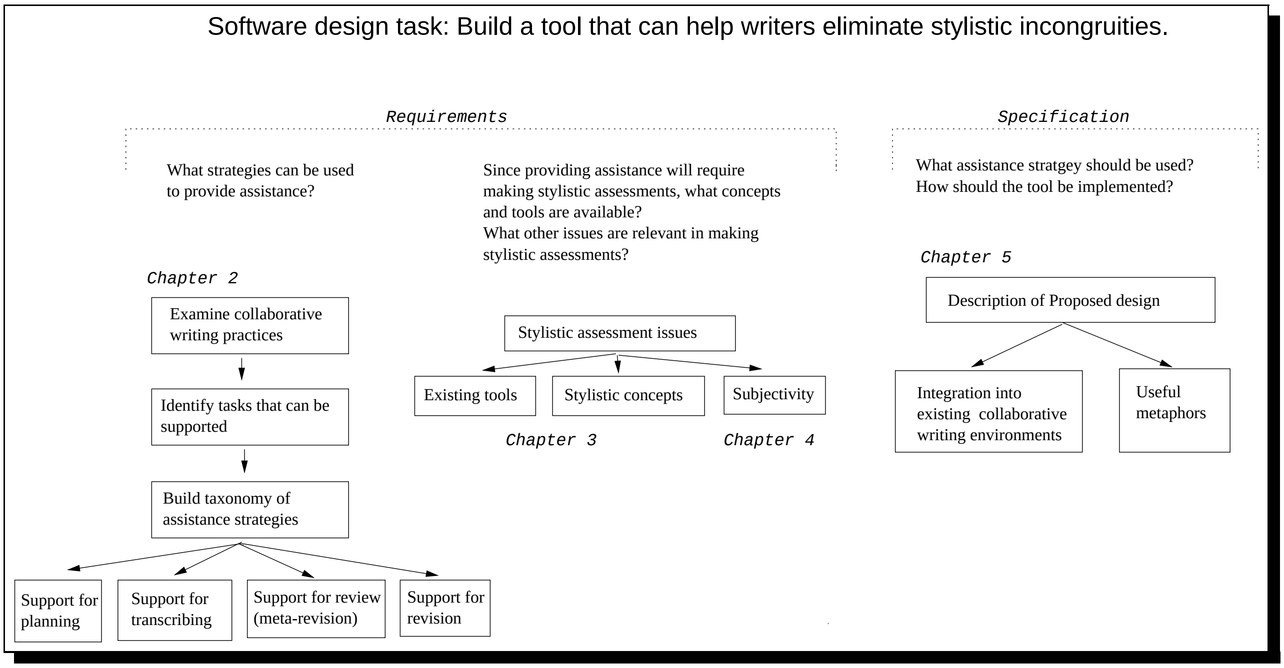


Figure 1.3: Organization of the thesis.

the functionality of the software should not require major research advances (e.g., requiring automated reasoning based on the metaphors in a text to infer the author's intended imagery).

Effective Assistance Strategy Although an intuitive form of assistance would be to locate and help repair stylistic incongruities that already exist in a text, the most effective strategy for providing assistance isn't known. Perhaps it would be more effective to provide assistance during another stage of the writing process, such as planning. It also isn't known how the tool should present the assistance to the user. Perhaps users do not want assistance in the form of a glorified spelling checker. Maybe they want a 'virtual' group member (computational process) who emails writing comments when the need arises. The tool should be useful to the users and not be a hindrance, but at the same time, deliver assistance using an effective strategy. The strategy must be flexible, since the users may be using one of several possible forms of collaborative writing.

Identify Relevant Concepts A good software design practice is to use concepts familiar in the application domain in place of implementation details when analyzing the software requirements. A vocabulary for computational stylistic assessment must be established. For the design, the tool interface should be based on an easily understood metaphor. These concepts must be identified.

Text vs. Document The object of the analysis is a text rather than a document. For our purposes, a document is a text that has been formatted and augmented with other carriers of information, such as typography, lists, tabular information, and figures. Although inconsistency in formatting is an important problem, it is excluded from the scope of the proposed tool.

Integration of Tool Writers do not need yet another piece of software. Rather, they need an additional facility within the environment that they currently use. Therefore, any proposed tool should be designed as an augmentation of the collaborative writing environment. While the area of developing specialized software applications as collaborative writing environments is growing, the use of these systems is not widespread. So while it is important to determine how to augment the facilities of future collaborative writing environments, the current forms of normative groupware must also be augmented.

Types of Text The proposed tool is intended to target the stylistic incongruities occurring in expository texts in the domain of business, scientific, and academic writing.

The Target User The users of this tool will not necessarily be writers. Stylistic tools that provide assistance during text revision may be used by editors or text planners, for example. Throughout this document, the terms 'user' and 'writer' are used almost interchangeably. However, the definition of 'writer' includes all those who are involved in the composition of the text and not necessarily the actual transcribers of the text.

1.4 Chapter Summaries

1.4.1 Chapter 2: Assistance Strategies

The objective of chapter 2 is to identify strategies for eliminating stylistic incongruity. First, the problem of stylistic incongruities is discussed with respect to collaborative writing practices. Supporting revision activity, an existing strategy for eliminating stylistic incongruities, is discussed next. Two obstacles in implementing this support strategy are identified; the type of assistance given does not reflect the type of help that reviewers need; and substantial research advances are still required.

The tasks required for composition are identified and three characterizations of the set of collaborative writing practices are synthesized. Collaborative writing practices are analyzed to determine what activities can be supported in order to either eliminate stylistic incongruities or prevent stylistic incongruities from appearing.

These results are summarized in a taxonomy of support strategies.

1.4.2 Chapter 3: Making Stylistic Assessments

To implement support strategies for transcribing, reviewing, and revising, a method of assessing the style of text is required. In this chapter, existing approaches are reviewed and their shortcomings are discussed. The meaning of a text's style is characterized and current definitions of style are discussed for comparison. This characterization stresses the reader's role in the assessment of style. A model for stylistic assessment, based on the construct/indicator framework, is introduced. Existing applications that make some kind of stylistic assessment are discussed. A body of past research in computational linguistics is synthesized to produce a corpus of stylistic constructs and corresponding indicators. For each construct, the construct description, the construct indicators, and the validity of the indicator is discussed.

1.4.3 Chapter 4: Audience Agreement on Stylistic Assessment

Many characterizations of style do not acknowledge the variability that may occur between readers' assessments of a text's style. But the reader plays a large role in the perception of style, which is both subjective and qualitative. In this chapter, the design, results, and conclusions of an exploratory study are described. The study was designed to reveal the degree of subjectivity in subject's stylistic assessments of a set of writing samples. Interestingly, the conclusion is that while subjects display a significant degree of agreement, there appear to be distinct patterns of stylistic assessment. For instance, the stylistic assessments by one group of subjects had a strong positive correlation with the authorship of the writing samples, while those of another group had a strong negative correlation. This study indicates that there is reader subjectivity and additional research is required in the future.

Chapter 2

Strategies for Assistance

2.1 Overview

The objective of this chapter is to identify the areas of difficulty in achieving stylistic congruity that collaborative writers experience. Preliminary work is examined that sets out an initial approach to providing assistance: assistance should be given in two stages while the text is being revised, first to help the authors detect inconsistencies and then to help them diagnose and repair those that are incongruous. Some problems with this strategy are discussed.

In order to discover other means of providing assistance (possibly targeting other aspects of the composition process), collaborative writing practices are examined in order to discover how stylistic incongruities are created in the collaboratively written document and a taxonomy is developed to summarize these areas of difficulty.

2.2 Assistance with What?

The first investigatory studies into collaborative writing both confirmed some intuitive ideas and uncovered some interesting findings. Studies by both Ede and Lunsford (1990) and Rimmershaw (1992) showed that it is uncommon for a single author to work in isolation and that a large number of professionals engage in some form of collaborative writing activity. While the study by Ede and Lunsford, although more detailed and comprehensive than that by Rimmershaw, can be criticized as only exploratory (the questionnaire on which it was based was not validated, nor was a pilot study conducted), it does provide useful information in the following two areas of interest for this thesis:

- characterizing the set of practices that are recognized as collaborative writing; and
- identifying difficulties that collaborators encounter.

In this study, it was discovered that nearly half of a professional's time is spent in some writing-related activity and 98% of those professionals surveyed regarded writing as important or very important to the successful execution of their jobs. An overwhelming majority of these professionals (87%) engage in some form of collaborative writing.

While the study found that collaborative group members generally find their collaborative writing efforts rewarding, it also uncovered a number of difficulties. The disadvantage of collaborative writing most often cited was the "tough task of making a common single style from numerous styles" (p. 60, (Ede and Lunsford, 1990)). But what is meant by a *common single style*? Ede and Lunsford have identified this difficulty as "achieving stylistic consistency" (p. 61), which is a commonly used term (e.g., (Glover and Hirst, 1995), (Mawby, 1991)). There are two kinds of stylistic inconsistencies, as Glover and Hirst (1995) first pointed out, and I believe that a definitional clarification is necessary to prevent confusion between the two. One type of stylistic inconsistency detracts from the quality of a text (i.e., bad inconsistencies), while the other kind provides the variation and texture in prose that make the writing interesting to read (i.e., benign inconsistencies).

To clarify the difference between the two types of inconsistency, it is best to distinguish *consistency* and *congruity*. We understand that *consistency* implies a similarity among the parts of a text, while *congruity* implies a similarity among the *character* of the parts, in a sense that this thesis explicates. So stylistic consistency and stylistic congruity are both properties of a text, at any level of granularity. A text is stylistically inconsistent if the parts of the text aren't similar, and a text is stylistically incongruous if the characters of the parts aren't similar. A particular text part cannot be stylistically incongruous on its own; rather, this is a property of a set of text elements. Therefore a text can be *stylistically congruous* even if is stylistically inconsistent, provided the inconsistencies are benign. In this case, the text parts, although different in style, are of the same character. For example, consider a text in which it is necessary to refer repeatedly to the same object. The referring expression could be inconsistent throughout the text (it could vary between pronoun and noun phrase, or between noun phrases using different synonyms), but the text can still be stylistically congruous. A *stylistically incongruous* text is one that contains a set of text elements that are stylistically inconsistent in a way that reduces the quality of the text — for example, the use of both very colloquial adjectives as well as technical adjectives to modify the same noun in a technical document. The stylistic effect of these elements is different and they do not have the same character. With these terms defined, the greatest difficulty in collaborative writing identified by Ede and Lunsford can be restated as *achieving stylistic congruity* or *avoiding stylistic incongruities* in the text being written. The underlying belief of this thesis is that a computational tool can help achieve these goals.

On the basis of this belief, we seek to develop some scheme to help writers produce stylistically congruous texts. The remainder of this chapter will begin to construct such a strategy or set of strategies. To do this, we first examine the assistance strategies of existing style checkers. Since the effectiveness of these strategies is relatively limited, we look to other ways to develop strategies for assistance. The approach used is to analyze collaborative writing practices in order to identify the aspects of composition with which writers experience difficulty. To do this, we consider research on both singular and collaborative writing. Although the composition process of singular writers has been modeled, no equivalent model for collaborative writers exists; therefore, an approximation is made by synthesizing several different viewpoints of collaborative writing. The results of this analysis are then presented in a taxonomy in the last section of this chapter.

In later chapters, these results will be used in order to propose some preliminary design directions for the creation of a computational tool. For now, the following observations are given, as they played an important role in shaping the type of analysis used in this chapter:

- The crucial precursor of the design of this tool is an understanding of the writer's needs. Therefore, an understanding of the difficulties writers experience throughout the composition process is of more value than the understanding, even if detailed, of a single problem specific to one subtask of the composition process. A broad understanding allows selectivity in the design of the tool. Then, determining which needs of the writer to be addressed can be part of the design decision. To continue in the direction of the current research (for providing assistance in revision activity) would be to assume the targeting of a specific need of the writer as an *a priori* design decision. This may be fatal to the user's acceptance of the eventual tool, as software acceptance depends upon its support of the natural activities of the users rather than forcing unfamiliar routines (Jones, 1995).

- A computational application to combat stylistic incongruity cannot be autonomous. The form of the application will be a tool to help writers tackle areas of difficulty and there are many ways to do so. No assumptions should be made before the analysis, such as that the support should be following the paradigm of a spelling checker.

- Just as the members of a collaborative writing group need not be all writers, the potential users of a computational tool need not be the actual group members who transcribe the text. Therefore, areas of difficulty in planning and reviewing are also potential targets for computational support.

2.3 Developing Strategies for Providing Assistance

2.3.1 An Initial Strategy

In analyzing their failure to achieve stylistically congruous text, the collaborative writers who participated in Ede and Lunsford's study identified the cause of their problem as one of 'melding individual styles.' The causes of this recurrent problem were identified variously as a group member with "their own writing style which they are not willing to give up" or a set of group members with their own "distinct and well-developed individual styles" (pp. 60–61). One solution, identified by Ede and Lunsford (albeit implicitly), is to "negotiate a common style." However, this solution is not easily implemented. During the initial stages of the composition process, writers have enough difficulty meeting the constraints of the writing task without imposing an additional constraint about adherence to a particular writing style. To cope with the constraints, writers draw upon established routines as a strategy for writing (Hayes and Flower, 1980a), so requiring writers to conform to an unfamiliar style may not even be possible at all. This leaves the negotiation to take place during the writing or after the writing has taken place. What actually does happen, implicit in a respondent's comment — that "the editing process was made tedious" — is the more common solution to the problem of stylistic incongruity: revision, revision, and more revision. A common style is hammered out when the text is being revised. This strategy is not reserved only for collaborative writers, since singular writers also revise their texts heavily.

It is this pattern of activity that is supported by the approach assumed by Glover and Hirst (1995), (1996). To achieve the "goal of helping collaborating writers achieve consistency of style," an approach was used that I classify as *supporting revision activity*. This was viewed as a task consisting of two components. The first would be to discover stylistic inconsistencies and to discern between the bad and the benign. This corresponds to distinguishing between the inconsistencies that are incongruities and those that are not. The second component would present these findings in a manner that would enable the authors to correct the inconsistencies, by providing suggestions that are both comprehensible and useful. Comprehensible means that the information is understandable by the users and usable means that the information is not so abstract so that it cannot be applied towards a solution.

In order to maximize the effectiveness of this strategy, it would be desirable to provide assistance with repair, as well as providing suggestions. This implies a support strategy with the following facets:

- Helping writers detect stylistic incongruities by first detecting stylistic inconsistencies and then determining which ones are stylistically incongruous.

- Helping writers understand the stylistic incongruities by presenting diagnostic information and by providing comprehensible and usable suggestions.

- Helping writers fix the problems by providing assistance with repair and by helping the user to make the required change.

So this strategy assumes that writers have difficulty identifying stylistic incongruities, and for the stylistic incongruities that they do identify, they have trouble with diagnosis and repair. This assumption is true, but not quite complete enough. Researchers studying writers who must revise documents have identified three separate but interdependent abilities: detection, diagnosis, and repair (Schriver, 1992). Therefore, this support strategy decomposes the larger task of revision similar to the way that writers do. But there are two important problems with this strategy. First, writers require more assistance as they progress through the subtasks of revision than this strategy provides. For example, a writer being able to detect a stylistic problem on their own doesn't mean that they won't need assistance with later diagnosis or repair. Second, several large research advances are required in order to implement this type of support strategy. These are discussed in the following two subsections.

2.3.2 Cumulative Assistance is Required

Researchers studying the revision activity of single writers note that the problems encountered are cumulative (Schriver, 1992), (Kelly and Raleigh, 1990). This means that writers may be able to detect problems, but not be able to diagnose them; be able to detect and to diagnose problems, but not be able to repair them; or, in the worst case, not be able to detect the problems at all, much less diagnose and repair. This pattern suggests that the writer's need for assistance is cumulative; the farther the progression through the revision steps, the more difficulty authors have. Among collaborative writers, these problems with the revision process can only be intensified, since negotiation and communication might be required among the collaborators as they move between the subtasks of detection, diagnosis, and repair.

Computational strategies, such as the ones spelling checkers use and Glover and Hirst propose, are similar to the strategies used by humans, but they are different in important ways. Both the computational strategies and the human writer strategies decompose their problems similarly, into the detection, diagnosis, and repair subtasks. However, computational strategies are *forward dependent*, which means that the ability to detect is dependent on the ability to diagnose. For these computational strategies, the ability to perform detection is entirely dependent on an operational definition of the different types of errors, which subsumes diagnosis. It is by using the definition of the problem that the computational strategy is able to detect the problem. The definition of a spelling error to a spelling checker is embedded with information that is used to detect as well as diagnose (e.g., the message "word not found" indicates that a spelling error is defined as a word that has unsuccessful lookup; or the suggested correction of "until" for "untli" indicates that checking for letter transposition is part of the diagnosis). The computational definitions used for more-subtle spelling errors, such as malapropisms, are also embedded with diagnostic information (see the computational definition of Hirst and St-Onge (1995)). Human performance on these subtasks, on the other hand, is not forward dependent. As discussed above, many writers can detect problems in their text, whether spelling or stylistic, but are not able to say why.

Since the Glover and Hirst strategy also requires the same type of operational definition for stylistic incongruity, the ability to detect is forward dependent. The sequence of steps used to detect stylistic incongruity will implicitly include a diagnosis. For instance, a sentence might be assessed as stylistically incongruous not only because it is inconsistent, but the inconsistency is deleterious for a particular reason.

For this reason, the strategy will not be able to support the cumulative needs of the writers. This approach will not be able to diagnose problems that it cannot identify. In other words, it will not be able to provide diagnostic information even if a user specifically requests it (if the problem is not part of the tool's operational definition of stylistic incongruity). In an ideal scenario, however, a writer could use a computational tool at the point where assistance is required. For instance, this point could could be after a problem has been found, but has not yet been diagnosed.

Since detection requires the least amount of support, relative to diagnosis and repair, the drawback of this support strategy is serious. This is not to suggest that assistance with detection would not still be very useful, but the tool should be flexible enough to provide, on demand, the other kinds of assistance too, such as diagnosis or repair. A spelling checking strategy means that useful diagnosis and repair assistance comes only with the ability to detect every problem that a human reviser could conceivably detect as well.

Furthermore, a spelling checker strategy can only provide repair assistance with problems that are diagnosable. For example, a spelling checker provides repair assistance by allowing the user to quickly replace the misspelled word with a selection from a list of candidates created using the diagnostic information. It also provides assistance by ensuring that all instances of an error are corrected consistently. Users may desire the same capabilities from a stylistic repair tool. A user may desire repair assistance with a segment of text, even if the tool hasn't detected or diagnosed any problems. This repair assistance also should ensure that a writer's particular repair is consistent with the other repairs and is not introducing any new stylistic incongruities in the text. Obviously, determining whether stylistic incongruities are similar in nature is more difficult than identifying similar spelling errors.

Overall, the assistance provided by the spelling checker strategy is inadequate for all the steps of the revision process. A tool using this strategy can only attempt to diagnose what it detects, and can only attempt to help repair what it is able to diagnose. For a specific problem, it is more likely that a user will require assistance with repair and diagnosis, but for any given text, writers need more assistance with diagnosis and repair than detection. Users may need help diagnosing problems that they can detect, or help repairing problems that they can diagnose. The strengths of the assistance strategy are mismatched with the needs of the writers. This mismatch, illustrated in Figure 2.1, occurs because writers need the most assistance with repair and diagnosis of arbitrary problems, while the computational strategy only does so for particular problems. An better strategy would concentrate the assistance in the areas where the users require it — in the later subtasks of revision.

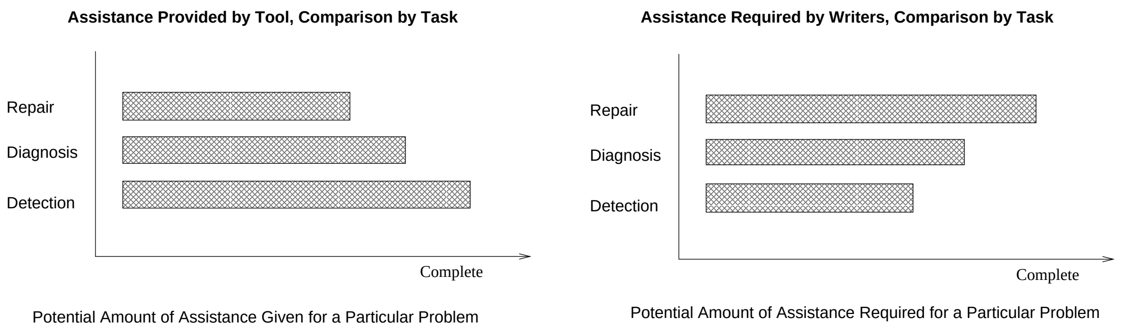


Figure 2.1: A comparison of assistance given by the proposed strategy to assistance required by writers performing revision activity.

2.3.3 Substantial Research Advances are Required

In order to implement this type of strategy to detect and repair stylistic incongruities during text revision, Glover and Hirst (1995) identify a number of advances that would first be required. These are listed in Figure 2.2.

Advance Required	Revision Subtask Targeted
• Ability to know what kinds of things do and don't count as undesirable inconsistences	Automatic **Detection** of stylistic incongruities
• Ability to detect these things computationally	
• Ability to articulate stylistic problems in terms that the user can understand	Generation of **Diagnostic** information to user
• Ability to suggest to user, again in simple terms, how stylistic problems can be corrected	Generation of **Repair** advice for user

Figure 2.2: Advances required for Glover and Hirst's support strategy.

These research advances would be substantial. There are many obstacles even to achieve the first research advance. First, stylistic inconsistencies cannot be reliably detected, let alone stylistic incongruities. An experiment conducted by Glover (1996) used the assumption that stylistic inconsistencies correspond to authorship boundaries and then explored the value of stylostatistical tests in determining stylistic inconsistencies. The experiment was designed to create a large set of pseudo-collaborative texts. In the experimental tasks, subjects wrote two halves of a document (Part 1 and Part 2), corresponding to the task of summarizing the halves of a television show, shown on two separate occasions. A set of pseudo-collaborative texts was produced by merging every subject's Part 1 with all the other Part 2's (including the one authored by the Part 1 author). The stylostatistical tests — statistical counts of quantifiable features of the text — had moderate success as determinants of which pseudo-collaborative texts were authored by the same person. This work was preliminary and several questions still need to be addressed:

- Can these tests be used to detect inconsistencies that occur within a segment written by a single author (since a collaboratively written text is often composed of individually written segments)? Surely stylistic incongruities occur within a segment written by a single author and not just between segments authored by different writers.

- Even if all the stylistic inconsistencies can be reliably detected, it still must be determined which inconsistencies are incongruities. How can this be done?

- Does authorship correspond to stylistic inconsistency? The assumption used by Glover and Hirst (1995) was that an author's style is more like her own than like anyone else's. One of the results of an exploratory study conducted for this thesis is that readers are poor judges of authorship (Chapter 4). Perhaps the assumption about authorship isn't useful for the problem of stylistic incongruity.

- A style tool ideally would be an integral part of a collaborative writing environment. The collaborative writing environment maintains (or could be made to maintain) editing histories of the various components of the text. Therefore, the authorship of the segments of the text would already be known. Wouldn't it be better to take advantage of this information?

2.3.4 Other Strategies for Eliminating Stylistic Incongruities

Supporting revision activity in this way is so intuitive as an approach to helping writers eliminate stylistic incongruities that it is easy to fail to see that it is but one strategy of several possible. In fact, due to the problems discussed in the last two sections, it is unlikely to be the best strategy.

There are other approaches that could help collaborative writers produce stylistically congruous text. One strategy could be to support better the crucial planning stage of composition, since planning is linked to style and therefore stylistic congruity. This approach may help writers to reduce or to avoid costly repairs and rewriting by using a preventive approach to eliminate stylistic incongruities before they occur.

Additionally, stylistic incongruities could be detected while they are being created. An interventive support strategy would alert the writer when the text begins to show incongruities. This approach has the advantage of eliminating stylistic incongruities before they become deeply ingrained throughout large segments of text by successive modifications. This approach may help writers correct stylistic incongruities during the actual transcribing stage of the writing process.

To meet the goal of identifying the areas of difficulty in achieving stylistic congruity for collaborative writers so that an effective support strategy can be devised, collaborative writing practices are examined. It would be convenient if a model existed that unified all the different practices, but this is not the case. The approach proposed by Posner (1991) was to taxonomize the different facets of collaborative writing activity, but her taxonomy is not comprehensive enough and is more like an enumeration of the different collaborative writing strategies than an abstraction of the nature of collaborative writing. Other researchers, such as Sharples et al. (1993) and Ede and Lunsford (1990), enumerate different patterns of collaborative writing. The analysis to be presented in section 2.4 is based on this research. To draw together these different views, instead of simply choosing one of them, I asked the following questions for each:

- Of what subtasks does the composition process consist? How does the transition between the subtasks take place?
- If the set of subtasks is considered to be the workload of the collaborative group, how is the workload distributed among the members?

- How do the subtasks relate to the eventual stylistic congruity or incongruity of the text?

In answering these questions, I found commonalities in the three views of collaborative writing. In section 2.4, I synthesize the existing work by constructing a 2^3 space of possible forms for collaborative writing practices. In sections 2.5, 2.6, and 2.7, I examine the subprocesses of the collaborative composition process to uncover areas of difficulty in achieving stylistic congruity.

2.4 Analysis of Collaborative Writing Practices

The goal of this section is to establish a useful way of looking at the collection of practices which make up collaborative writing activity. The way we need to look at collaborative writing activity may be different from the viewpoint required for other types of research. The picture of collaborative writing activity we need is process oriented; we want to know how stylistic incongruity arises during the process of creating a text collaboratively. A complete model of collaborative writing activity is still an ongoing research goal, but I believe that sufficient components of research exist with which to construct a useful picture.

One such component is derived from established research on the composition process of singular authors. Singular writing can be considered a special case of collaborative writing. A singular writer is a collaborative group of one, with most of the communication and negotiation obstacles removed (e.g., a dialogue among collaborators now becomes an internal dialogue and tasks that can be accomplished in parallel by a group must be done in sequence by a singular author).[1]

Models of singular writing have been developed for at least a decade and are more established than models of collaborative writing. But from this research, we can know the basic subtasks of composition that also face a group of collaborative writers. This set of subtasks is given by the hybrid model of composition, shown in Figure 2.3. This model captures the salient features of established cognitive-processing models[2] and reflects the

[1] A common view is to consider collaborative writing as a souped-up version of singular writing, as an activity with all the same components as singular writing, but made more difficult by the necessity of communicating and cooperating with other people. I believe a better view is to consider singular writing as a special case of collaborative writing, as an activity with many of the dimensions that collaborative writing requires collapsed down to a trivial case.

[2] As opposed to other theoretical viewpoints, such as psycholinguistic, linguistic, and developmental

findings of Nold (1981) and Flower and Hayes (1981) (see also Hayes and Flower, (1980a; 1980b)). This model is widely accepted[3], but, as with any model, is not without critics (e.g., see (Hartley, 1991)).

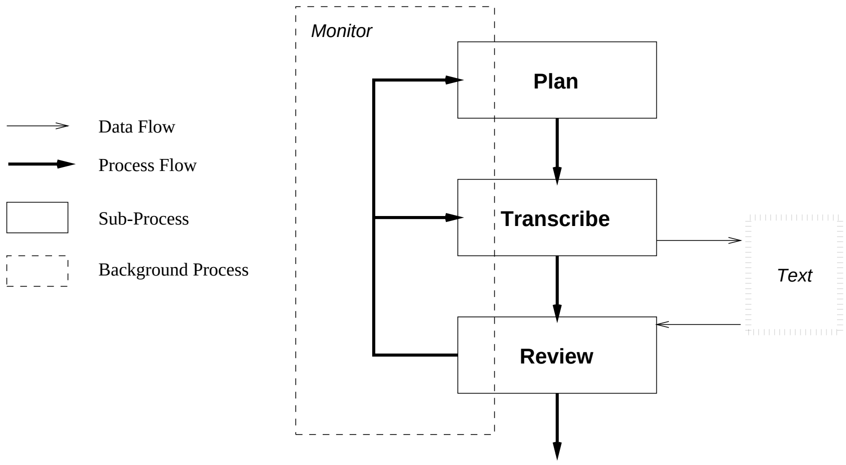


Figure 2.3: A hybrid model of the cognitive processing model of composition.

The underlying hypothesis of this model is that writers must strategize in order to achieve a solution to a communication problem. The model then elaborates on the nature of this strategy which is basically to plan, to transcribe and to review the text in order to see whether it meets the goal as conceptualized in the plan. This goal-directed approach to writing (Hayes and Flower, 1980b) is consistent with the research premise of computational stylistics, which holds that a writer uses style with particular goals in mind (e.g., DiMarco (1993) and Hovy (1988)). This is worth noting, since we assume that the strategies that writers use to achieve their communicative goals also include strategies for style.

Broadly speaking, the solution, the realization of the communicative goals, is achieved by *planning* a solution, *carrying out* the plan, and *reviewing* the results to judge whether they meet the criteria for a good solution (Nold, 1981). These steps correspond to the sub-processes of *planning*, *transcribing*, and *reviewing*, but these sub-processes do not necessarily correspond to subtasks which are carried out in sequence. Rather, they are carried out both iteratively and recursively. For example, to repair a problem discovered while reviewing a

models of composition.

[3]Certainly, the process used to produce the text depends on the nature of the text (Hartley, 1991). Therefore, a style tool based on this model may not be applicable to poetic or expressive writing.

text, the writer may need to plan, transcribe, and review a modification.

The controlling and guiding force for these subprocesses is the *monitor*. The function of the *monitor* is to control the sequence of the writing subprocesses (Hayes and Flower, 1980b). It acts as a background process, monitoring the status of the writing process and triggering a subprocess switch when required. This abstraction is a metaphor that works well for singular writers, since the monitor in singular writing has access to the mental representations of all the sub-processes, whether active or not, and presumably corresponds to some kind of cognitive process. But the notion of a monitor corresponding to an internal cognitive process doesn't scale up very well to collaborative writing. Since the tasks are being performed by different group members, each person has their own private view of the text and none has the ability to trigger a change in subprocesses in the other group members. How can a monitoring process exist to guide the group to switch to a review sub-process? A group member can switch subprocesses locally, but globally the subprocesses of the collaborative composition are guided by the collaborative writing strategy. For example, a singular writer may review their writing at any time during transcription and make modifications of any part, but this is not as easily done for collaborative writers.

Several researchers have observed and enumerated the collaborative writing strategies that groups use. Instead of investigating how each of these strategies affects the eventual stylistic congruity of the final text, I look for areas of difficulty common to each. These areas of difficulty may occur more severely or frequently in some particular patterns and less in others, but our goal is to characterize the general difficulties in order to design a tool with broad coverage. The different writing strategies can be categorized by the way in which the tasks associated with the composition process are distributed. To complete each sub-process, planning, transcribing, and reviewing, the collaborative writing group can use one of the following two strategies:

- The whole group performs the subtask. This includes work in the form of discussion and negotiation. All group members have access to the communication taking place (even though one member of the group may be appointed to record the results of the discussion, e.g., the *scribe* (Posner, 1991)). (Case **G** in figure 2.4 below)

- An individual or small sub-group performs the subtask. A set of individuals or small sub-groups all perform the subtask, but not together. This case also applies if there is only one text segment (e.g., each member writes a text segment, or each reviews a text segment). (Case **I** in figure 2.4 below)

Consider now the 8 cases (2^3) where each of the three subtasks, planning, transcribing, and reviewing, are completed using one or the other these strategies. All the patterns of collaborative writing identified by Ede and Lunsford (1990), Posner (1991) and Sharples et al. (1993) can be placed in this categorization, as shown in Figure 2.4. Each characterization has been assigned a key, corresponding to the legend at the bottom of the figure.

Planning	**Transcribing**	**Reviewing**	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
G	G	G		Sc, Jnt	Rec
G	G	I	5	Sc	Rec
G	I	G	1, 2	Sc, SiW, SeW	P
G	I	I		Sc, SeW	P, S
I	G	G			
I	G	I			
I	I	G	3, 4	SeW	P
I	I	I	6, 7	SeW	P, S

Legend

$\mathcal{A}$ = (Ede and Lunsford, 1990):	1, 2, 3, 4, 5, 6, and 7 correspond to their 'seven organizational patterns' (pp. 63–64).
$\mathcal{B}$ = (Posner, 1991):	Sc = Scribe, SeW = Separate Writers Strategy, SiW = Single Writer Strategy, Jnt = Joint Writing Strategy (pp. 51–55).
$\mathcal{C}$ = (Sharples et al., 1993):	Rec = Reciprocal Strategy, S = Sequential Strategy, P = Parallel Strategy (pp. 14–16).

Figure 2.4: A Taxonomy of Collaborative Writing Strategies.

Ede and Lunsford

Ede and Lunsford (1990) described a set of seven organizational patterns:

1. Group plans and outlines. Each member drafts a part. Group compiles the parts and revises the whole.

2. Group plans and outlines. One member writes the entire draft. Group revises.

3. One member plans and writes draft. Group revises.

4. One person plans and writes draft. This draft is submitted to one or more persons who revise it without consulting the writer of the first draft.

5. Group plans and writes draft. The draft is submitted to one or more persons who revise it without consulting the writers of the first draft.

6. One member assigns writing tasks. Each member carries out individual tasks. One member compiles the parts and revises the whole.

7. One person dictates. Another person transcribes and revises.

Posner

Posner (1991) describes four writing strategies for text creation, which is one of four facets of a taxonomy of collaborative writing. The other facets are document control methods, activities, and roles. The strategies are as follows:

- **Single Writer:** One person is responsible for the writing of the entire text. The group members participate in the planning activity by providing ideas and helping in the brainstorming process. The group members also participate in the review process by providing comments on the text.

- **Scribe:** One person records the discussion results from group activity. The type of group activity, according to Posner, is typically brainstorming and planning meetings. For the remaining tasks, another writing strategy is used. For this reason, the scribe strategy is more of a member role during a subtask than a strategy for collaborative writing.

- **Separate Writers:** Several group members are each responsible for a segment of the text. In order to assign these responsibilities, the planned text is conceptualized so that this segmentation can take place, and Posner foes not explain this. Although she stresses the importance of the integration stage, during which the text is revised, she does not specify who performs the integration or the revision. I assume that the planning and revision subtasks can be performed by either an individual or by the entire group.

- **Joint Writing:** In a collaborative group using this strategy, the entire group composes together. This includes the planning, transcribing, and reviewing subtasks.

Sharples, Goodlet, Beck, Wood, Easterbrook and Plowman

Sharples et al. (1993) describe three general strategies for collaborative writing.

- **Parallel Working:** The writing is divided into subtasks, such as writing segments of the text. Sub-groups or individuals complete the subtasks in parallel and with knowledge about the other segments. An integration phase follows this, which includes the revision subtask. It is not specified whether the planning or review subtasks are performed by an individual or by the group; I assume that it can be either.

- **Sequential Working:** The entire task is divided into a set of subtasks, which are completed in sequence (in stages) by individual group members. The result of each stage is passed to the individual responsible for the next subtask. The subtasks can be writing segments of the text or revising. It is not clear whether the whole group or an individual does the initial planning; I assume it could be either.

- **Reciprocal Working:** The group members work together during the planning and transcribing of the text. It is not specified explicitly how the reviewing is performed; again I assume it could also be done by the group or by an individual.

Summary

Although it is convenient to use specific aspects of these models for this purpose, it is important to realize that these models each have a different focus. The model of Ede and Lunsford is not rigorous; it simply characterizes the results from an exploratory study. Posner's model is based on a relatively small number of interviews and pays attention to the equality of the division of the collaborative tasks. The model of Sharples et al. is part of larger research that focuses on task, group, communication, and external representation issues.

In the next section, the sub-activities of planning, transcribing, and reviewing in collaborative writing groups are discussed. The description of these subtasks from Ede and Lunsford, Posner, and Sharples et al. are examined. The goal of the examination is to uncover aspects of the collaborative activity that may cause or help eliminate stylistic incongruities.

2.5 Planning

In Section 2.4, the composition process was described, following the research of Flower and Hayes, as a process through which a solution is found to a communication problem. In order to find this solution, writers employ a number of strategies. In this section, we expand on *planning*, an important part of the composition process, as it embodies the solution-finding strategy that writers use.

In general, people use the heuristic strategy of planning when faced with a complex and under-constrained problem. Using this strategy entails constructing a plan or outline of the desired solution, creating the solution, and then evaluating the solution against the plan. For a communication problem, using this strategy means constructing an outline of a text that represents a valid solution (planning), so that the text can be created by translating the plan into sentences (transcribing), and then can be evaluated against the plan (reviewing). But constructing an outline of a text that represents a valid solution to the communication problem is a difficult task, let alone the other subtasks of the composition process. The reason for this is that the communication problem is complex and under-constrained (Flower and Hayes, 1981) (Hayes and Flower, 1980a; Hayes and Flower, 1980b; Hayes and Flower, 1986).

This means that although there are constraints on what the solution text should look like, there are not so many that they can be used to rule out a great number of the many different possible outlines facing the writer during the planning stage. For any communication problem, there are many ways to write a text to convey what is desired. This factor, that a text should convey what is desired, is an example of a constraint on the solution and I describe it as the *primary constraint.* In general, authors are faced with a number of such restrictions. They must compose a text that communicates some particular information, that must follow a certain layout or that conforms to some size, and that must use a certain vocabulary. These constraints can both help and hinder an author and handling them is a 'juggling act', to use Flower and Hayes's metaphor. They can help to narrow the field of possible solutions, but they also create a cognitive burden for the writer on whom they are imposed.

We know from past research that planning helps singular writers in the following ways:

- It helps the writer to decompose the problem into manageable pieces.

- It helps the writer to operationalize the solution (to specify the solution in terms of the steps required to achieve it).
- It helps the writer to prioritize the subtasks.

For collaborative writers, planning helps in these ways as well. But collaborative writers have additional needs. A group of collaborative writers must make two types of plans: the plan for accomplishing the task as a group, and the plan for the communication that the text is intended to convey. Most collaborative writing research concentrates on the first type of plans that groups make — identifying the regular patterns of activity.

For the second type of plan, the text outline, most collaborative writing research either just acknowledges that it takes place or omits this important sub-process. For example, Ede and Lunsford acknowledge the planning activity in each of the seven patterns (in the last two patterns, it is implied that it is done by an individual). In Posner's taxonomy, the planning activity is omitted as a step in composition in the separate writers strategy. In the model of Sharples et al., the planning activity is also under specified in the parallel and sequential strategies, although communication issues, which are relevant to planning, are explored in detail. So even though planning is acknowledged as a part of the collaborative activity, its role is not given the attention that it merits. In a discussion of how to best support collaborative authoring, Newman and Newman (1992a) argue that writing "models that deal with collaboration are confined to the solution of a structured problem, generally using some form of central co-ordination" (p. 24). This is a serious criticism because structured communication problems are but one of many types facing collaborative writing groups. Also, making plans for solutions to structured problems is easier than unstructured problems, since the form is less flexible.

In the following sections, we describe several aspects of planning that have consequences for stylistic congruity. First, we describe why stylistic congruity cannot simply be defined as one more constraint on the solution to the communication problem facing the group, thus trivializing the problem (section 2.5.1). In the subsequent section, we describe why the text plan is important to stylistic congruity and why it is difficult to make an adequate text plan (section 2.5.2). In the final subsection, we discuss why it be essential that the plan is shared among the group members (section 2.5.3).

2.5.1 Additional Constraints Don't Ensure Consistency

A naive strategy for achieving stylistic congruity is to simply make stylistic congruity a constraint that is part of the communication problem. This way, the group could ensure that stylistic incongruity is avoided by simply planning properly in the first place. For example, writers must handle constraints on the length of the text, on the layout of the text, and on the vocabulary used. It certainly would be convenient if the collaborative writers could just be given an additional constraint of having to write in a certain style. Unfortunately, this strategy fails.

The failure is due to two shortcomings. First, it is not possible to specify a target writing style other than in coarsely grained terms, such as "write with a formal style" or "write in a conversational style." These goals are too broadly defined, since it is possible to achieve the target writing style, and yet to be stylistically incongruous. In addition to this, these terms are relativistic. These qualitative terms, such as *formal*, can have different meanings for each member of a group. An experiment conducted for this thesis demonstrates that stylistic assessment is extremely subjective and has great variability (Chapter 4).

The second shortcoming is due to the incompatibility of stylistic constraints and the way writers work. Writers draw on routine or well-learned procedures as a means of reducing the cognitive burden of juggling the multiple constraints imposed during the composition process (Hayes and Flower, 1986). Introducing another constraint would not be compatible with this strain-reducing technique. For these reasons, it is ill-advised to impose an *a priori* style. Rather, the strategies that are used naturally by writers in their genre must be supported.

There is one constraint that could be imposed on writers that is related to stylistic congruity. Mawby (1991) suggests that collaborative writing environments should have the means to ensure that consistent terminology is used throughout a document. Her work specifies three capabilities that the environment should offer a collaborative writer: to be able to identify and to view the same word or phrase used across a document; to be able to use a defined dictionary of agreed-upon terms in conjunction with a thesaurus and to receive suggestion for agreed-upon terms if others are typed instead; and to be able to update the dictionary as new terms are selected or existing terms are modified.

These functions are more relevant to the actual transcribing of the text rather than the planning of the text's content, but such a facility is relevant to the planning subprocess. When first planning a text, writers must decide if a vocabulary of special terms should be defined (e.g., when writing a technical document). Deciding to define a vocabulary may avoid usage inconsistencies, which are one type of stylistic incongruity. Defining such a vocabulary is another task for the group to perform, unless it already exists. Adherence to the usage presents another constraint that writers must satisfy, but can be supported during the transcription by computational facilities. In the planning subprocess, the decision must be made, introducing another constraint, namely the creation the dictionary of terms (either from scratch or by adapting an existing one).

2.5.2 Building a Concept of the Text's Intended Communication is Difficult

Earlier in section 2.5, planning was described as a strategy for developing a solution to a communication problem. In using this strategy, authors construct an outline or a plan of the text and then evaluate the solution against the plan. In this subsection, the nature of the text plan is discussed.

An important aspect of the text plan is that it is a conceptualization, as opposed to some tangible, physical entity. As a concept, it may or may not have some kind of external representation, such as a written outline or a schematic. Additionally, this concept has to be constructed. According to Flower and Hayes (1981), the planning done by singular authors consists of two tasks: generating content and organizing content. Generating content alone is not sufficient for creating a good text plan; rather, it must be organized with the intended reader in mind. For this reason, subject-matter experts can be notoriously ineffective writers; they have great amounts of knowledge, but cannot distinguish what is important to readers and organize the information for them. Therefore, the plan consists of more than what the text should convey; it also captures *how* it should be conveyed. The text is a communicative act and has pragmatic content.

Because the text plan encompasses so much, the term *plan* is not a very descriptive label for the product of the planning activity. Since the author determines not only content, but envisions how the content should be conveyed, the concept constructed during the planning process is effectively a conceptualization of the text's communication — or more accurately, the communication that the author intends to achieve through the text. The term *intended* *communication* is used as an elaboration of the term *plan*, but both refer to the concept that is constructed during the planning process.

The concept of the text's intended communication greatly affects the style of the text produced by the composition process. The style of the text is part of the author's communication and communication in the social world is structured by basic patterns (Newman and Newman, 1992b), (Nold, 1981), such as:

- How you talk to `<role>`,
- How you talk in `<place>`,
- How you manage your identity, what risks you will take in exposure to other `<of status>`, what your desired `<persona>` is.

Authors use stylistic means to convey these pragmatic factors that are part of the text's intended communication. Since the text's intended communication is achieved with the text's style (among other things), it also plays a role in stylistic incongruity. Inconsistent style can convey inconsistent pragmatic information, and these inconsistencies result in something 'going wrong' with the communication, to use Austin's term (1962) — stylistic incongruities are *infelicities*. An ineffective conceptualization of the text's intended communication foreshadows stylistic incongruity.

There are several aspects about a text's intended communication that make it difficult to construct, especially for collaborative writers. First, of the two tasks — generating content and organizing content — the latter is more difficult. In order to organize content for an intended reader, it is necessary to understand how to best achieve the communicative goals for a given communication problem. This includes clearly identifying the target reader abd tailoring the communication to them. But this is difficult, as the bevy of research on various text genres, such as persuasive, didactic, and expository writing. One strategy that collaborative writers use is to draw on an expert for these abilities (e.g., the *consulted* strategy in Posner's taxonomy), but this strategy is far from failsafe.

A second difficulty arises from the lack of any framework in which to discuss the pragmatic factors. Even among researchers, little agreement exists. Luckily, a completely verifiable and precise framework is not required, just a serviceable one. For example, the NLG application PAULINE has the ability to take pragmatic factors into account when generating text; and the following framework was used to categorize and represent pragmatic factors (Hovy, 1988):

- Conversational atmosphere: time, tone, conditions
- Speaker (interlocutor characteristics): knowledge of the topic, interest in the topic, opinions about the topic, emotional state
- Speaker-hearer relationship (interlocutor characteristics): depth of acquaintance, relative social status, emotion
- Speaker-hearer relationship (interpersonal goals): to affect hearer's emotion toward speaker, to affect relevant status, to affect interpersonal distance
- Hearer (interlocutor characteristics): interest in the topic, opinions about the topic, language ability, emotional state
- Hearer (interpersonal goals): to affect hearer's knowledge, to affect hearer's opinions of topic, to involve hearer in conversation, to affect hearer's emotional state, to affect hearer's goals

Similarly, this representation could be used by collaborative writing groups. Instead of serving as an input to a natural language generation application, it (or some modification of it) could serve as a framework for the collaborative writers' discussions.

These pragmatic aspects are a part of the solution, not part of the problem definition, as Flower and Hayes suggest. Flower and Hayes (1980) (p. 40) identify *rhetorical constraints* which must be satisfied.

> Whatever writers choose to say must ultimately conform to the structures posed by their *purpose* in writing, their sense of *audience*, and their *projected selves* or imagined roles. In essence, writing is also a speech act and therefore subject to all the constraints of any interpersonal performance.

But this constraint implies that pragmatic structures are posed separately from what a writer chooses to say, which is incorrect. The rhetorical constraints must be constructed as part of the solution to the communication problem; they cannot exist a priori to the solution, serving to guide the writer and helping narrow the field of possible solutions.

2.5.3 A Common Conceptualization of the Text's Intended Communication is Essential

The writer's conceptualization of the text's intended communication is important, as it guides the composition process. For singular writers, it is the concept held by one individual that guides the composition process, from start to end (which incidentally is also the responsibility of one individual). This is not the case for collaborative writing. Rather, a collection of concepts, held by various group members work in concert to guide collectively the composition process. Ideally, the same concept would be shared by all the group members. This way, the guidance would be consistent and the group would work constructively together towards a common goal, but this is not always the case. For example, a reviewer may make comments about a text draft that the transcriber can't or won't incorporate, arising from a misunderstanding of the purpose of the text. Of course, in a group, the members could discuss any unclear comments and possibly repair the misunderstanding, but this underlines the point that a lack of a common concept of the text's intended communication results in inconsistent actions within the group of collaborative writers (which then require communication to repair).

The members of a collaborative writing group can easily have different ideas about the text's intended communication. First, every writer has their own predisposition to a type of solution to a communication problem. In order to establish a common concept of the text's intended communication, the concept must be constructed and negotiated. This negotiation requires communication and communication requires a representation. There are potential problems with both of these steps. First, the construction of the text's intended communication is a difficult task, even for singular writers (see section 2.5.2). Collaborative writers must negotiate and be persuaded to buy into the concept and to give up their pre-existing beliefs. Second, writers lack the vocabulary and framework to discuss the pragmatic aspects of the text's intended communication. Furthermore, different types of communication problems place different demands on the collaborative writers with regard to the amount of negotiation required. The communication problems of technical documents or instruction materials are relatively structured (Newman and Newman, 1992b), whereas the problems of academic papers, which aim to reconceptualize or redefine knowledge, are more difficult. Newman and Newman (1992b) identify three modes of collaborative authorship, and these can be distinguished by the type of negotiation required among the collaborative writers in order to establish a common concept of the text's intended communication.[4] Each mode is described below:

1. **Literature:** The authors must negotiate a common conceptualization of the text's intended communication. This is an input to the text production process (which may be

[4]There are some differences in terminology. For example, Newman and Newman's term "definition of reality" basically denotes the concept of the text's intended communication.

singular, shared, etc). Since all the collaborators understand and agree, there is no need for renegotiation. Revisions are minor polishing, rearrangements or adjustments.

2. **Documentation:** A common conceptualization of the text's intended communication pre-exists and negotiation is not required.

3. **Critical Discourse:** The shared concept of the text's intended communication is renegotiated during the production of the text, resulting in major revision (e.g., reconceptualizations). This requires rich communication channels.

To support the negotiation required for these modes of collaboration, especially "Critical Discourse," rich communication channels are required. Ideally, the collaborative writers would be involved in face-to-face negotiation. Some collaborative writing patterns have very poor channels of communication for this important subprocess of composition. For example, in the sixth organization pattern identified by Ede and Lunsford, the writing tasks are simply assigned to and carried out by individuals in the group. There is no feedback mechanism, no way to ensure that the concept of the text's planned content is properly understood. It seems reasonable, however, that collaborative writers will choose a pattern of collaboration that supports the communication required by the negotiation entailed by their particular problem.

2.6 Transcribing

The transcription process, or the sentence generation process, involves the translation of the text plan into written prose. This process is guided by two types of knowledge (Nold, 1981):

- knowledge about language, and
- knowledge about the text's plan.

Both types of knowledge affect writing style. For example, an author's knowledge about language shapes the vocabulary and syntax used. Additionally, aspects of the intended communication, such as the author's desired persona, can be conveyed subtly and through the text's style. Therefore, knowledge about the text's plan also shapes writing style.

These two types of knowledge, taken together, guide the translation of the text's plan into sentences. A cohesive text, as we know, is more than a sequence of sentences strung together. The local sentence-level choices that must be made during transcription rely on the overall structure of the planned text and require a mental representation of both the text transcribed so far and the text plan. In collaborative writing, there are several factors that can adversely affect the sentence-level decisions that are made. In the remaining sections, I describe three factors that affect the transcription process that impact the overall stylistic congruity of the text.

2.6.1 Discontinuity from the Initial Planning Subprocess

In some collaborative writing practices, there is a discontinuity between the initial planning subprocess and the transcription subprocess. For example, Ede and Lunsford identified a pattern in which one individual makes the text plan and then assigns tasks to the other group members (organizational pattern number 7). The concept of the text's intended communication is developed during the planning and must be communicated to the transcribers; it is one of the two types of knowledge that they require. But not all aspects of a text's intended communication are easily communicated. The pragmatic aspects discussed earlier are especially pertinent to style, but are difficult to state explicitly. There may be an effective discontinuity even if the group plans together, but doesn't discuss these aspects. In the absence of a specific discussion, each group member may rely on their own version of these pragmatic aspects or their own defaults. Transcribers need continuity from the planning subprocess in order to have adequate knowledge about the text plan.

2.6.2 Lack of Global Perspective

As they translate a text plan into prose, writers need a global perspective of the text — a constructed mental representation of the text. Without an adequate global perspective, an author is prone to produce a text with poor organization and coherence (Severinson Eklundh, 1992). The text will likely contain inconsistencies as well. The global perspective guides the transcription subprocess, as well as the planning and reviewing subprocesses. But it is during transcription that writers particularly use word processing environments that can cause particular difficulties in constructing an adequate global representation. The difficulty stems from the writer's inability to handle a conflict arising from the nature of the word processing medium (Severinson Eklundh, 1992). The two opposing sides of the conflict are the following:

- Using the word processor affects the composition process. More specifically, writers tend to plan less when using a computer than when writing on paper. Additionally, both due to the reduced planning and the ease of modifying text, writers tend to revise more when using a word processor than when writing on paper. This intensifies the need for an adequate global perspective.

- When using a word processor, the writer's global perspective of the text is diminished. Severinson Eklundh argues that using this medium hampers the construction of a sufficient mental representation of the text and the changes taking place, because writers reading their own work have difficulties using spatial cues in the word processor's representation of the text. Spatial cues are essential to building a global perspective of a text because they support the reader's memory and orientation in the text. Skilled readers pick up on linguistic signals that provide global information about the content and structure of a text; the physical appearance of a text is important, as these cues are associated with spatial locations. It is difficult to use spatial cues in a text being developed on a word processor because of the scrolling required and the lack of fixed position of the text on the page. Scrolling is required to compensate for the display limitations; only part of the text can be viewed at a time and writers need to move back and forth several times to review what they have written. Paper copies, on the other hand, can be spread out, giving the writer a view of the entire text.

Although Severinson Eklundh's research targeted singular writers, the transcribers in a collaborative writing project are also faced with this conflict. For collaborative writers, the task of building a global representation is confounded even more, since not all segments of the text (from which the global representation is constructed) may be shared within the group. The lack of sharing is also related to another difficulty that collaborative writing groups experience, which is being "out-of-step" (section 2.6.4 below).

2.6.3 Variation in Language Knowledge

In a *segmented* approach to collaborative writing, the planned text is divided into segments. The writing of these segments is then assigned to group members or sub-groups. Examples of this strategy include Ede and Lunsford's organizational patterns 1 and 6, Posner's separate writers strategy and Sharples et al.'s parallel and sequential strategies. This approach contrasts with the strategy where a single group member is responsible for the transcription, such as Ede and Lunsford's organizational patterns 2, 3, 4, and 7; Posner's single writer strategy and possibly Sharples et al.'s reciprocal strategy.

In the unsegmented approaches, an individual or sub-group transcribes the entire text, so "one individual's style of writing is present in the text" (p. 52, (Posner, 1991)). As mentioned earlier, there are two types of knowledge guiding the transcription: language knowledge and knowledge about the text plan. In unsegmented approaches to transcription, there is one transcriber and only one transcriber and hence, only one, consistent source of the knowledge guiding the sub-process. Although this might not be sufficient in order to achieve stylistic congruity (i.e., stylistic incongruity occurs in spite of this), some degree of consistency is necessary.

For the knowledge that guides the transcription process to be consistent, the knowledge (or the property of possessing the knowledge) should be *homogeneous* in the group of collaborative writers. Since various segments of the text may be transcribed by different group members, all of these members should share the same knowledge about language and the text plan in order to achieve consistency in translating the text plan to prose. In section 2.6.4, we will describe why it is difficult to establish a concept of the text's plan that is shared homogeneously. But in addition to this difficulty, there is natural variation among the group members in their knowledge about language. In other words, language knowledge is *heterogeneous* among the group members. So even if knowledge about the text plan were shared completely, in a segmented text approach, the transcription of the text segments would be guided by inconsistent knowledge due to the inconsistencies in knowledge about language.

In the next section, an additional cause of stylistic incongruity is discussed arising from *heterogeneous* knowledge about the text plan among the members of the collaborative writing group.

2.6.4 The "Out-of-Step" Phenomenon

In a segmented text approach, each member "may have different perceptions of what they should be writing, and what their colleagues are producing, based on earlier plans and drafts." (p. 16, (Sharples et al., 1993)). From this inconsistency in perception, the group becomes "out-of-step." When a group becomes out of step, the various segments of the text start to diverge from the initial concept of the text's intended communication. This corresponds to a divergence in the transcribers' knowledge about the concept of what the text should convey. Since the style a transcriber uses in their prose is constructed from their perceptions, being out of step is very likely a condition which results in stylistic incongruity.

Common knowledge may evolve from planning activity, but not if there is discontinuity from the planning process. To reduce occurrences of being out of step, the group must negotiate a common knowledge about the text plan initially and keep open channels of communication to ensure the knowledge, as it evolves, remains common. One common practice in collaborative writing that addresses this need is the exchanging of plans and drafts of segments. If the segments start to diverge, the other group members provide feedback.

2.7 Reviewing

The purpose of the reviewing process is to evaluate the text to see if it meets the criteria for a good solution (Nold, 1981). Reviewing is more than evaluation though; in the case of a negative evaluation, a revising subprocess should be triggered so that the problems can be repaired. The reason that revision takes place is that authors want to fix the faults that they find. Revision doesn't necessarily follow from review. If problems in a text are not found or cannot be repaired, then revision cannot occur.

The skill of the author is reflected in their ability to perform this subprocess. Expert writers are more adept at finding and repairing faults in their own texts than novice writers. Novices, on the other hand, view revising as simple rewording, rather than reorganizing or reconceptualizing with the intended reader in mind. "Revising is not a sub-process in the same way planning, transcribing and reviewing are; rather it is the *retranscribing* of text already produced" (p. 68, (Nold, 1981)).

If difficulties in planning or transcribing the text result in stylistic incongruities, then the reviewing process is the last opportunity of the composition process to eliminate them and to meet the goal of stylistic congruity. If stylistic incongruities are not detected and repaired in this process, then the goal of producing a stylistically congruous text cannot be met. For this reason, leaving the goal of achieving stylistic congruity to be satisfied only during the revision process is not a good strategy. Since there are many difficulties in achieving this, it is likely to fail. Not only are detection and repair difficult tasks for writers, but the nature of collaborative writing confounds the reviewing activity. For a group of collaborative writers, the initiation of the review is not as flexible as for singular writers. Additionally, local revision activity requires a global context that is often not available. In the remaining sections, these issues are discussed.

2.7.1 Detection is a Difficult Task

In order to detect problems in a text, it must be evaluated with respect to the intentions of the author (Nold, 1981). In order to do this, it must be judged against the intended meaning, which requires that "writers must *conceive* their text's meaning (for an audience) separately from their *intentions*" (p. 74, (Nold, 1981)). Authors in general are not skilled at revision; authors often fail to detect problems (Kelly and Raleigh, 1990), (Shaughnessy, 1977). Writers need help not only to detect, but to diagnose accurately and to build a repertoire of remedies (Kelly and Raleigh, 1990).

The abilities to examine the text for faults and to take the role of the audience are late developing and are the sign of a skilled writer, as opposed to unskilled writers who view revision as "finding the right word" (Nold, 1981).

The ability to conceive the text's meaning for an audience is so crucial that Nold identifies this as one of the two most interesting issues for research in revision activity: tracing the development of the process of reviewing; and inferring the underlying representations (meaning, audience, writer) and knowledge (language, conventions) against which writers may be evaluating their texts. Understanding the underlying representations would be very useful for designing a computational tool that attempts to detect stylistic incongruities.

Hayes and Flower (1986) also argue that the complexity of the revision process is bounded by the depth of the preceding planning subprocess. This is true for collaborative writing groups as well. There are several patterns of collaborative writing in which the member responsible for the reviewing of the text was not involved in the transcribing or in the planning processes. For such a reviewer, the revision process is bounded by their knowledge (or lack of) of the planning subprocess. To detect problems, they must have clear idea of the concept of the text's intended communication in order to have something against which to evaluate the text. So in addition to the difficulty of the detection task, the reviewer also faces the difficulty of trying to construct this concept, which is difficult to represent and communicate.

2.7.2 Repairing is an Even More Difficult Task

Even after problems in the text have been detected, writers have trouble repairing their text (Nold, 1981), (Kelly and Raleigh, 1990), (Schriver, 1992). Writers have difficulty in diagnosing the problem and may have to resort to a strategy of rewriting. Less skilled writers often must resort to this hit-and-miss strategy; without understanding the cause of the problem, they simply rewrite the offending sections of text. Skilled writers, on the other hand, are able to rewrite or revise, and they choose the best strategy depending on the number and the nature of the problems in the text (Hayes and Flower, 1986).

2.7.3 Local Review Cannot Trigger Global Review

The reviewing subprocess can be initiated at any time and many times during the composition process (Nold, 1981). It is not simply during the final stages of the composition process. In fact, Hayes and Flowers (1980a) suggest that the monitor of the subprocesses has a predisposition to initiate a review over the other subprocesses. The review subprocess is continually at work, ensuring that the author's goals are met. In singular writing, the author is free to review any part of their text at any time. Collaborative writers do not have quite the same freedom. The segmentation of a text poses some obstacles for the reviewing process. Without first reassembling all the segments, a writer can only review their own segments. This local review is less effective than a global review for several reasons.

First, there may be problems in the text segment that are not detectable without the context provided by the remainder of the text. For example, a text segment may be a piece of fine prose, but its style does not match the styles of the other segments. By doing only local reviews, a writer may actually exacerbate the problem. Successive polishing will be guided by the writer's knowledge of language and the text plan, but if this knowledge is inconsistent with that of the other group members, it could worsen the problem.

Second, once the text has been segmented, the start of the global review process is dependent on the pattern of collaborative writing being used and is not necessarily dependent on the actual need for a review. It is not clear which dependency is better. According to the findings of Ede and Lunsford (1990), the success of collaborative writing depends on sticking to a well outlined plan. On the other hand, if the start of the review process is delayed, then any required revisions must be delayed as well, which leads to the problems described above.

The timeliness of the review is important, as it can trigger the revision activity required to keep the text from diverging stylistically. If the review is not initiated in time, then any required revisions must be delayed. During this delay, authors may spend a lot of time polishing, even if very basic modifications are still required. Successive polishing can exacerbate the stylistic incongruity since the author will use a revision strategy consistent with the knowledge used during the transcription, which was the initial cause.

2.7.4 Local Revisions Must be Made in Context

In some patterns of collaborative writing, the review of the text segments is performed by team members who are not involved in the planning or transcribing (see section 2.7.1 for difficulties arising from this). These members are not responsible for repairing the problems, only identifying them and making comments (which may or may not include a diagnosis of the problem). The responsibility for the repair then falls back on the original transcribers of the segments.

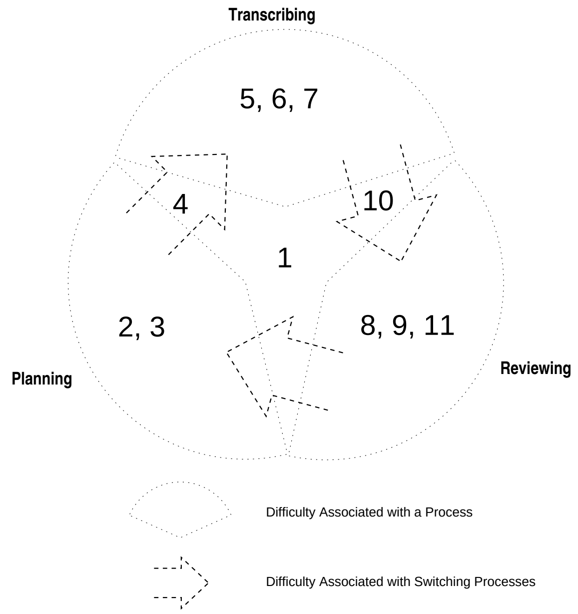
In order to perform the repair effectively, these writers must conceive the problem in the same way as the reviewer. With this understanding, the writer must choose a repair strategy to best correct the problem. But even this is not enough, since the text repairs may introduce new stylistic incongruities if the strategies for repair were not guided by common knowledge about language and the text plan.

To repair misunderstandings, the members of the group need to discuss and negotiate. This is achieved through the communication channels provided by the cycling back and forth of the text with added comments. This could be supported by the collaborative writing environment. For example, Mawby (1991) identified the need for collaborative writing environments to support commenting, including the capability of being able to handle comments on the comments from the other collaborators without affecting the original text of the document.

2.8 Summary

In this chapter, many collaborative writing practices have been examined. These practices have been classified according to the way in which the subprocesses associated with the composition process are handled by the group. This classification also was used in the analysis to find areas of difficulties for the collaborative writers. There are many potential difficulties in achieving stylistic congruity facing a group of collaborative writers; some of these difficulties are associated with performing a particular subprocess (which singular writers face as well), while others are associated with achieving the transition between subprocesses. The latter type of difficulty is more likely to arise for collaborative writers than singular writers due to the lack of a centralized monitoring subprocess.

These areas of difficulty are presented in a taxonomy in Figure 2.5. The difficulties are associated with one of the two possible types: process specific or transition specific. Then the difficulties are associated with a subprocess of the composition process. The numbers in the figure correspond to labels which are further explained in the table within the figure. In Chapter 5, these results are used in the discussions of the preliminary design.



Label	Area of Difficulty
1	Additional constraints don't ensure consistency (section 2.5)
2	Building a concept of the text's intended communication is difficult (section 2.5)
3	A common concept of the text's intended communication is essential
4	Discontinuity from the initial planning (section 2.6)
5	Lack of global perspective (section 2.6)
6	Variation in language knowledge (section 2.6)
7	Out-of-step phenomenon (section 2.6)
8	Detection is a difficult task (section 2.7)
9	Repairing is a difficult task (section 2.7)
10	Local review cannot trigger global review (section 2.7)
11	Local revision must be made in context (section 2.7)

Figure 2.5: A Taxonomy of Areas of Difficulty in Achieving Stylistic Congruity.

Chapter 3

Making Stylistic Assessments

3.1 Objective

In Chapter 2, several areas were described in which collaborative writers encounter difficulty achieving stylistic congruity. These problems occurred at all stages of the composition process and many of them involved difficulty in assessing the style of the text, both while it is being written and while it is being reviewed. A writer needs to assess the style of a text in different ways throughout the collaborative writing process. In some cases, a writer just needs to be aware of the stylistic nature of the other text segments that are being produced by the other collaborative writers in the group. Other times, a writer needs to detect the divergence of styles between text segments. Sometimes, a writer needs to assess a text with the goal of detecting stylistic incongruities. If any these areas of difficulty are targeted in the design of an software application, then the existing capabilities of computational stylistic assessment are relevant for the design.

In previous sections, we have been careful to explain what we mean by stylistic incongruity, but we have appealed to the reader's intuition when referring to style. In section 2.2, we described stylistic incongruity, its effects on the reader, and the conditions under which it is produced. Style, on the other hand, was only described as an expression of the author, and the pragmatic aspects of the author's communication were given as an example of part of the expression. At that point, we did not define what is meant by style, and, as this section will show, we cannot do so as easily as we defined what is meant by stylistic incongruity.

First, there is a methodological issue. The definition of stylistic incongruity is one level of abstraction higher than the definition of style. Characterizing stylistic incongruity as opposed to style is like characterizing an earthquake as opposed to the earth. If we have experienced earthquakes and know some rudimentary properties of the earth's crust, then earthquakes can be satisfactorily characterized, at least enough to describe the problem, to say why one would like to avoid them, and to say what steps are likely to precede your contact with them (e.g., moving to a city situated over a fault line). This is what we have done so far for stylistic incongruity. But, if you want to detect the type of earthquake in progress; or that one is going to happen; or, with some kind of Herculean force, if you want to prevent one from happening, then more than rudimentary knowledge about the how the earth works is required, as well as the right tools. So we need to know not only what is meant by the term *style*, but also how it works and any relevant tools in order to have the analogous abilities for dealing with stylistic incongruities. We have a functional model of the earth, and we assume that such a model can exists for style, but since it hasn't yet been developed, our ability to completely model stylistic incongruity is constrained.

In the next section, the various models of style are discussed, both theoretical and practical. In the final section, the construct/indicator model is introduced in order to synthesize theoretical and computational research and to provide a framework for improving stylistic assessment for the needs of future applications.

3.2 Existing Definitions of Style

3.2.1 Defining "Define"

Before we describe research related to theories of style, a methodological clarification is required. The theoretical issues involved with defining style involve a lot of territory, such as the philosophy of communication and literary theory. As well, definitions of style originate in diverse fields of research and have been developed to suit different needs. These factors make for a landscape of research results that is not easily navigated. There is no established metalanguage with which to discuss style, to describe stylistic effects, or to describe the role of subjectivity in the perceptions of the readers of a text, so it is extremely difficult to make a guidebook for this terrain, much less neatly categorize and classify its landmarks. For example, the word "definition" means different things for the different definitions of style and represents either a monumental task or a trivial one. The following statement about consciousness (another difficult concept to define) is quite relevant:

> It is supposed to be frightfully difficult to define the term. But, if we distinguish between analytic definitions, which aim to analyze the underlying essence of the phenomenon, and the common sense definition, which just identifies that we are talking about, it does not seem to me at all difficult to give a common sense definition of the term[1].

The point of the distinction seems to stress the importance of first identifying the phenomenon being studied and then to analyze it. As illustrated by Enkvist's often-cited observation (1973), "style is a concept as common as it is elusive: most of us speak about it, even lovingly, though few are willing to say precisely what it means (p. 11)," this advice hasn't really been heeded. There is a lack of common-sense definitions and instead, an abundance of incomplete or unsatisfactory analytic definitions.[2]

3.2.2 Identifying What We Are Talking About

After reading reams of papers spanning the last thirty years, I (along with most others) noticed that there are nearly as many different ways to 'define' style as there are papers, and that the papers shared few commonalities. The first commonality, described in the previous section, was the lack of precision in the terminology. If the authors said "we define..." then "define" could mean several things, such as: to model, to describe, or to attach a label to. Because of this, a vast array appeared of what the authors were identifying as the topic of their discussions, as shown in Figure 3.1.

The listing in this figure isn't intended to be exhaustive or to contain only mutually exclusive definitions; many of these definitions of style entail or overlap other definitions. For example, defining style as the author's choice entails the definition of style as a function ("based on") the intention of the writer. One way to look at all these definitions is that they are all elaborations of one very basic 'common-sense' definition. Such a common-sense definition is too broad to be useful, but the elaborations are the start of 'analytic' definitions and try to get at the underlying essence of style. One possible common-sense definition is

[1] John R. Searle. The Mystery of Consciousness. *The New York Review of Books* XLII(17):60–66, Nov. 2, 1995.

[2] Since this is the case, the term 'stylistic theory' will be used instead of the term 'stylistic definition'.

Style:
Defined as (Defined = "Specified in terms of an object")

- the variation with respect to a property or quality of a text
- a dimension of language variation
- information; part of the communication
- something to be chosen or selected
- a classification or characterization
- the reaction of the reader
- something that exists in the mind of the reader, a constructed mental representation
- the expected norm by the reader

Defined as (Defined = "Specified in terms of a process")

- the author's choice or selection
- a mapping from the topic or subject matter to specialized language

Defined as (Defined = "Specified in terms of a quality")

- a property that is desired and must be obtained
- a quality of a particular type of prose
- a quality of a genre
- the amount of variation from a norm
- the amount of variation that has a purpose

Given as a function (function = "based on, depends on, or determined by")

- the topic or subject matter
- the reader's interpretation
- the reader's emotive effect
- the norm of a genre
- the purpose of the text
- the intention of the writer
- the writer's expectation of the reader

Given as a function (function = "with a purpose, done with intention")

- a functional variety

Given as a function (function = "a mapping")

- of the topic of subject matter of discourse to a text type or specialization of language

Figure 3.1: A Thumbnail Sketch of Various Definitions of Style.

that style is information, part of the author's communication, and, therefore, part of the text's meaning. If all the definitions of style shared this foundation, then each definition's answers to the following questions could form a basis for categorization:

- What is being communicated?
- Why is it being communicated? What purpose does it serve?
- How is it communicated?

Even this broad definition is incompatible with a whole class of stylistic theories that Cluett (1976) identifies as *Theories of Ornate Form* (p. 5). Fundamental to these theories is the "acceptance of the possibility of synonymity and ... the idea that style must be separable from the linguistic medium out of which it is produced" (p. 5, (Cluett, 1976)). For the possibility of synonymity to exist, an author must have the ability to produce multiple texts, each with different styles, but all conveying the same communication. Since style is separable from content, style and content are equated with two independent sets of choices that must be made by an author while writing. In ornamental theories of style, the stylistic choices that are made by an author (explicitly or otherwise) do not carry any consequence for the meaning of the text. These theories are rejected by computational linguists who espouse the belief that stylistic choices do affect the author's communication (see DiMarco (1990), Hovy (1988)).

Theories of ornate form aside, the remaining views of style still vary widely and defy the neat three-step classification given above. But if theories of ornate form are rejected, then the task of categorizing all the different definitions of style is made easier. The concept of style, in some form, must have a role in the text's communication and therefore must be accounted for in a functional analysis of communication. This functional model of communication can be used as a basis for categorization.[3] A functional model of communication will have a decomposition into parts, with an explanation of each part's capabilities and their interrelations. For example, Hartmann (1981) uses the so-called communication model as a basis for categorizing the differing views of style (more specifically, the differing views that define style as a dimension of language). The parts of the communication model are: encoder, topic, decoder, code, medium, context, and message. On the basis of the relative emphasis that may be placed on these parts, he distinguished the following categories of style and used them to classify widely differing views of style:

1. The encoder's rhetorical choice;
2. The genre of a subject;
3. The decoder's reaction to a particular text;
4. A kind of code or dialect;
5. A norm imposed by a medium or a role;
6. A contextually determined variety; and
7. A text idiom.

Hartmann also sorted these categories into a rough chronological order. They are an appropriate means of classifying views of style in addition to those in which Hartmann was

[3] As opposed to general overviews, (e.g., (Spencer et al., 1964)) or chronological overviews (e.g., (Green, 1992)).

interested. In the next section, the principal definitions of style are discussed with respect to these major categories.

3.3 Theories of Style

3.3.1 The Encoder's Rhetorical Choice

This characterization of style is the oldest and most common. It began in classical rhetoric. Original rhetorical theory, considered synonymous with stylistic theory in many minds (p. 5, (Cluett, 1976)), has made many important contributions.

> Rhetoric from the *Ad Herennium* through the early Renaissance considered discourse (written) under three heads: Invention, or what is said, the topics; Disposition, or the ordering and arrangement of what is said; and Elocution, or the tricking out of the properly disposed matter with tropes and figures. (p. 5, (Cluett, 1976))

Elocution is especially relevant as it concerns the identification of linguistic devices that are used, namely tropes and figures. From this, the first metalanguage with which to describe style was created. This metalanguage is still used (e.g., in handbooks of rhetorical terms, such as (Lanham, 1991)).

The motivation of this classical viewpoint was "how *can* one communicate?" This contrasts sharply with "this is how one *should* communicate," the viewpoint of the stylistic prescriptivists. Style to prescriptivists is still a matter of choice between alternatives, but in this case, there is a correct choice and an incorrect choice. Proponents of this view hold that a text with style is one that has achieved the universal and correct mode of expression (DiMarco, 1990). In order to help achieve this, one should adhere to a set of basic prescriptive rules, conveniently located in several handbooks (e.g., such as (Strunk and White, 1979)). Unfortunately, adherence to these prescriptive rules does not guarantee the avoidance of stylistic incongruities for collaborative writers. The examples in Chapter 1 follow the rules set out in Struck and White (1979) and yet are stylistically incongruous.

Another view holds that style mirrors the author's personality, that "how one communicates is a *reflection* of their unique personality." The style of a text is determined by this reflection and is described in terms of features. The features result from choices an author makes, both intentionally and unintentionally. Doležel (1969), for instance, asserts that "stylistic features are apparently consciously controlled to only a limited degree" (p. 10). Cluett (1976) gives the definition that "literary style is that set of propensities that define an author's voice" (p. 8) and that "a writer's style is an aspect of whatever distinctiveness he [!] possesses and therefore is an extension of his personality" (p. 6). Since the skill of an author determines the intentional choices and the sub-conscious mind determines the unintentional choices, this reflection can be used as a fingerprint. Stylometry is the science that describes and measures the author's fingerprint. This is the underlying assumption in what Cluett (1976) describes as "Individualistic Theories." Work in this area is motivated by the desire to understand the nature of the fingerprint, the dimensions of the fingerprint, and how differences in fingerprints can be described. Statistical studies are designed to reveal the stylistic features that account for stylistic differences between two texts. For example, Cluett (1976) inventories different authors' sets of propensities with respect to syntactic features. Similar to this is the work by Biber (1988), which characterized the style by genre (e.g., newspaper, academic prose) in terms of syntactic propensities. There are several different applications of the authorial fingerprinting, such as authorship attribution, forensic studies, and imitation studies.

One motivation is to confirm the claimed or the assumed authorship of written texts. By far, the largest attention is focused on Shakespeare and Marlowe, as attest to by the huge number of studies (e.g., see (Brainerd, 1973)). An example of a stylometric measure serving as basis for authorship is the the Thisted-Efron Authorship test, which was evaluated and found to be effective (in terms of observation and theory) (Valenza, 1991). The domain for this test is drama and poetry, and it is limited to comparing texts of the same genre and aggregates of samples rather than single samples. Related to the area of authorship attribution is forensic studies, which are motivated by the desire to discover which author wrote a particular text.

Another motivation is to discover how a style can be imitated. Imitation studies differ from authorship attribution studies. The authorship and authenticity is not in question, since it is already known that the work is an imitation. The imitator deliberately sought to imitate texts and therefore text should "share certain salient and objectively identifiable characteristics" (p. 228, (Irizarry, 1989)). Standard stylometry can be used to help a critic distinguish the ways in which the author achieved the mimicry (Irizarry, 1989). These mimicry studies have the potential to help authors achieve stylistic consistency. If a text element, such as a paragraph, is incongruous with the rest of the text, one might simply use mimicry techniques to modify an offending paragraph here and there. Unfortunately, the stylometric analysis used in the mimicry studies consists of a series of statistical tests, based on lexical and syntactic occurrences and this type of information is not very usable to a writer making revisions.

3.3.2 The Decoder's Reaction

According to the viewpoint that style is the encoder's rhetorical choice, and assuming that style is not a purely ornamental feature of text, the choices relating to style carry a consequence for the author's communication. But the meaning of the resulting communication is not simply received by the reader; rather, it is constructed (Garnham and Oakhill, 1992). Since style is part of the text's meaning, it must also be constructed. This is the basis of another group of viewpoints that hold that style is defined in terms of the decoder's reaction to a particular text. Hartmann (1981) says "the interpretive task of the stylistician is to isolate those individual features which produce a certain emotive effect on the decoder" (p. 264). This viewpoint acknowledges the subjectivity of the audience's assessment of style, and it is more appealing than the view that style is a matter only of the author's rhetorical choice both for its completeness and verifiability, since correlations of stylistic features with the audience's impression can be verified by empirical studies.

There are many literary studies that attempt to isolate individual features in a text that produce an emotive effect on the critic, but they are not very applicable. These studies are generally criticized as too intuitive and impressionistic to carry much weight (Hartmann, 1981), (Winter, 1969). Additionally, these literary studies have a narrow domain, namely literature rather than everyday text. Additionally, to adapt techniques used in literary stylistic analyses to computational applications would be inherently incompatible with the whole premise of literary criticism; that is, it is the skill and insight of the critic that produces the analysis and not some algorithmic process.

On the other hand, empirical studies are shaped by well defined elements. A particularly relevant study was carried out by Carroll (1960) and extended by Dale (1977). In an effort to quantify aspects of literary style, Carroll asked a set of human judges to rate 150 different texts, using a set of 68 different measures. This data, after being subjected to factor analysis, revealed that the stylistic judgements varied with respect to six meaningful dimensions. Although the identification and naming of these dimensions is the main contribution of the study, an interesting side-effect of the empirical basis of the study is relevant here. Of the 68 different measures, roughly half were objective measures, such as sentence length, and half were subjective measures. For the six dimensions identified, one consisted solely of subjective measures and another consisted solely of objective measures. The remaining four dimensions were found to be based on a combination of both objective and subjective measures. There was some lack of agreement among the judges (calculated using linear correlation) for the subjective measures for a substantial portion of the sample texts; however, some subjective measures were more stable than others. It would certainly be useful if the objective measures of these dimensions could be used as indicators of the values of the subjective measures. These measures cannot completely predict human judgement, but they could be used to make stylistic assessments that correlate with human judgement. This model for making stylistic assessments of subjective measures will be further discussed in section 3.5.

3.3.3 A Combination of the Encoder, the Decoder, and Other Factors

Crystal and Davy (1969) proposed a more multi-dimensional viewpoint of style. In their view, style was a function of social context. Stylistically significant linguistic features were chosen for their extra-linguistic purpose. Therefore, this viewpoint of style included the additional dimensions of medium, context, and message (Hartmann, 1981), in addition to dimensions of encoder and decoder on which the previously mentioned stylistic viewpoints concentrate. Their goals were to develop a metalanguage in order to describe these features; to identify the stylistically significant features; and then to classify these features based on their function, although these goals were only partially realized.

3.4 Stylistic Analysis in Existing Applications

Even though there isn't much common ground or consensus in the collection of research areas that take a theoretical approach to defining style, the development of software applications that deal with style has proceeded. In this section, several computational applications are discussed.

Approach	**Purpose**	**Method**	**Audience/ Genre**	**Application Name**
Subjective	Writer's Aid	*a priori* rules	Absolute	CorrecText
				Reader
				RightWriter
			Relative	Grammatik
			Relative	CorrectGrammar
		Heuristic	Relative[a]	Ruskin
		Grammar	Absolute	AECMA Simplified English Checker
	Stylistic Instruction	*a priori* rules	Absolute	McRuskin[b]
Objective	Writer's Aid	*a priori* rules	Absolute	Writer's Workbench[c]
				PowerEdit
	Stylistic Instruction	Grammar Subset	Absolute	STASEL
	NLG	Heuristic	Relative	PAULINE[d]
	MT	Grammar	Absolute	STYLISTIQUE[e]

[a]In addition to specifying the audience and genre, the user may also specify the purpose of the text.

[b]The focus of this project was to determine how to present the diagnostic and repair advice to the user. As a result, only "overly long" sentences were detected.

[c]In order to help interpret the results, a second tool, mkstand, can be used to develop a stylistic standard.

[d]Implicit in the heuristics used to generate the text is a stylistic assessment that is ultimately valuative as well.

[e]This method also attempts to connect the evaluative information with valuative judgements of clarity, dynamism, etc.

Figure 3.2: Overview of existing software with stylistic assessment capabilities.

These applications serve many different purposes and have forms other than the common 'style-checking' feature that often appears as an additional facility in word-processing applications. Even though the implementations may vary, they share the property of having to make some kind of stylistic assessment.

In Figure 3.4, a list of these applications is presented. I have made a classification based on the following dimensions: Approach, Purpose, Method, and Audience/Genre (the fifth column lists the application names and is not a dimension).

The dimensions of Approach and Audience/Genre characterize the underlying stylistic theory (implied by the implementation, if not stated explicitly). The dimension of Purpose is used to distinguish between the different types of applications and to highlight the different purposes served by their respective stylistic assessments. The dimension of Method is used to characterize the different implementations for making computational stylistic assessments.

3.4.1 Approach

In section 3.2, a number of different viewpoints of style were presented. In particular, stylistic prescriptivism was discussed. As a view, stylistic prescriptivism holds that style is a desirable quality of writing that is achieved only when a correct mode of expression is found. Not surprisingly, many software applications have been created to help authors with this pursuit, usually in the form of providing feedback during the writing process. From the point of view of the user, the feedback provided by the applications includes an evaluation of the quality of the writing with respect to this view of style. For this reason, applications embodying a philosophy of stylistic prescriptivism are described as *subjective*. The feedback might describe a text (or part of it) as unacceptable, bad, vague, wordy, or some other subjective evaluation that conveys that the text is problematic and should be fixed.

Embodying a less-judgemental approach are the *objective* applications. Again, the underlying viewpoint is that style is a quality of the text, but the difference is that this viewpoint does not include a value judgement about the appropriateness of the style. The assessments made by evaluative applications attempt to capture the salient aspects of style and present them to the user, who is then left to interpret this information.

A key issue for this type of stylistic assessment is determining which aspects of style are salient. This determination is often based on convenience (e.g., the stylistic indicators that can be computed cheaply, like average sentence length) rather than any theoretical model of style. For example, a common indicator of style is the average number of words per sentence, even though there is no basis to believe that this information is useful to the user (it might actually be confusing). Some applications achieve their valuative approach by taking an evaluative approach and then making a judgement based on this.

3.4.2 Purpose

As mentioned earlier, there are applications other than style-checkers that make stylistic assessments. For example, there are applications in machine translation, natural language generation, and intelligent computer-assisted language instruction which make use of stylistic assessments.

For a machine translation application, it is important to assess the style of the source text in order to produce an analogous style in the target language (DiMarco, 1990). The stylistic assessment alone is necessary but not sufficient for achieving this. The translation application must also understand the cultural and language differences between the source and target languages in order to determine exactly what the analogous style is, if it exists, or the closest match.

For a natural language generation application, it is important to understand that the style of a text is shaped by its intended effect and meaning. The form of the stylistic assessment for this type of application is a representation of the style-dependent factors of the text's intended communication, as well as a representation of the knowledge required to apply this information to the generation process.

For applications in intelligent computer-assisted language instruction, it is important to be able to provide instruction to language learners about stylistic issues in the target language. An intelligent computer-assisted language instruction application uses the review stage of the composition process as a natural place to provide language instruction about the user's written prose and to communicate instructional feedback in a way that is effective (both understandable by the user and easily applied to further writing). For example, the stylistic instruction tool STASEL (Payette and Hirst, 1992) uses this opportunity to help language learners. STASEL uses a hybrid approach of flagging traditional indicators of stylistic problems, such as wordiness and passive sentence structure, and using a grammar-based approach to detecting stylistic constructs (see section 3.4.4 below descriptions of flags and grammars). These applications, unlike the writer's aids, give some thought to how the information should be communicated to the intended user, in addition to what should be communicated. For example, the application, McRuskin (McGowan, 1992), focuses instead on how stylistic diagnostic and repair advice should be communicated to the intended users.

In addition to these three different application domains, stylistic assessment is also important to a whole class of software applications designed as writer's aids. The most common form of a writer's aid is the style checker, a stylistic analogue to the spelling checker. This type of application, either a stand-alone application or a facility integrated into a word processor or desktop publishing environment, aims to help a writer produce stylistically correct[4] text by assisting in the detection, diagnosis, and repair of stylistic

[4]The tool's definition of 'stylistically correct' theoretically should include stylistic congruity, but in practice the definitions are not that sophisticated.

problems. The stylistic assessment is typically in the form of a report, with problem areas identified and accompanied with offerings of advice for repair.

3.4.3 Audience/Genre

In addition to the issue of prescriptivism, there is another theoretical issue that all software applications address. In section 3.3.1, a group of definitions were discussed that consider style with respect to the choices that an author makes. For these types of definition, the role of the audience or text genre in style is downplayed. For other definitions, the audience or genre plays a key role in the definition. Correspondingly, the software applications acknowledge the role of the audience or genre to different degrees. With respect to these factors, the stylistic assessment performed by an application might be *absolute* or *relative*. For *absolute* stylistic assessments, these factors play no role in the stylistic assessment. For *relative* stylistic assessments, these factors are do play role — they help determine if the style of a text is correct. The relative stylistic assessments implement this influence differently. Typically, the factors of audience/genre are represented as a pre-determined category or a combination of several categories (e.g., audience is represented as "friend" or "business colleague" and the genre is represented as "business correspondence" or "personal correspondence"), although there are more sophisticated representations. For example, the natural language generation application PAULINE makes its concept of style dependent on a wide range of pragmatic factors which subsume the simple audience/genre distinction. In all cases, however, the user must help the application with the stylistic assessment by selecting from a set of predetermined settings for these factors. These relativistic concepts of style are different from the absolute concepts because the user's choice affects the way in which subsequent stylistic assessments are performed.

3.4.4 Method

In the previous sections, the three dimensions of Approach, Purpose, and Audience/Genre were described. These dimensions were described first because they capture a lot of the variation among the different types of software applications that perform stylistic assessments. In this last section, the different software applications are described with respect to the last dimension, Method. The different ways of performing and implementing stylistic assessment, which are based on flags, heuristics, or grammars, are described.

Flags

In order to make stylistic assessments, many software applications, such as writers' aids and stylistic instruction tools, make use of *flags*. A flag is an identifiable occurrence in a text. For instance, a spelling error is a flag of sloppiness on the writer's behalf and therefore a flag of a sloppy writing style. A colloquial word (as defined by a dictionary) is a flag of an informal writing style. Of course, there is a multitude of possible flags, so the software application tries to make use of the flags that are more directly related to some kind of stylistic property.

Writers' aids aim to help writers detect and repair stylistic problems. Applications use a set of flags that are thought to be related to stylistic problems, in order to detect such problems and to provide repair information. If the flag is detected, then the occurrence is considered to be the cause of a stylistic problem, and removing or modifying the offending occurrence is considered an appropriate repair. For example, a colloquial style supposedly corresponds to a text in which slang and informal terms are found. The WordPerfect style checker would identify "The issue is not all that important" as colloquial and would propose the substitution "not very" for "not all that".[5] In Figure 3.3, a set of common flags is listed.

There are some advantages to using flags. In many cases, they are simple to detect. Additionally, many flags do correspond to stylistic problems. However, there are several drawbacks.

The quality of a flag-based evaluation is constrained by the quality of the flags. The effectiveness of the flags, in turn, may be constrained by the underlying theoretical assumption that all stylistic problems do indeed have a correspondence with some kind of flag that is observable within the text (rather than with how the text is processed or perceived by the reader).

Additionally, flags are used because they correspond to some stylistic quality of interest, whether it is a stylistic problem or an important aspect of style (e.g., formality). Software applications make assumptions about the value of what these flags indicate. For example, the applications assume the flags correspond to certain types of style that are problematic and need to be repaired. Usually, this judgement is justified by style manuals or guides, which, in turn, make assumptions about the writer's goals. But these assumptions are not

[5] See the on-line documentation for the 'Grammatik' facility within WordPerfect version 6.0. for additional examples.

- Abbreviations
- Archaic words
- Beginning a sentence with a conjunction
- Clichés
- Colloquial words
- Commonly confounded words and phrases
- Consecutive nouns
- Consecutive prepositional phrases
- Contractions
- End of sentence prepositions
- Inappropriate prepositions
- Informal expressions
- Jargon
- Lack of sentence variety
- Misspelled foreign expressions
- Misspellings
- Misplaced modifiers
- Misused words
- Multiple negation
- Open and closed spellings (use of hyphens in phrases)
- Overstated language
- Overuse
- Paragraphs with only one sentence
- Passive voice
- Pejorative language
- Pretentious language
- Questionable usage
- Redundancy
- Second-person address
- Sexism
- Split infinitives
- Stock phrases
- Ungrammatical expressions and sentences
- Use of foreign phrases
- Use of 'that' and 'which'
- Use of trademarks
- Use of 'well'
- Use of words that end with -wise or -ize
- Vague adverbs
- Vague quantifiers
- Weak modifiers
- Wordiness, wordy expressions

Figure 3.3: Summary of Common Flags for "Style Checkers".

always appropriate and may conflict with the writer's actual goals. For example, a writer's aid might flag all passive sentences, since style manuals advocate clarity as a stylistic goal. But the writer may be creating a business document with the intent of being evasive and may need to use passive sentences.

One way to avoid this conflict is to allow users to turn certain flags off. Along the same line, many software applications allow sets of flags to be selected for particular texts. These sets serve as profiles and are either pre-defined to correspond with particular audiences or genres, or are customizable by the user. But there are some problems with this practice. First, the applicability of flags to a particular audience or genre has only a heuristic basis and has no theoretical justification. Users who face the task of creating profiles with sets of flags must also make heuristic judgements as well, but the typical user of such tools does not have the experience or expertise to make these kinds of judgements.

The flags themselves may be of poor quality, since some flags cannot be detected reliably. For example, stylistic assessments are usually intertwined with grammar checking. Many stylistic flags are determined by the occurrence of undesirable syntactic constructions (e.g., too many prepositional phrases). The performance of these tools is so poor that the information upon which these flags rely is faulty or inapplicable. The user must be able to distinguish between the faulty information, and the useful information and since such a small proportion of the stylistic problems are identified, the tool cannot be relied on (Bolt, 1993). Also, the relationship between flags and the style of the text is not always clear. For example, it is not clear to which stylistic property the flag based on the average number of words per sentence is related nor whether this information is particularly useful.

There is no way of knowing how to compare texts on the basis of the stylistic assessments provided by these flags. For example, given two segments of a text, how does one know which flag differences are important and which flag differences are inconsequential? Additionally, these flags only serve to detect stylistic properties at the sentence level. For problems of style occurring between sentences, it is necessary to interpret a collection of sentence-level flags.

Another problem is the validity of the relationship between the flag and the stylistic property to which it allegedly corresponds. For example, most “style checkers” produce readability scores. These statistical tests are considered indicators of the text’s understandability, but there is serious criticism of the validity and adequacy of these indicators (Baker et al., 1988).

Heuristics

In the previous section, several applications were described that use flags as indicators of higher level abstractions, such a stylistic problem or a relevant stylistic property (such as formality). We showed some difficulties with this approach, namely the quality of the flags, the validity of the flags as indicators, and the assumption that flags exist for all the desired stylistic properties of interest. In addition to these problems, users are faced with the task of deciding which flags are relevant. These applications provide little help in this way. The only guidance provided might be in the form of predefined sets of flags, each associated with a particular audience or genre type.

In contrast, heuristic-based stylistic assessments attempt to select flags for stylistic assessments a little more intelligently. Applications using heuristics require as an input a representation of the context, within which the style of a text will be evaluated. These applications have a large set of heuristic rules that specify which flags are relevant for a given context. With the representation, the application can determine which heuristic rules are relevant for any particular text, and therefore, which flags should be applied.

For example, the writer's aid Ruskin ((Holt and Williams, 1989), *cited in* (McGowan, 1992)) uses a heuristic approach for stylistic assessment. In order to use this style checker, the user must first specify a number of contextual variables. From these variables, "Ruskin is able to build up a so-called 'ideal model' of what a document of that type should consist of in terms of 'good style' " (p. 299, (McGowan, 1992)). The contextual variables are classified into the following five categories:

1. Audience,
2. Purpose,
3. Subject,
4. Use (how the text will be used), and
5. Author.

Ruskin also has a set of production rules, in the form of IF-THEN rules. For example,

(1) IF THE AGE OF THE AUDIENCE IS LOW, THEN USE FEW LONG SENTENCES.

The values of the contextual variables determine whether the IF part of the rule is satisfied, and, therefore, determine which production rules are applied to a text. Notice that the THEN part of this particular rule is basically prescribing the use of a flag based on sentence length. Flag-based writer's aids would detect all long sentences, but the user would need to determine whether long sentences are inappropriate for a particular text. The heuristic-based approach, on the other hand, embodies the knowledge that long sentences are inappropriate for young audiences.

Hovy's (1988) natural language generation application PAULINE also used a heuristic approach, but with a different focus. This focus used heuristics embodying knowledge about style in order to generate text rather than to assess the style of a pre-existing text. Like Ruskin, PAULINE has a representation for the text's context (upon which the applicability of heuristics is based); this consists of 23 pragmatic features, each of which is a range of three values. These pragmatic features can be categorized as pertaining to interpersonal goals (e.g., to teach, or to earn the reader's respect) or conversational settings (e.g., describing the tone of the conversational atmosphere).

This application uses two sets of heuristics; one set is used to determine the target *rhetorical goals of style* based on the pragmatic settings and the other set is used to create a text with a style that satisfies a given set of such goals. The rhetorical goals of style serve as an intermediate level between the low-level decisions that a text generator must make (thereby creating a text with a particular style) and the high-level specification of the author's pragmatic goals. They are intended to capture the qualities that determine style. Hovy proposed a set of 12 rhetorical goals of style: formality, simplicity, timidity, partiality, detail, haste, force, floridity, color, personal reference, open-mindedness, and respect (p. 34, (Hovy, 1988)). He points out that classifying all possible styles is an impossible task, and he acknowledges that his classification is but one possible classification and might be incomplete or inconsistent (pp. 32–33, (Hovy, 1988)). He does claim that these goals contain the common rhetorical styles and that most others are refinements or extensions of them. Notice that the task of developing this set of rhetorical goals of style is similar to the stylostatistician's task of identifying all the dimensions along which a text's style varies.

The first set of heuristics, developed to link the many combinations of pragmatic settings to a corresponding set of rhetorical goals of style, was created empirically. With these heuristics, the generator would 'know' to create a text with a colloquial, arrogant, or forceful style, given a certain pragmatic representation. Of more interest is the second set of heuristics. These heuristics link text-level qualities to particular rhetorical goals of style. This link could also be useful for other applications by mapping back from text-level qualities to rhetorical goals of style (provided that the rhetorical goals of style are relevant).

The text-level qualities, classified in terms of the decisions that the text generator must make, are given in terms of:

- topic collection,
- topic organization,
- sentence organization,
- clause organization, and
- phrase and word choice.[6]

For example, in order to make a text seem more formal, Hovy's heuristic includes the following: use many adverbial clauses; build parallel clauses within sentences; use the passive

[6] In the subsequent section 'Grammars', the machine translation application STYLISTIQUE will be described; this application maps text-level qualities in terms of the last three categories, sentence organization, clause organization, and phrase and word choice, to goals of style (an analogous set of goals to the rhetorical goals of style).

voice; use more complex tenses such as the perfect tenses; and avoid ellipsis (p. 85). This heuristic for achieving formality could be used to assess the formality of a text, assuming that it is possible to detect computationally the presence of the text-level qualities that it suggests. That is, the heuristic is 'reversed' to map text-level qualities to a rhetorical goal of style instead of its original mapping from rhetorical goals of style.[7]

These heuristics have a potential use in applications that make stylistic assessments, but a number of issues need to be addressed.

Because the heuristics are only general 'rules of thumb,' they do not have a theoretical basis. Although they do have some basis in empirical studies and in established research (e.g., (Brown and Levinson, 1988) was used to develop indicators of formality), in general, it is not known how accurate or correct they are. The validity of the heuristics needs to be established through empirical study. Also, the set of heuristics covers only a subset of the rhetorical goals of style.

In addition to these practical issues, there are some theoretical concerns. These heuristics give only one way of possibly many to achieve a particular goal of style. Therefore, the 'inverse' of the heuristic will properly assess a rhetorical goal of style if it was achieved in the same way as specified in the heuristic rule. So as an indicator, the inverse heuristic may be incomplete. Additionally, decisions made by the generator can be difficult to reconstruct from a text. For example, to produce formal text, the generator should produce sentences that have causal, temporal, or other relations to other sentence topics (p. 85, (Hovy, 1988)). It would be difficult to detect this computationally and may even require some level of natural language understanding.

Grammars

In the previous two sections, applications using flags and heuristics were described. These applications shared the foundation that the basis of stylistic assessments is the relationship between text-level flags or indicators and stylistic constructs (such as stylistic problems, instances of bad style or a particular quality of style, such as formality). These applications differed in the way in which the flags or indicators were given relevance. For flag-based

[7] Notice that the mediating layer that the rhetorical goals of style provides is crucial. The heuristics in Ruskin cannot be 'reversed' in the same way. If Ruskin's rules were to be 'reversed,' the rule would link text-level qualities with contextual variables rather than a stylistic construct.

writer's aids, the user must decide which indicators are relevant. For heuristic-based applications, the heuristic rules try to embody this knowledge rather than imposing this task on the user (but instead require a specification of the text's context from the user in order to make the right decisions).

In this section, stylistic assessments based on grammars are discussed. For these applications, the relevance of the indicator is assumed *a priori*. Rather, the focus is on the detailed and rigorous specification of how the values of the indicators are derived.

The Boeing Simplified English Checker, developed at the Boeing Advanced Technology Center (Hoard et al., 1992), is a computational tool that produces a stylistic assessment for a given input text (as well as a grammatical assessment). Its purpose is to provide a detailed report on a text's grammatical and stylistic deviations from AECMA (Association Européene des Constructeurs de Materiel Aerospatial) Simplified English, which is an international writing standard for aircraft maintenance manuals, intended to help improve readability. AECMA Simplified English has a wide variety of grammar and style restrictions, ranging from broad expository considerations to simple syntactic prohibitions.

The notion of style in AECMA Simplified English is more restricted than regular, written English and incorporates the idea that style is the adherence to a norm. This definition of style is quite different from the definitions used by previously discussed applications. Even though deviation from AECMA Simplified English is not an issue for general stylistic assessment, this tool has two contributions. Since adherence to AECMA Simplified English can serve as an indicator of understandability, the methods used by the Boeing Simplified English Checker could be reused for performing general stylistic assessments. Additionally, the Boeing Simplified English Checker enforces consistency in low-level textual elements, which is related to stylistic congruity.

The assessment is done by a parser, using rules of the AECMA Simplified English grammar. Since the motivation for creating the standard was to improve readability, it mandates the avoidance of both semantic ambiguity and communication through stylistic content. Even syntactic mechanisms which affect style, such as prepositional attachment, are strictly prescribed. Thus, the Simplified English Checker does not perform any type of semantic or pragmatic analysis. For example, it does not check certain mandated criteria, such as that verbs show action whenever possible or that instructions be as specific as possible, since this requires human judgement. It does check mechanical features such as paragraph length, compound nouns, articles, passivization, and dependent clauses. For example, the consistency of labels and noun group composition is checked. Consistency at this level has been identified as important for achieving stylistic congruity ((Farkas, 1985), p. 29, (Mawby, 1991)).

There exists another grammar-based application, whose definition is more useful. STYLISTIQUE, developed by DiMarco (1990), is an application that incorporates stylistic assessment into machine translation. Instead of evaluating style as the deviation from a prescribed norm, this tool performs stylistic assessment in terms of the author's stylistic goals. At the basis of this application is the identification of these *stylistic goals*. These stylistic goals, which capture the communicative intent of the author. They are divided into three dimensions: *clarity* and *obscurity*; *abstraction* and *concreteness*; and *staticness* and *dynamism*. These goals are important, since it is not always the case that an author wants to achieve the goals assumed by stylistic prescriptivists and by style manuals. Rather, the style of the text is chosen deliberately and is part of the author's communication.

The other component of the application is the grammar, which is a detailed specification under which these stylistic goals are achieved. Conceptually, the stylistic grammar is intended to capture lexical, syntactic and semantic aspects of style. In practice, the stylistic grammar has been implemented to capture syntactic aspects of style, although a partial account of semantic style has also been implemented by Ryan et al. (1992).

The evaluation of a text with respect to these stylistic goals is carried out by using three stylistic grammars in sequence; the grammar of primitive elements, the grammar of abstract elements and the grammar of stylistic goals.

The grammar of *primitive elements* is a precise specification of the correspondence between grammatical English sentences and a corresponding representation in terms of *primitive stylistic shapes*, which includes terms denoting regular syntactic constituents, such as adjective or sentence, and terms denoting combinations of stylistic effects and syntactic constituents. This representation is constructed in tandem with a syntactic parse.

The grammar of *abstract elements of style* serves to correlate the representation of primitive shapes to abstract elements of style. Abstract elements of style capture various stylistic effects due to position, balance and dominance (quantifiable properties of sentences with qualitative effects). In the last step, the grammar of *stylistic goals* maps combinations of abstract elements of style to corresponding stylistic goals, such as clarity, obscurity, abstractness, concreteness, staticness, or dynamism.

The other grammar-based approach to assessing stylistic goals was implemented by Ryan et al. (1992). The stylistic goals of this grammar (with settings in brackets) are as follows: *emphasis* (emphatic, neutral, flat); *clarity* (clear, neutral, obscure); and *dynamism* (dynamic, neutral, static). In this grammar, the basis for evaluating these goals was semantic, rather than the syntactic basis used by STYLISTIQUE. The particular semantic information used was the pattern of focus among the sentences in a paragraph. In an abstraction similar to STYLISTIQUE's, a set of grammar rules was used to specify the correspondence between the various possible patterns of focus and the defined stylistic goals (again, with the intermediary level, the abstract elements of style).

Both of these applications use the relationship between indicators and stylistic goals as the basis for assessment. The difference between these applications and the flag-based applications is that the relationship is much more sophisticated and the identified stylistic goals have a basis in stylistic theory. They do have some similarities, however. First, only intra-sentence indicators are used. Second, the relationship between these sentence-level quantifiable properties and the stylistic qualities (whether stylistic goals or properties) has not been empirically verified.

3.5 A Construct/Indicator Model of Stylistic Assessment

In sections 3.3 and 3.4, many views of style were described, both theoretical and practical. Figure 3.4 shows a construct/indicators model that provides a framework for the information upon which stylistic assessments are based and for describing the qualitative aspects of style. The components of the model and their names are drawn from the research field of experimental design (and are defined analogously). These components will be described in the following sections.

3.5.1 Computational Detection

There were many differences among the various applications described in section 3.4, but all of them used, as the basis of the stylistic assessment, a relationship between quantifiable clues in the text and qualitative, stylistic abstractions such as stylistic goals, stylistic problems, or stylistic qualities. For some applications, this relationship was simple and

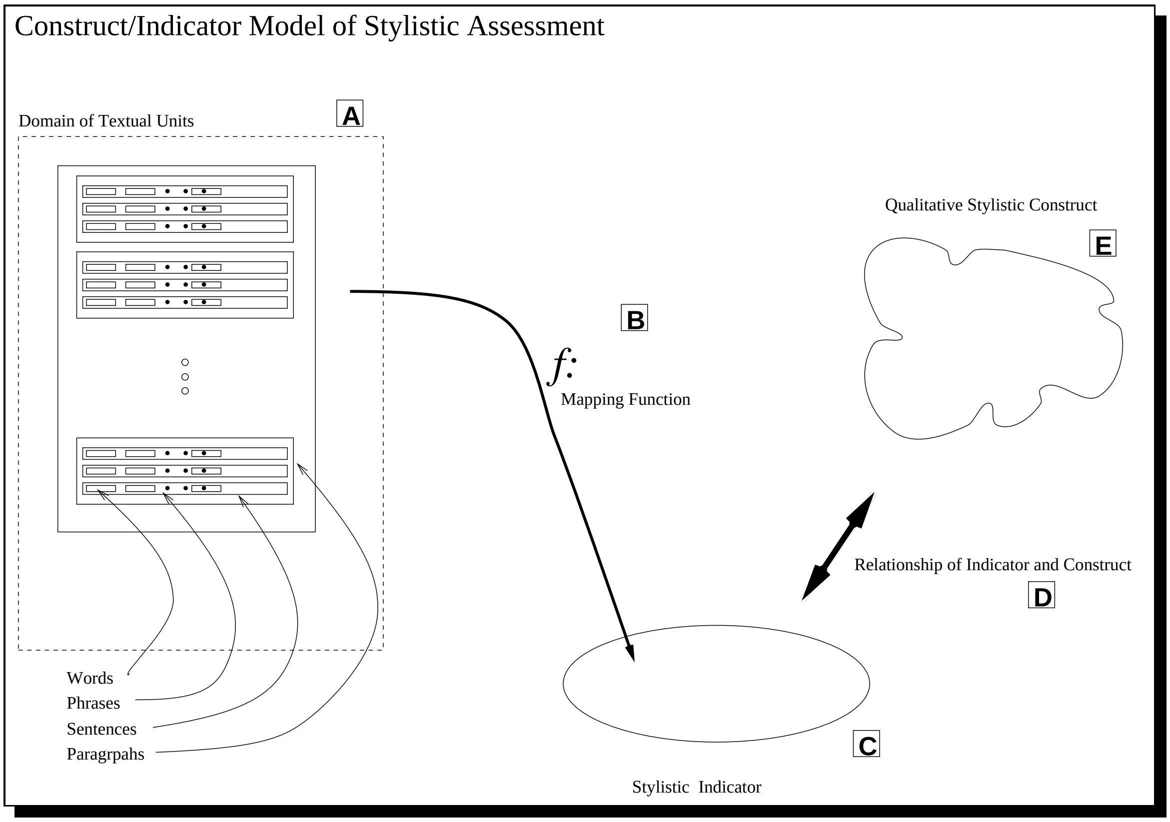


Figure 3.4: Stylistic indicators and stylistic constructs.

straightforward (e.g., flags) and for others, it was detailed and specific (e.g., grammars). This relationship (labelled D in the figure) is shown in the model as the connection between two components, the *stylistic indicator* (labelled C in the figure) and the *stylistic construct* (labelled E in the figure).

3.5.2 Stylistic Constructs

There are many different terms that are used to describe the outcome of a stylistic assessment, such as formality, partiality, or clarity, but these terms are all labels for qualitative concepts. A stylistic construct (labelled E in the figure) is an abstraction of some quality of the text's communication, invented in order to capture some aspect of a text's style. Since stylistic constructs are qualitative, they cannot be measured directly. Rather, aspects of a stylistic construct are measured indirectly via a stylistic indicator.

3.5.3 Stylistic Indicators

An indicator (labelled C in the figure) is some quantifiable aspect of the text that is thought to correspond in a meaningful way to a stylistic construct. For example, an established indicator of the formality of a sentence is the etymology of its content words. The value of the indicator, which can be either Latinate or Germanic, was shown by Levin and Novak (1991) to correspond to the formality of a sentence in a systematic way, and this was verified using empirical studies. As additional justification, Levin and Novak gave a historical explanation of this indicator. This argument lends support to the validity of the indicator.

3.5.4 Validity

Exploiting the relationship of the stylistic indicator to the stylistic construct provides a means of performing stylistic assessments, but this relationship must be based on a systematic correlation. The quality of this relationship is described in terms of its *validity*. A valid indicator is one which provides a reliable measurement of the intended stylistic construct.

Care must be taken to ensure that the indicator is measuring what it is expected to, and not some different but related construct. For example, sentence and word length has always been considered a reliable indicator of text understandability, forming the basis of the Flesch Reading Ease formula (Baker, 1988). The premise is that the longer a sentence and the more words in it, the more complex it must be, and the more complex a sentence is, the less understandable it is. However, the white space on a page of printed text (a construct that has a simple indicator) correlates in systematic way with the length of the sentences on the page. Smith and McCombs (1971) demonstrated by varying the amount of white space on a page and holding fixed the sentence lengths of the text on the page, that white space rather than sentence length is an indicator of text understandability. Thus, sentence length is an indicator of white space, rather than text understandability, although white space and text understandability are also related.

In Figure 3.4, the relationship between indicator and construct is shown as a two-directional line (labelled D in the figure). The quantifiable aspect of the text serving as the indicator may be the actual cause of the perceived stylistic construct or an associated side-effect. Therefore, stylistic indicators can be causes or effects of stylistic constructs.

3.5.5 Computability of the Stylistic Indicator

In section 3.4, some of the writer's aids described attempt to make use of the relationship between the construct of stylistic clarity and an indicator based on syntactic complexity. Even though a value for this indicator exists for every sentence, its calculation is difficult task for a computational process. The lack of an adequate parser means that such stylistic indicators cannot be measured readily. That is, the mapping from the sentences to the syntactic parse could not be performed accurately. In Figure 3.4, the relationship between the text (labelled A) and a stylistic indicator (labelled C) is shown as a mapping function (labelled B). So, for a successful stylistic assessment, not only are relevant indicators and constructs needed, but the mapping function must be of reasonable time complexity so that it computable (computable in the practical sense, not the mathematical sense). Note that the text is represented as a set of units, all of which can serve as the domain of this function (e.g., words, clauses, sentences, paragraphs, etc.). The domain for existing stylistic indicators has been based on words (e.g., the indicator of etymology for formality), on sentences (e.g., grammars used as indicators of stylistic goals), and on paragraphs (e.g., patterns of focus serving as indicators of stylistic goals).

3.5.6 Discussion

This model has several advantages:

- Separating the quantifiable aspects of texts from the qualitative aspects of style;
- Representing the relationship between text and style and identifying the property of validity as a criterion; and
- Identifying computability as a criterion for stylistic indicators.

However, there are important issues which cannot be addressed by this model, since they are outside its scope:

- Determining which constructs are relevant for useful stylistic assessments;
- Determining which indicators are valid;
- Determining the most economical indicators (balancing the tradeoff between computability and precision);

- Determining if there exists some stylistic constructs that defy measurement based on this model;
- Determining the role of subjectivity in the relationship between the indicator and the construct; and
- Understanding the relationship between constructs and the impact on their respective indicators.

Using this construct/indicator model, the various methods of performing stylistic assessment can be categorized. Flag-based stylistic assessments used relatively simple construct/indicator relationships. The constructs are often poorly articulated and often are given in terms of their indicators. Additionally, the only justification for the validity of the relationship is derived from style manuals. Heuristic-based stylistic assessments use a more sophisticated construct/indicator relationship. Additionally, the heuristic rules are used to identify the relevant constructs with which to evaluate a text, given a specification of the text's context. Heuristic-based applications also have more-clearly articulated stylistic constructs (such as the rhetorical goals of style), which are clearly separated from the indicators. Grammar-based applications are similar to flag-based applications, although the constructs are abstracted away from the indicators and the indicators are given in rigorously defined terms, which are expensive to compute. In spite of the attention to the rigorous specification of the indicator, there is little attention given to the validity of the relationship of the indicator to the construct.

3.6 Existing Stylistic Constructs and Stylistic Indicators

The construct/indicator model of stylistic assessment can be used to explain the stylistic assessments performed by existing software applications, but it also serves to provide a framework with which to consider the large body of existing research on stylostatistics. In this section, an inventory of stylistic constructs and stylistic indicators is presented.

Using the meta-language defined for this model, the contributions of the various stylostatistical research papers can be characterized as one or more of the following:

- an effort to define a stylistic construct;

- an effort to define a stylistic indicator (which may include an implementation); and
- an effort to establish the validity of the relationship between particular indicators and constructs.

Most research falls into the second category. Much of the research implicitly assumes the stylistic construct of author fingerprint and surprisingly little attention is paid to the validity of indicators, although there are some exceptions (such as (Snelgrove, 1990), (Dale, 1977)).

Even though the stylistic construct of fingerprint is commonly assumed, an inventory is still useful, since the stylistic indicators often can be reused. In the remaining sections, the following constructs are considered:

- Abstractness/Concreteness
- Archaicness/Trendiness
- Authorial Fingerprint
- Clarity/Obscurity
- Colour
- Emotional Tone
- Floridity
- Force
- Staticness/Dynamism
- Understandability
- Vocabulary Richness
- Word Difficulty

There were great differences in the completeness of analysis for each of these constructs. Many constructs had intuitive definitions, while others were carefully defined. For each construct, some had corresponding indicators carefully defined, while other indicators were merely conceptualizations. Overall, there was a surprising lack of regard for validity. The following questions were asked for each construct:

- What does this construct mean? How is it defined? Is it described intuitively or precisely? Do people agree on the meaning of this construct?
- How is this construct measured? What are its indicators? Has the validity of this indicator been explored? Can the indicator be measured computationally?

Abstractness/Concreteness

The constructs of abstractness and concreteness have been measured in text by several different researchers. These constructs were chosen by DiMarco (1990), following Vinay and Darbelnet (1958), to represent the opposite ends of one of three dimensions that were selected to capture the style of text, independent of its language (p. 50). Concreteness is associated with sentences "that express an effect of immediacy by emphasizing a particular component" (p. 133) and abstractness is associated with sentences in which there is a general lack of modification. The construct of concreteness, as well as meaningfulness, was studied by Pavio et al. (1968) as a property of nouns. For Pavio et al., the construct of imagery was postulated as the most relevant psychological attribute underlying the linguistic abstractness-concreteness dimension (p. 2). The correlation between imagery and the abstractness-concreteness dichotomy was proven in a series of experiments (see (Pavio et al., 1968)).

By using the stylistic grammar developed by DiMarco (1990), indicators of abstractness and concreteness can be defined on the basis of syntactic features of text. In addition to syntax, there are a number of lexically-based indicators. Benjafield and Muckenheim conducted experiments to develop lists of subjective assessments of proverbs (Benjafield et al., 1993) and words (Benjafield and Muckenheim, 1989) with respect to the constructs of imagery, concreteness, goodness, and familiarity. Similar trials were conducted by Pavio et al. (1968), where the interconnections of the constructs were also explored. In a large experiment using human judges and subjected to factor analysis, Carroll (1960) found several objective measures (indicators) of a stylistic construct that was labelled abstractness, including the proportion of noun clauses, determining adjectives, and pronouns.

Archaicness/Trendiness

Lexicographers have devised several approaches to the problem of labeling 'older' word usage. This is one of several style values for words that are included in dictionaries (Hartmann, 1981). This stylistic feature was identified as an important consideration in the task of lexicalization in natural language generation (Stede, 1993). The stylistic effects from these words extend to text, influencing the style of the entire text. Stede (1993) points out that old words can be exhumed to achieve specific effects, for example by calling the pharmacist an 'apothecary'. Conversely, using terms that have been recently coined gives the impression of trendiness. Stede points out that this stylistic dimension also holds for non-content words (p. 456).

Stede describes a natural language generation system in which the indicators of these constructs at the lexical level are used to create the overall style of a text. The canonical indicator of this construct is to directly use a value that has been assigned to a particular word. Dictionaries in general are full of such assignments. However, using these indicators is fraught with shortcomings. First, there is a lack of common labelling practice, so the schemes are inconsistent (Hartmann, 1981). This means that several sources of information are not readily integrated, including thesauri and other lexical resources (e.g., WordNet). Second, these values reflect the subjective opinion of the individual lexicographer. Third, as Stede recognizes, these values must also be validated.

Authorial Fingerprint

This construct is used to describe the quality of a text that serves as a unique identifier of the author. This quality captures the essence of an author's personality. This construct relies on the metaphor that style serves as a fingerprint for the author and so is the foundation of authorship attribution studies. "Stylometry attempts to capture quantitatively the essence of an individual's use of language" (Dale, 1977).

Several different types of statistical tests serve as indicators of this construct. The development of these indicators has a long history (e.g., basing the indicators on word and sentence length originated in the late 1800's (Smith, 1983)). The pursuit of developing these indicators thrived in the 1970's, obviously due to the availability of computers. A typical stylometric approach is EYEBALL (Ross Jr. and Rasche, 1972), a subcomponent of the larger CALAS (Computer Assisted Language Analysis System) (Gervasio et al., 1986). The system performs stylistic analysis based on eight ratio measures, which in turn are based on verb, phrase, and clause categorizations. A critical study by Smith (1983) of stylostatistical measures found that word-based frequency counts failed frequently to accurately correlate with authorship. The sentence length–based measures are endorsed only as confirmatory measures; no conclusions based exclusively on them can be made. In any case, the author's style often varies over their career. Additionally, the measures must be used on texts that are the same literary type.

Other, more sophisticated indicators of author fingerprint have been developed, such as neural nets (Matthews and Merriam, 1993) and multivariate analysis (Ledger, 1985).

Clarity/Obscurity

The construct of clarity is a common one in style guides and writing handbooks. In fact, writing text with this quality is implicitly assumed to be the author's only desire. This assumption has some merit, since one does hope that the desired communication is clearly conveyed. This construct is related to understandability. On the other hand, there are instances, however, when an author aims to achieve obscurity. The constructs of clarity and obscurity were chosen by DiMarco (1990) to represent the opposite ends of one of three dimensions which were selected to capture the style of text, independent of its language (p. 50).

Although the style handbooks and writing handbooks describe how to achieve clarity (e.g., Strunk's (1979) advice to 'omit needless words'), these rules cannot be used to determine whether clarity has been achieved. In contrast, DiMarco's stylistic grammar describes precisely, in syntactic terms, when clarity has been achieved (DiMarco, 1990), (DiMarco and Hirst, 1993). This codification serves as a precise indicator; however, its validity has not been studied.

Colour

Hovy (1988) describes the stylistic construct of colour. A colourful text is one in which the author adds references to personal experience. These references include examples, idioms, and statements of personal evaluation. Colourful text is personalized, which helps achieve an atmosphere of informality and reduces the interpersonal distance between the author and the audience.

Hovy describes some heuristics for generating text with colour. These heuristics describe features in text, such as: the inclusion, in addition to the topic of the text, of instances similar to the topic; the use of idioms instead of general statements; the inclusion of sentences describing personal evaluations; the making of adjectival clauses of instances; and the use of metaphoric and idiomatic phrases and words.

Emotional Tone

In addition to evoking conceptual processes, literature also evokes an emotional response in the reader. The quality of emotion expressed in a text is a construct that Anderson and McMaster have subjected to computational analysis (1982), (1986). In order to define this construct, they use dimensions of emotional tone established in psychology: pleasure, arousal, and dominance. The psychological theory argues that these three dimensions are necessary and sufficient to describe any emotional state. Anderson and McMaster were interested in using emotional tone to chart the ebb and flow of emotional tension in literature. Although emotional effect is not the goal of the text in our targeted domain, it is still worthwhile to consider how a quantitative indicator was developed for this subjective construct. It is also a good example of the importance of establishing indicator validity through experimentation.

As an indicator of the three dimensions of emotional tone in a passage, the values of the Heise words in a text are used. The Heise words are special words in a dictionary that have been assigned weightings in each of the dimensions, called connotative meaning scores. As an indicator of the more specific construct of emotional tension in a sentence, Anderson and McMaster (1982), (1986) developed an formula based on an aggregate of the word scores. The validity of these indicators was verified in an experiment. The emotional tension score was also shown as being able to reveal a "gripping episode" or emotional catastrophe in a text. Although the indicator of emotional tension produces a value for each sentence, the information is most useful when plotted over the sequence of sentences in a text. The emotional tone scores were also be presented in a chart of emotional state transitions.

Floridity

This construct was developed by Hovy (1988) to capture the flowery quality of some styles. Although this construct is understood intuitively, it is poorly defined. Hovy defines florid style in a text as using unusual words. Stede (1993) suggests that floridity is used to sound sophisticated, which is different from formality.

So far, the only indicators developed have been based on values associated with flowery words (Hovy, 1988), (Stede, 1993). Therefore, detecting words that are marked as florid serves as an indicator of florid style. This indicator is not particularly useful, since no resources exist to correlate this quality with words, nor has the validity of this type of indicator been established.

Formality

The quality of formality captures the amount of interpersonal distance between the author and the audience that is conveyed in a text, where the appropriate amount of interpersonal distance depends on the social setting. Therefore, the definition of formality is always relative to audience, but generalizations can be made. For example, an informal or colloquial style conveys a closeness between the author and the audience. Because of this closeness, it is acceptable to use slang as well as personal phrases and words. A formal text conveys a distance between the author and the audience. Hovy uses this stylistic construct prominently, which is partly based on research in politeness (see (Brown and Levinson, 1988)). He uses it to achieve his pragmatic goals, using the assertion that in language, formality is one of the strongest carriers of non-literal information we use (Hovy, 1990).

Hovy (1988) uses several heuristics to generate text with more or less formality, as required. To create formality in text, he uses the heuristics such as the following (pp. 82–88): create long sentences, especially those with causal, temporal, or other relations to other sentence topics; use many adverbial clauses, placed towards the beginning of sentences; use passive voice; use complex verb tenses; and avoid ellipsis. Therefore, detecting these features can serve as indicators of formality, but the validity of these indicators needs to be established.

One indicator of formality has been established as valid. Levin and Novak (1991) found that the etymology of lexemes in sentences is a reliable indicator of the sentence's perceived formality. Latinate sentences are considered more formal than Germanic sentences. Sentences using low-frequency Germanic words are perceived as more formal that those with high-frequency Germanic sentences, although this frequency effect was not observed for Latinate sentences.

Force

Hovy (1988) uses the construct of force to describe a text that has a direct, straightforward style that carries momentum. This construct is also related to the stylistic construct of timidity, which is the unwillingness to include personal opinions. Hovy lists a number of heuristics that can be used to create forceful text. For example, use short, simple sentences; use the active voice; use simple, plain words and phrases rather than flowery or unusual ones. Thus, an indicator of forceful style can be short, simple sentences in the active voice. As well, text that uses flowery language is likely not to be forceful.

Staticness/Dynamism

DiMarco (1990), following Vinay and Darbelnet (1958), describes French as tending to be more static than English, as there is a predominance of the noun over the verb. The constructs of staticness and dynamism, albeit intuitively defined, capture the movement or action of the style of a text. Staticness is associated with adherence to the standard and order; dynamism is associated with invigoration and deviation from the norm. These constructs were chosen by DiMarco (1990) to represent the opposite ends of one of three dimensions which were selected to capture the style of text, independent of its language (p. 50).

DiMarco's stylistic grammar describes precisely, in syntactic terms, when a text is static or dynamic (DiMarco, 1990), (DiMarco and Hirst, 1993). This codification serves as a precise indicator; however, its validity has not been studied.

Understandability

The construct of understandability is of interest for several areas, such as pedagogy, as well as stylistic analysis. This understandability of a text means its comprehensibility, rather than its readability. This construct is relevant to style, since the style of a text can be used to achieve clarity or to achieve obscurity. Therefore, this construct, while not the same as clarity or obscurity, is related. Another related construct is text complexity and simplicity. The stylistic construct of simplicity was also described by Hovy (1988).

The most common indicator used to measure understandability is a reading formula. A reading formula is a regression equation used to predict the level of comprehension by the audience of a text, on the basis of predictor variables and a comprehension measure. The predictor variables are chosen so that they are causally related to comprehension. These formulas (there are over 100 in use today) rely on a model of reading as passive decoding (Baker et al., 1988). Therefore, the predictor variables capture lexical and syntactic characteristics. The biggest shortcoming of readability formula is their dependence on the simple decoding model. Meaning, however, is not found just in the text, but is generated from the text representation and world knowledge. Readability formulas have been shown to correlate weakly or even negatively with human assessment of readability (Bruce et al., 1981), (Selzer, 1988). In spite of this, readability formulas are still widely used.

Vocabulary Richness

The stylistic construct of richness of vocabulary is thought to characterize an author's style (Baker, 1988), since it is a reflection of the author's vocabulary. The PACE measure, developed by Baker (1988), was designed to be an indicator of vocabulary richness. This indicator is used to measure an author's ability to select new words as the length of a text increases. Baker (1988) claims that this stylistic construct correlated strongly with the authorship of Shakespeare and Marlowe. Additionally, several other measures of vocabulary richness have been developed. For instance, statistical measures, based on the size and number of words used only once, were developed by Honoré (1979). Other statistics include factors such as: vocabulary overlap between text segments, introduction of new words , and the closeness between two texts in terms of the number of steps to change one to the other (Ule, 1982). Since this construct can be cheaply measured using statistical techniques, it should be evaluated for relevance to stylistic incongruity.

Word Difficulty

The construct of word difficulty is related to, but different from, text understandability. The difficulty of a word is defined with respect to intelligence testing, but correlated with word frequency, and hence familiarity. This construct has been greatly studied in psychology.

Since differences in word difficulty among text segments can easily be perceived by readers, the role of this construct in stylistic incongruity should be explored.

The common perception is that frequency corresponds to familiarity and less familiar words are more difficult than more familiar words. Since word frequency in text doesn't always correspond exactly to familiarity, the correlation between frequency and difficulty must be verified as reliable. An experiment to demonstrate this was successfully performed by Rudell (1993). Therefore, for the Kučera and Francis words (a set of words with their corresponding frequencies taken from a 10^6 word corpus), the frequency measures can serve as a reliable measure of word difficulty.

Chapter 4

Audience Agreement on Stylistic Assessment

4.1 Introduction

In this chapter, the issue of the agreement within an audience, in terms of their subjective assessment of style, is discussed within the context of building a computational tool to assist writers, especially those writing collaboratively, to eliminate stylistic incongruities. To measure audience agreement and investigate some factors that might govern the assessment of style, an exploratory study was conducted. The results of this study are that authorship does not necessarily coincide with assessments of stylistic similarity, that there is a significant amount of inter-subject agreement, and that sentence count is not a good predictor of stylistic similarity. This chapter also discusses the procedure of using Monte Carlo simulations to assess the significance of the results and the validity of the similarity measures used.

4.1.1 Exploratory Study Design

Task Selection As a starting point, we assumed that the detection of stylistic incongruities is a sub-skill of the more general skill of stylistic awareness. Overall stylistic awareness is more foundational, since the detection of stylistic incongruities can't presumably be done without this meta-skill. To discover each subject's stylistic assessment, we wanted to give them the greatest possible amount of latitude in judgement. We did not want to evaluate the assessments against an *a priori* assessment, since the existence of such an *a priori* standard itself is being investigated.[1] For this reason, the task given to a group of subjects was a free-sort[2] of writing samples, where the set of writing samples was controlled as much as possible such that only the style of the samples varied (e.g., they did not have great semantic differences, or the semantic differences would not be detectable by the subjects). The premise was that the resulting sorting arrangement would be a reflection of a subject's stylistic assessments.

Goals The first goal of this study was to develop a way to make comparisons between the subject's stylistic assessments. This measure must reflect the degree of agreement or disagreement between any two subjects. Furthermore, the measure must describe the degree of agreement or disagreement among multiple subjects.

The second goal was to discover the extent of the influence of authorship in subject's stylistic opinions. Even without looking at the overall agreement among subjects, we first wanted to see if a subject's decision to place a group of writing samples together (thereby judging them to be stylistically similar) was related at all to whether the writing samples had the same author. An assumption used in a previous study was that samples of writing by the same author (and taken from the same text) should be more stylistically similar to each other than to writing samples with different authors. We suspected that perhaps subjects would not agree with each other in terms of exactly which writing samples are similar, but for the samples they did choose as similar, these samples would largely be by the same author. Additionally, we wanted to see if this would also be true of writing samples by the same author, but taken from different texts.

A third goal was to determine the degree of agreement, if any, between all the subjects' stylistic assessments. We realize that there might be different strategies for sorting the writing samples (perhaps due to the subjectivity of stylistic assessments), and that there might be a high level of agreement within each strategy. If these separate strategies do exist, lumping all the subjects together would blur the degree of agreement; however, this is an exploratory study, and as a first step, we start with the most general case. More sophisticated models of the audience can be explored in the future.

[1] Such a *a priori* standard is that the style of a text is perceived in the same way by different readers.

[2] A free-sort is a task in which the subjects are instructed to sort a set of items into piles according to some criterion. The criterion for our task was *stylistic similarity*. This is explained in more detail in Section 4.2.3.

In the case of a high level of agreement among the subjects, another question to investigate is whether sentence count as a stylistic indicator is a possible explanation for the sorting assessments done by the subjects. Although this is a trivial measure, subjects, for lack of a better strategy, may place writing samples of the same size (number of sentences) together. For example, the subjects may place all the short writing samples together.

4.2 Experiment

Before this experiment was conducted, a pilot study was conducted to determine the length of time required to complete the experiment. Three students participated in the pilot study, including native and non-native speakers of English. Due to the length of time that subjects required to complete the study, the testing materials were altered to contain shorter writing samples. This pilot study confirmed the expectation that the writing samples were sufficiently difficult, in terms of the required background domain knowledge, to make the semantic content of the samples more or less opaque. We didn't want the sorting to be based on the semantic content of the samples.

4.2.1 Subjects

Subjects ($n = 11$) were solicited by e-mail within the University of Toronto computer science graduate student community. Their participation was voluntary. They were told that the experiment involved sorting a set of writing samples according to their assessment of the samples' *writing style*. The participants all were native speakers of English and were either graduate students or holders of graduate degrees. The participants were pleased to participate and curious about the experimental results.

An interesting research area for future exploration of the differences in the perception of style in native speakers of a language compared to non-native speakers of a language. Since determining native vs. non-native effects were not desired for this experiment, only native speakers of English were permitted to participate.

4.2.2 Materials

One set of materials was prepared, containing 24 writing samples. The 24 writing samples consisted of 8 subgroups of three samples each. For each subgroup, the three writing samples each consisted of a single paragraph extracted from an academic paper on the philosophy of the mind/body problem. The paragraphs were chosen so that they did not contain any glaring out-of-paragraph references or contextual references to the original paper's overall discourse structure. Each of the chosen paragraphs consisted of between 1 and 8 sentences. The 8 papers from which the samples were chosen were selected so that the subject matter would be sufficiently opaque to a "lay" reader (i.e., a reader not familiar with philosophical writings on this subject). The hope was that semantic clues could not be easily found to assist in the sorting procedure. The set of papers was also chosen to represent the writing styles of the following 7 authors: Thomas Nagel, Sydney Shoemaker, Ned Block, Frank Jackson, Collin McGinn, Jerry Fodor (3 paragraphs from each of two different papers), and Michael Posner.

4.2.3 Procedure

The subjects were each given a stack of small slips of paper, each slip containing a writing sample. They were instructed to sort the writing samples into piles so that each pile contained a different writing style. The subjects were told to use their own intuitive sense of writing style. They were reassured that any permutation, ranging from one pile of 24 samples to 24 piles, each containing one writing sample, was acceptable.[3] The subjects were allowed to take up to an hour to complete the free-sort.

We give the name *sorting arrangement* to the resulting configuration (the number of piles and the size and content of each pile). Since we do not care about the specific order in which the piles were made, nor about the order in which the writing samples composing a particular pile were added, it is equivalent to talk either about a subject's particular *sorting arrangement* or a subject's *partitioning* of the given writing samples.

4.3 Statistical Analysis of Data

We are now in the position to answer the main questions of this exploratory study:

- What is a valid measure of stylistic agreement between subjects?

[3] The likelihood of a subject producing either of the extremes is extremely unlikely; rather, these details were given to help the subjects understand the range of the space of possibilities. Additionally, it was desired to assure the subjects that a pile containing a single, stand-out writing sample would also be acceptable without overtly suggesting such an arrangement.

- Do readers perceive texts by the same author as stylistically more similar than texts by different authors?
- Do the stylistic assessments of our reader audience (group of subjects) show a high degree of agreement?
- Can a simple indicator, such as sentence length, be used to explain the strategy used by the subjects to produce their sorting arrangements?

4.3.1 Preparation of Data

Each free-sort (or sorting arrangement) of writing samples into piles corresponds to a partition of a set S of 24 unique elements, where a partition is a set of non-empty subsets of S, called *cells*, whose union is the set S. The set of all possible sorting arrangements corresponds exactly to the set of all partitions of 24 unique elements. The term *cells* of a partition is interchangeable with the term *piles* of a sorting arrangement. For our data, each subject's partition was represented by matrices and alternatively, by vectors. More specifically, a 24×24 $(0,1)$-incidence matrix was created such that the (i,j) element was set to 1 if and only if writing samples i and j were placed in the same pile by the subject. Otherwise, element (i,j) was zero. Additionally, the sorting arrangement was represented by a $(0,1)$-vector with $\binom{24}{2} = 276$ elements. Each element corresponds to an (i,j) pair of writing samples and is assigned the value of 1 if and only if the writing samples i and j have been placed in the same pile. These representations are equivalent; the version used was a matter of requirement (e.g., for the computational Monte Carlo simulations, the less redundant vector representation was implemented; for illustration in this thesis, the more visual and intuitive matrix representation was used). Additionally, representations were created to correspond to the sorting arrangement of 6 piles of 3 writing samples each and 1 pile of 6 writing samples, which corresponds to the actual authorship of the writing samples. These representations are labeled the *authorship matrix* or the *authorship vector* respectively.

4.3.2 Measuring Stylistic Agreement

The agreement between the stylistic assessments, as reflected in two sorting arrangements, should be determined by degree of similarity to each other. The measure of similarity that was used in this study was a *distance statistic*. The smaller the distance between two sorting assessments, the more similar the subjective opinions. A distance of zero between two assessments implies that they are identical and the respective subjects share the same stylistic perceptions about the group of writing samples that they sorted. The remainder of this section discusses the selection of 3 distance statistics.

Most existing distance statistics would depend on the categorization of the $\binom{24}{2}$ writing sample pairs in terms of the two partitions $\mathcal{A}$ and $\mathcal{B}$ being compared (Hubert and Levin, 1976). The categorization of the set of writing pairs can be made using the following terms:

1. m_1 writing sample pairs that are placed together both in $\mathcal{A}$ and $\mathcal{B}$.
2. m_2 writing sample pairs that are not placed together either in $\mathcal{A}$ and $\mathcal{B}$.
3. m_3 writing sample pairs that are placed together in $\mathcal{A}$ and apart in $\mathcal{B}$.
4. m_4 writing sample pairs that are placed apart in $\mathcal{A}$ and together in $\mathcal{B}$.

Many linear combinations of m_1, m_2, m_3, and m_4, normalized or unnormalized, are proposed as measures of homogeneity[4]; however, the single quantity m_1 determines m_2, m_3 and m_4 (Hubert and Levin, 1976).

In a past experiment, the *gamma measure of proximity* (Γ) was used to measure the similarity of subject's sorting arrangements to a hypothetical norm (Teshiba and Chignell, 1988). This gamma measure of proximity is well established in mathematical psychology as a measurement of similarity between sorting arrangements (Hubert and Levin, 1976). The major change in this study from the previous experiment is that only our subject's final arrangements were considered and not the process by which they determined the final arrangement. In the past experiment, each subject produced a hierarchy of elements. However, with respect to the measurement of distance, this difference is unimportant as the final sorting arrangement of each subject in this study can be thought of as a flat, single-level hierarchy.

The Γ measure of proximity is given as the cross-product term from the numerator of the Pearson product-moment (Hubert and Levin, 1976), (Hubert, 1978):

$$\Gamma = \sum_{i,j} q(o_i, o_j)c(o_i, o_j). \tag{4.1}$$

The elements from the i^{th} row and j^{th} column of the two *proximity matrices* to be

[4]There are many provided in (Hubert and Levin, 1976), but will be omitted here for brevity.

compared (in our case, the $(0,1)$-incidence matrices that we previously defined) correspond to the elements $q(o_i, o_j)$ and $c(o_i, o_j)$. Additionally, one can see that the Γ measure calculates m_1 for $(0,1)$-matrices (occasionally called the "Matching Coefficient", e.g., see (Borgatti et al., 1992)). Since this distance statistic is already established and in use, it was used in our study.

Since the Γ measure is the numerator of the Pearson coefficient[5], it stands to reason that typical measures of distance between vectors could also serve as distance measures. The $(0,1)$-incidence matrix can easily be transformed into a $(0,1)$-vector. For this reason, the Pearson coefficient (ρ), a measure of linear correlation between vectors, was used as well. We also tried using Euclidean distance (Δ) as a distance statistic, since it can be a spatial measure between vectors.

We decided to use each of the three distance statistics in this study — the Pearson correlation (ρ), the Gamma measure (Γ), and Euclidean distance (Δ) — in order to analyze the data in different ways.

Confounding Within the Δ Measure

After completing the analysis, we noticed some strange findings produced by the Δ-based calculations. Upon investigation, it was discovered that the Δ-based distance statistic was being affected in a systematic way by an unrelated side-effect, resulting from the way in which our data was being represented.

The Δ measure, being based on Euclidean distance between $(0,1)$-vectors, is more a measure of the differing "bits". The vectors representing our subject data, in general, are very sparse, with very few 1's. We note that the number of 1's in a $(0,1)$-vector grows exponentially with the cardinality of the piles in the corresponding sorting arrangement (a pile of size k adds $k!$ 1's). Therefore, partitions with cells of large size will have substantially more 1's in their (0,1)-vector representation than partitions with cells of smaller size (in the worst case, all cells will have the same size). Therefore, there is a substantial amount of confounding in this measure. The results of our Δ-based calculations are included here as illustration however, but because of this confounding, it would be unwise to draw conclusions from them.

[5]This is how the mathematical psychology research community seems to refer to *linear correlation*.

Measuring Very Similar Sorting Arrangements

The sorting arrangements that our subjects made were not very similar to each other. However, we wanted to examine the behaviour of our distance statistics when measuring the similarity between sorting arrangements that are very close to each other. In the future, we may conduct studies that produce such data and may need valid measures.

The closest distinct pair of partitions possible in our sorting task would reflect the judgements of subjects who each placed the writing samples in essentially the same piles. The only difference would be that one subject made a finer distinction between the styles of the writing samples in one pile and further divided that pile into two smaller ones. This is a less serious difference of opinion than the difference shown between two subjects who agreed on the arrangement of all the writing samples, save one writing sample which has been placed in a different pile. This is a more serious difference because it indicates a disagreement, while the first difference indicates a refinement, but not an overt disagreement. But these cases of disagreement are less serious again than the difference between two partitions in which two elements have been swapped between cells. Overall, the latter two scenarios — a difference of a *move* and a *swap*, respectively — represent more serious differences because the subjects show a difference of opinion, where as the first scenario — a *divide* — while not showing total agreement, doesn't necessarily show a disagreement either.

The distance statistic should correspond to the degree of disagreement as well as agreement. Since disagreement might be given in terms of *move*, *swap*, and *divide* operations, an ideal measure would count the minimum number of these operations separating two partitions. Such a statistic requires careful definition, since these operations can be expressed in terms of each other and alone are not adequate to express the difference between any two partitions (e.g., new cells may be required). While such a statistic would have the advantage of giving a meaningful measurement, it unfortunately does not exist in a computationally feasible form. In fact, the analogous problem in graph theory[6] is an open research problem. Even if a heuristic could be devised which approximates this distance statistic, it may not be valid; it could not guarantee a minimal number of operations and it may produce values with the property that $d(A, B) \neq d(B, A)$.

[6]Each partition can be represented as a graph of cliques (complete graphs), where the nodes represent a writing sample and an edge indicates that the two writing samples have been placed in the same pile. The analogous problem in graph theory is, given graph G_1 and G_2, how many steps are required to transform graph G_1 into graph G_2?

To see if the Γ, ρ, and Δ distance statistics captured these finely grained differences, a set of partitions was constructed. This set of partitions represents a set of sorting arrangements that are similar, but differing from each other to different degrees. We wanted to see if the Γ, ρ, and Δ distance statistics could capture the varying degrees of agreement. In Figure 4.1, the 5 close, but different, partitions (labeled A, B, C, D, and E) of 8 writing samples (labeled $\alpha, \beta, \chi, \delta, \epsilon, \phi, \gamma$, and η) are shown, along with their corresponding $(0, 1)$-incidence matrices.

Partition A	**Partition B**	**Partition C**	**Partition D**	**Partition E**
αβχδεφ γη	αβχ δεφ γη	αβχδε φγη	αβχδεγ φη	αβχ δεγ φη
1 1 1 1 1 1 0 0	1 1 1 0 0 0 0 0	1 1 1 1 1 0 0 0	1 1 1 1 1 0 1 0	1 1 1 0 0 0 0 0
1 1 1 1 1 1 0 0	1 1 1 0 0 0 0 0	1 1 1 1 1 0 0 0	1 1 1 1 1 0 1 0	1 1 1 0 0 0 0 0
1 1 1 1 1 1 0 0	1 1 1 0 0 0 0 0	1 1 1 1 1 0 0 0	1 1 1 1 1 0 1 0	1 1 1 0 0 0 0 0
1 1 1 1 1 1 0 0	0 0 0 1 1 1 0 0	1 1 1 1 1 0 0 0	1 1 1 1 1 0 1 0	0 0 0 1 1 0 1 0
1 1 1 1 1 1 0 0	0 0 0 1 1 1 0 0	1 1 1 1 1 0 0 0	1 1 1 1 1 0 1 0	0 0 0 1 1 0 1 0
1 1 1 1 1 1 0 0	0 0 0 1 1 1 0 0	0 0 0 0 0 1 1 1	0 0 0 0 0 1 0 1	0 0 0 0 0 1 0 1
0 0 0 0 0 0 1 1	0 0 0 0 0 0 1 1	0 0 0 0 0 1 1 1	1 1 1 1 1 0 1 0	0 0 0 1 1 0 1 0
0 0 0 0 0 0 1 1	0 0 0 0 0 0 1 1	0 0 0 0 0 1 1 1	0 0 0 0 0 1 0 1	0 0 0 0 0 1 0 1
	divide	*move*	*swap*	*split and swap*

Figure 4.1: Sorting arrangements with differences dues to *divides, moves* and *swaps.*

In Table 4.2, the values for a some of the inter-partition distances are given, as measured by each of the three distance statistics. As a measure of performance, we considered how each distance statistic ranked the similarities of the partitions A, B, C, D, and E.

Distance	**Distance Statistics**				**Ranking of Similarity**
Statistic (d)	$d(A,B)$	$d(A,C)$	$d(A,D)$	$d(A,E)$	
Ideal	1 divide	1 move	1 swap	1 divide + 1 swap	(A,B), (A,C), (A,D), (A,E)
ρ	.5000	.5168	.1250	0	(A,C), (A,B), (A,D), (A,E)
Γ	7	11	10	4	(A,C), (A,D), (A,B), (A,E)
Δ	3	$\sqrt{7}$	$\sqrt{2}$	$\sqrt{15}$	(A,D), (A,C), (A,B), (A,E)

Figure 4.2: Summary of distance statistics ρ, Γ and Δ.

Overall, each of the three distance statistics show that the similarity of two partitions differing by 1 operation (a divide, move or swap) is greater that the similarity of two partitions differing by 2 operations. This is a basic criterion for a valid distance statistic.

However, the divide, move, or swap operations result in different degrees of dissimilarity for each distance statistic. For instance, the ρ distance statistic finds a divide operation to cause more dissimilarity than a move operation. The Γ distance statistic orders the operations in ascending order of resulting dissimilarity as move, swap, and then divide. The Δ distance statistic orders the operations as swap, move, and then divide[7]. This is slightly different from what we would like, but it will not adversely affect our results unless the partitions happen to be extremely similar. This is not the case for our subject data, since the average ρ distance is 0.04.

To explain this, we note that the Γ and Δ distance statistics are both *discrepancy based*. Each statistic is derived as a function of the number of discrepant elements of the proximity matrices. But, as the $(0,1)$-incidence matrices in Figure 4.1 illustrate, the operations such as *divide, move* and *swap* cause a number of changes in the corresponding $(0,1)$-incidence matrices that is disproportionate to the seriousness of the disagreement that each of these operations represent. For example, a divide operation represents a difference in partitions where subjects don't actually disagree, but it adds the most discrepancies. In particular:

- A *divide* operation adds $m \times n$ discrepancies. In the (0,1)-matrix, 1's must be removed to show that elements are no longer in the same pile. The values of m and n are the cardinalities of the two smaller cells resulting from the divide.

- A *move* operation adds $m + n - 1$ discrepancies. The matrix must reflect the fact that an element is no longer in one cell and has been placed in a new cell. The value of m represents the size of the cell the element is joining and $n - 1$ represents the size of the cell that the element is leaving minus 1.

- A *swap* operation adds $2(m+n-2)$ discrepancies. The matrix must reflect the new position of two elements. Two cells are affected, say of size m and n, because they gain a new element and lose an old element. For each cell, changes are required to show it now contains its new element ($m - 1$ and $n - 1$ respectively, so $n + m - 2$ in total). The same number of changes are required to show it no longer contains the old elements, for a total of $2 \times (m + n - 2)$ discrepancies.

[7]The Δ measure actually increases as the similarity decreases, as opposed to the ρ and Γ measures which increase as similarity increases. This ordering was taken into consideration in the ranking.

From this analysis, a *divide* (or *split*) operation will add discrepancies at a faster *rate* than the discrepancies added by a *move* or *swap* operation. For judgements of stylistic similarity, we would like a divide or split to cause discrepancies at a slower rate than a move or swap. However, to reiterate, this would only be an important distinction for assessing partitions that are very similar, which is not the case for our data. The Δ distance statistic, in addition to the confounding discussed in Section 4.3.2, has the disadvantage of mirroring this property exactly (i.e., it is perfectly correlated with the number of discrepancies, while the Γ measure reflects this more indirectly). The ρ statistic is not a simple reflection of discrepancy and is not considered a discrepancy-based distance statistic.

Summary

In this section, we have examined three different distance statistics. The Δ distance statistic has problems with confounding. The Δ-based results are included in the reporting of our results, but only for illustration. No conclusions should be based on the Δ distance statistic. For very similar partitions, Γ and ρ may not reveal very fine distinctions, but this isn't an issue for our data. These measures serve as the basis for the following analysis.

4.3.3 Measuring Stylistic Agreement Within an Audience

Now that we have established ρ and Γ as measures of similarity between 2 subjects, we wish to extend the idea of similarity between 2 subjects to similarity among a group. Measurement of a group's agreement is a function of the distances between each of the subjects sorting arrangements, or the **inter-subject distances** (ISDs). For our study, there were $\binom{11}{2} = 55$ inter-subject distances upon which to base the measurement of agreement between the members of the audience sample. The mean of the ISDs was used as an indicator of overall agreement in a group.

Assessing the significance of the ISD values is not straightforward, however (for groups of 2 or more). The theoretical frequency distribution of these dependent variables, the inter-subject distances, was not known. It seemed best not to assume a normal distribution (keeping in check the often exuberant application of the Central Limit Theorem) due to the blatant lack of independence among the variables.

There is not even a procedure for evaluating the size of an ISD measure between two partitions. A method for assessing the significance of the Γ distance statistic (and potentially the others as well) has been proposed (Hubert and Levin, 1976) and is in use (e.g., available in the statistical analysis software UCINET (Borgatti et al., 1992)). The procedure is paraphrased as follows:

> One way to evaluate $\Gamma_{\mathcal{A},\mathcal{B}}$ (the distance statistic between the sorting arrangements of subjects $\mathcal{A}$ and $\mathcal{B}$, each represented as a 24×24 $(0,1)$-incidence matrix.) is to calculate a set of many $\Gamma(\mathcal{A},\mathcal{B}')$ values, where there are $n!$ $\mathcal{B}'$'s, derived by reordering the rows and columns of $\mathcal{B}$ simultaneously. All the distance statistics between $\mathcal{A}$ and each of the possible rearrangements of $\mathcal{B}$ are used to create a distribution of values. If the statistic $d_{\mathcal{A},\mathcal{B}'}$ is larger than $d_{\mathcal{A},\mathcal{B}}$ in only a few instances (say, less than 5% or 10% of the time), then $\mathcal{A}$ and $\mathcal{B}$ can be considered to be statistically significantly close.

For the above procedure, the ISD is evaluated only within the context of other sorting arrangements which are possible by keeping the same number and cardinality of piles, but permuting the elements. This is what the process of simultaneously reordering the rows and columns of $\mathcal{B}$ achieved. We find this result somewhat restricted, as our subjects made arrangements with a range of different numbers and cardinalities of piles. Instead, to assess the significance of the mean ISD, several Monte Carlo simulations were used. Each simulation gathered data corresponding to a very large set of pseudo-trials (from $2,000$ to $5,000$) in which the mean ISD for a set of $n = 11$ randomly generated sorting arrangements was calculated. We choose the size $n = 11$ to correspond to our set of subjects. Since we measure the mean ISD using three different distance statistics, for each Monte Carlo simulation, data must be generated to produce the distributions corresponding to each measure in order to assess the significance of each mean ISD. There were three Monte Carlo simulations conducted, each corresponding to a different method of selecting random partitions. Different methods of random partition selection were necessary, since we did not want to randomly select a sorting arrangement from the entire space of possibilities. The entire space of possibilities was so unlike the range of possibilities facing a human subject, given their cognitive constraints. Three relevant subsets of the entire space of sorting arrangements were identified. These are further described and justified in Section 4.4.1 and are summarized in Table 4.3.

	Restriction for Subset	Space of Possibilities	Pseudo-Trials Conducted
	none	$\approxeq 4 \times 10^{17}$	*none - not relevant*
I	6 ± 2 cells	$\approxeq 1.2 \times 10^{17}$	$4,548$
II	6 ± 1 cells	$\approxeq 3.8 \times 10^{16}$	$1,963$
III	6 ± 1 cells and *balanced*	$\approxeq 4.0 \times 10^{14}$	$2,010$

Figure 4.3: Summary subsets of partition data used for Monte Carlo simulations I, II, and III.

4.3.4 Results

Perception of Authorship

For each subject's partition, the similarity was calculated to the authorship vector. The influence of authorship is given by a measure of the distance between the subject's partition and the authorship partition (the "distance-to-author" measure). Each subject's "distance to author" measure was assessed for significance by comparison to the data from the Monte Carlo simulations. For example, we wanted to evaluate if a subject placed writing samples in piles that were by the same author more frequently than was attributable to random chance. The distribution of the mean inter-subject distances produced from each Monte Carlo simulation is akin to the distribution of mean distance between any two subject partitions. By using this data, we assessed, for each subject, where the subject-to-author partition distance was significantly closer than the typical distance between two subject partitions. Although this doesn't assess the subject-to-author distance against a distribution of distances between the author partition and randomly generated partitions[8], it still provides a good comparison. We still see that a significant number of subject to author partitions were not significantly more similar than this cruder approximation.

Table 4.4 summarizes these tests for significance (at a .95 significance level) with three different sets of data (each set corresponding to Monte Carlo simulations I, II, and III). The results for each distance statistic are included. The results using the Δ-distance statistic are included here as an illustration of the confounding that is possible (Section 4.3.2 contains the explanation of Δ confounding). Since the authorship partition had 7 piles and was very balanced, we would expect that Monte Carlo simulation III would be particularly relevant.

We see that, based on both the ρ and the Γ measures, subjects 6, 7, and 10's sorting

[8] In the future, the Monte Carlo simulations will be modified to produce this information as well.

Subject	ρ			Γ			Δ		
	I	II	III	I	II	III	I	II	III
1	●	●	●	●	●	❍	❍	❍	●
2	●	●	●	●	◗	❍	●	●	●
3	◗	◗	◗	❍	❍	❍	●	●	●
4	●	●	●	●	●	◗	●	●	●
5	❍	❍	❍	❍	❍	❍	❍	◗	●
6	●	●	●	●	●	●	❍	❍	◗
7	●	●	●	●	●	●	❍	◗	●
8	❍	❍	❍	◗	◗	❍	❍	❍	●
9	❍	❍	❍	❍	❍	❍	❍	◗	●
10	●	●	●	●	●	●	❍	❍	●
11	❍	❍	❍	◗	❍	❍	❍	◗	●

● - significant similarity
❍ - significant dissimilarity
◗ - no significant similarity or dissimilarity

Figure 4.4: Evaluation of similarity to authorship partition.

arrangements were significantly correlated with the authorship of the writing samples. Subjects 1, 2 and 4 are borderline due to the conflicting results from Monte Carlo Simulation III. For these subjects, a significant number of the writing samples by the same author were perceived to be more similar than the writing samples by different authors. These subjects had an awareness of the style of each author and tended to use this awareness in making their sorting arrangements. There were 2 subjects (Subjects 1 and 2) would seem questionably similar, but this is not the case since there is no significant similarity relative to the data from the third Monte Carlo trial.

Interestingly, the remaining subjects 3, 5, 8, 9, and 11 are significantly dissimilar (i.e., the likelihood of such dissimilarity occurring in a random partition is less than 5%).

Frequency Distribution for Pair Occurances in Same Pile
Split by : Authorship of Pair (Same/Different)

From (≥)	To (<)	Total Count	Total Percent	Different Authors Count	Different Authors Percent	Same Author Count	Same Author Percent
0.000	1.000	44	15.942	42	16.935	2	7.143
1.000	2.000	60	21.739	59	23.790	1	3.571
2.000	3.000	69	25.000	60	24.194	9	32.143
3.000	4.000	53	19.203	45	18.145	8	28.571
4.000	5.000	30	10.870	24	9.677	6	21.429
5.000	6.000	14	5.072	13	5.242	1	3.571
6.000	7.000	3	1.087	3	1.210	0	0.000
7.000	8.000	3	1.087	2	.806	1	3.571
	Total	276	100.000	248	100.000	28	100.000

Figure 4.5: Pairs of writing samples, organized by authorship and subject agreement of similarity.

Another way to look at our data is by all the 276 pairs of writing pairs (taking the 24 writing samples, 2 at a time). This data is summarized in Figure 4.5. We wanted to investigate whether, of all the pairs of writing samples that a given number of subjects placed in the same pile, a disproportionate number of the pairs were by the same author. If the authorship of the writing samples did not figure in the subject's assessment of the writing samples' stylistic similarity, then for a set of pairs placed in the same pile by the subjects, there should be a proportional number of same-authored and differently-authored writing samples represented. Of all the possible pairs of writing samples, 28 (10.14%) were by the same author and 248 (89.86%) were by different authors. Using the data from Figure 4.5, several calculations were made to test this hypothesis. There was no pair of writing samples that every subject placed in the same pile; the three most agreed-upon writing pairs were placed in the same pile by 7 subjects. Of the three pairs, one of them had the same author (33%). But in the set of next most agreed-upon pairs of writing sample, placed in the same pile by 6 subjects, not one of the pairs had the same author (0% of three). For pairs placed in the same pile by 2, 3, 4, and 5 subjects, the proportion of pairs that had the same author was 9/69 (13.04%), 8/53 (15.09%), 6/30 (20%) and 1/14 (7.14%) respectively. We see that these cases contain a disproportionate number of same author pairs. This is illustrated by Table 4.6. If authorship were not a factor, then the percentages for same-author and different-author pairs should be equal to each other in each row. This is not the case, as we can see that the same-author pairs are consistently overrepresented.

Number of Subjects	Pair % Placed in Same Pile (% of Total Pairs)	
	Same Author (28/276, 10.14%)	Different Author (248/276, 89.86%)
≥ 0	100% (28/28)	100% (248/248)
≥ 1	92.86% (26/28)	83.06% (206/248)
≥ 2	89.29% (25/28)	59.27% (147/248)
≥ 3	57.14% (16/28)	35.08% (87/248)
≥ 4	28.57% (8/28)	16.94% (42/248)
≥ 5	7.14% (2/28)	7.26% (18/248)
≥ 6	7.14% (2/28)	2.02% (5/248)
≥ 7	3.57% (1/28)	0.08% (2/248)

Figure 4.6: Pairs of writing samples, organized by authorship and subject agreement of similarity (with proportions given by authorship).

4.3.5 Inter-Subject Agreement

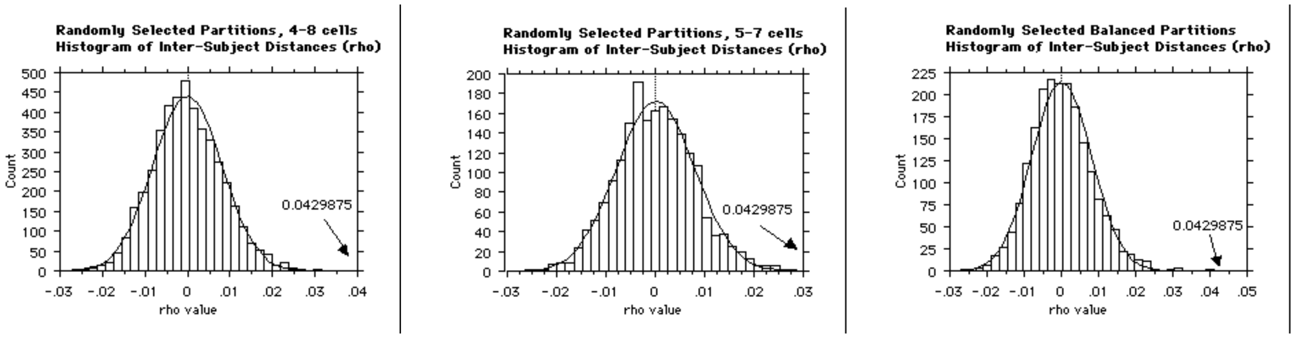
Figures 4.7, 4.8, and 4.9 summarize the data from the Monte Carlo simulations, organized by the distance statistics used.

In Figure 4.7, we see that the mean ISD among our subjects calculated using the ρ distance statistic ($\overline{ISD}_\rho = 0.0429875$) is significantly high. No pseudo-trial resulted in a higher ISD.

The mean ISD, calculated using the Γ distance statistic, was $\overline{ISD}_\Gamma = 11.9273$. In Figure 4.8, we see that it is significantly high, since our observed $\overline{ISD}_\Gamma$ was in the 0-percentile for simulations I and II (there were no values higher than what we observed). Only in the third simulation do we see that 0.4% of the randomly generated data has higher value.

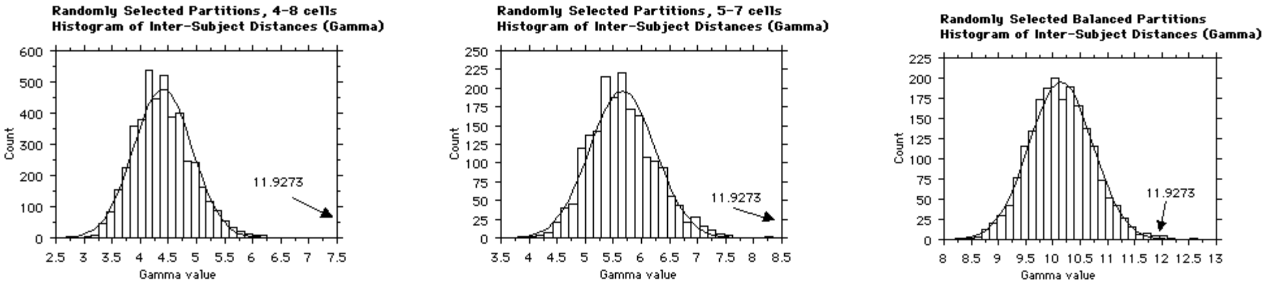
The confounding of Γ is clear in Figure 4.8. For simulations I and II, $\overline{ISD}_\Gamma$ was in the 100-percentile (all randomly generated data had higher or equal values). For simulation III, constrained to randomly generate only balanced partitions (and therefore with relatively few 1's), 1.5% of the randomly generated data is greater or equal to $\overline{ISD}_\Gamma$.

In summary, our subjects showed a significant amount of agreement. This result has been verified using two different distance statistics and three different Monte Carlo Simulations.



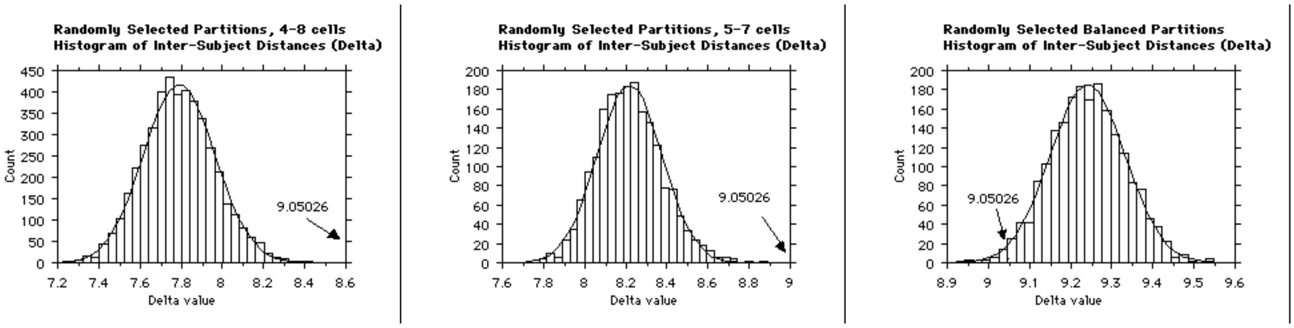
(a) Monte Carlo Simulation I (b) Monte Carlo Simulation II (c) Monte Carlo Simulation III

Figure 4.7: Inter-subject agreement using ρ distance statistic.



(a) Monte Carlo Simulation I (b) Monte Carlo Simulation II (c) Monte Carlo Simulation III

Figure 4.8: Inter-subject agreement using Γ distance statistic.



(a) Monte Carlo Simulation I (b) Monte Carlo Simulation II (c) Monte Carlo Simulation III

Figure 4.9: Inter-subject agreement using Δ distance statistic.

4.3.6 Sentence Count as an Indicator of Stylistic Similarity

We now consider sentence count as an explanation of the sorting arrangements produced by the subjects. It is possible, faced with no other ideas, a subject might group the writing samples by the size of the sample (where size is the number of sentences). The writing samples in our testing materials ranged from 1 sentence up to 8 sentences (mean sentences per sample is 4.33, with standard deviation of 1.79). Since each subject produced a number of piles, we calculated the mean and standard deviation of sentence count for each pile. A standard deviation of 1 or more represents a significant difference of sentence lengths in the writing samples that make up a pile (e.g., such as a pile of writing samples with sentence counts differing by 2 or more). Therefore, we would expect that sorting arrangements influenced by sentence length would have piles with low sentence count standard deviations (e.g., within the range of 0 to 1). As shown in Figure 4.10, our data does not resemble this type of distribution at all. It is interesting to note that of all the piles with sentence count standard deviations of 0, 16 of the 17 are single-element piles. From this, we can assume that subjects sorted the piles according to some criteria other than sample size.

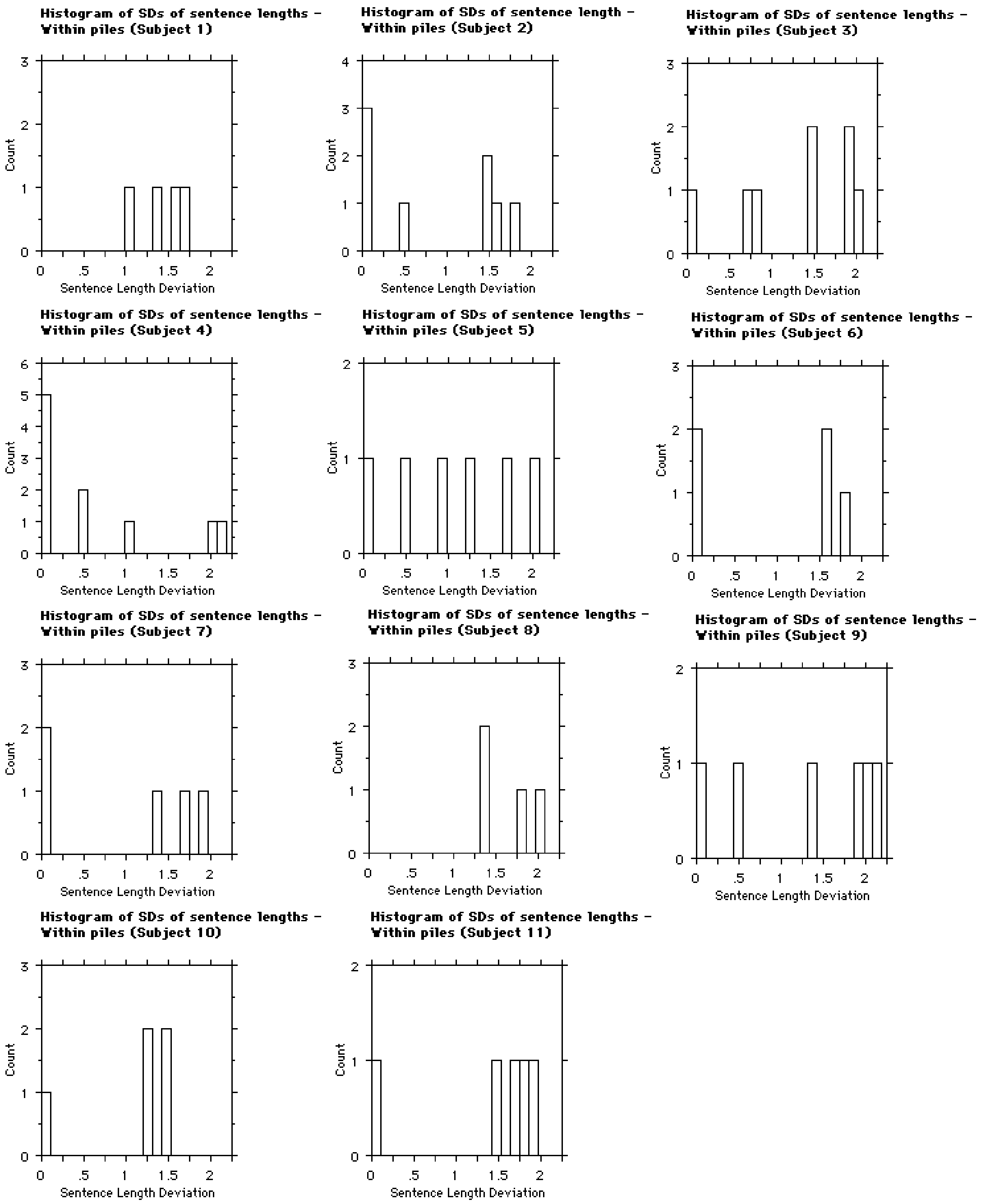


Figure 4.10: Distribution of the deviation of the lengths of the writing samples within the piles of the sorting arrangements, by subject.

4.4 Discussion

4.4.1 Restriction on Permutation Space

As mentioned earlier, every possible unique sorting arrangement has a one-to-one correspondence to a partition. The number of ways to arrange 24 writing samples into any number of piles of any length is best imagined as the space of all possible partitions of 24 unique elements. The size of the space of all partitions of n unique elements is given by Bell's number, which can be calculated using the following recurrence relation[9]:

$$B_n = \sum_{k=1}^{n} \binom{n-1}{k-1} B_{n-k}, \text{ where:} \tag{4.2}$$

$$B_{24} = 445,958,869,294,805,289 \approxeq 4 \times 10^{17} \tag{4.3}$$

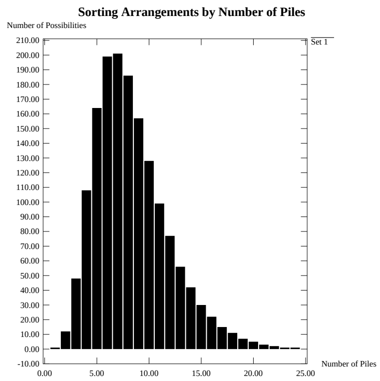


Figure 4.11: Partition space for 24 elements

Figure 4.11 shows the general distribution of the space of all possible partitions, broken down by number of cells with logarithmic scaling (to keep the graph on the page!). From the sorting arrangements that our subjects made, it is clear that human judges very likely are not randomly choosing from the set of all possible partitions. The mean number of piles for our subjects is 6 (standard deviation = 1.90). The likelihood of randomly selecting a 6-cell partition from this space of probabilities is under 2%. Rather, this result is probably explained by the constraints of short-term memory. We would like to randomly choose

[9] See (Cameron, 1994) for more information about Bell's number.

partitions from this space in order to produce a distribution against which to compare our subject's performance. However, we do not want to randomly choose partitions from the entire space since it is not representative of the space of possibilities facing a subject. According to our data, the number of piles in the subject's sorting arrangements mostly fell into the range of 6 ± 2 piles. It was from this range that the data for our first Monte Carlo simulation was drawn. As a refinement, we noticed that most of our subjects made 5 or 7 piles, so the set of partitions with 6 ± 1 piles was the basis for our second Monte Carlo simulation. Lastly, since our subject's sorting partitions are relatively balanced (where balanced means that the size of the cells is roughly equal), we wanted to randomly select a set of balanced partitions. Therefore, a third subset was considered which contained only balanced partitions with 6 ± 1 cells. For a partition to be considered balanced, it must satisfy the equation given below.

$$\text{max cell size} - \text{min cell size} \leq \frac{24}{\text{number of elements in partition}} \tag{4.4}$$

The resulting set of balanced partition in a space of possibilities which was, on the whole, more balanced than our data. In summary, the partitions for the three Monte Carlo simulations were drawn from the following corresponding subsets:

I 6 ± 2 cells (size $\approxeq 1.2 \times 10^{17}$)

II 6 ± 1 cells (size $\approxeq 3.8 \times 10^{16}$)

III 6 ± 1 cells and *balanced* (size $\approxeq 4.0 \times 10^{14}$)

4.4.2 Future Work

This exploratory study has revealed a number of interesting results. First, the authorship of writing samples has a significant effect on the stylistic assessments of those writing samples. The effect on subjects in this study was divided, since half made stylistic judgements that were significantly similar to the authorship of the writing samples and the other half made significantly dissimilar stylistic judgements. Future study is required to investigate this. This study should include a larger group of subjects as well as a larger set of testing materials.

Second, this study found that there was a significant amount of agreement among the subjects in terms of their stylistic assessments. One issue that was identified early in this study was to identify subgroups of subjects with especially similar stylistic assessments[10] (Section 4.1.1). As a preliminary step in this direction, a factor analysis was conducted. The goal of the factor analysis was to see if the observed values of many dependent variables (in our case, each subject's stylistic assessment is described by a $(0,1)$-vector of 276 inter-dependent elements) can be explained in terms of a smaller number of factors. Figure 4.12 shows that, in terms of two basic factors, our data falls into three clusters. This is suggestive of three underlying stylistic assessment patterns. Further investigation is required; perhaps two of the patterns correspond to the different effects of authorship.

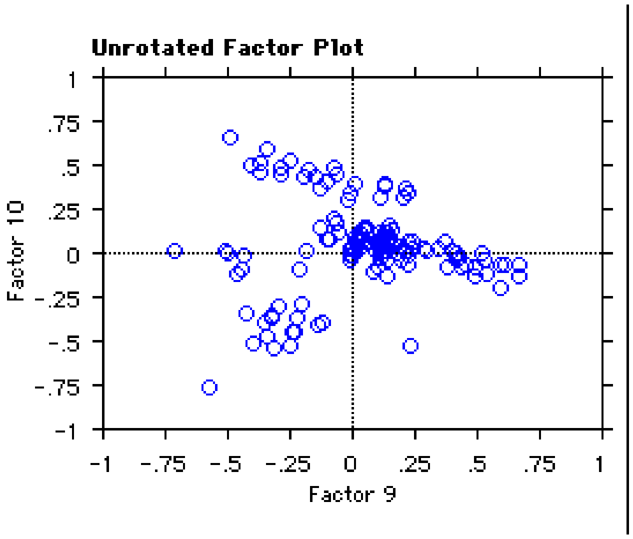


Figure 4.12: Factor analysis of the dependent variables describing the subject groups' stylistic assessments.

This exploratory study found that the size of the writing samples does not explain our subject's stylistic assessments. In the future, other less-simplistic indicators should be investigated, such as syntactic, lexical, or stylistic features. Of these potential indicators, investigating the stylistic features should have the first priority. Future work should include determining the correlation between the stylistic constructs of clarity/obscurity, abstraction/concreteness and staticness/dynamism, established by DiMarco (1990, 1993), and the

[10] The concern was that there may be such subgroups, but this would be obscured in looking at the overall similarity of stylistic assessments.

stylistic assessments made by a group of subjects. This should be straightforward, since the stylistic constructs have established syntactic indicators.

This exploratory study was motivated by the desire to learn more about the subjective nature of stylistic assessment. The eventual application of this knowledge will be in the design of a computational tool intended to provide assistance with stylistic revision. In order for such a tool to be useful, it must accurately emulate the response of the reader audience not only with respect to general stylistic judgement, but also for the detection of stylistic problems. Although examining the subjective assessment of style was a good starting place, this work should be extended to also include assessment of stylistic problems in text. The framework for classifying problems in text (according to locus and granularity) developed by Schriver (1992) could be used initially.

In general, for future work, only the ρ and the Γ distance statistics should be used as a measure of similarity. The Δ measure is prone to confounding, as discussed in Section 4.3.2. The Monte Carlo paradigm has been very useful; however, one main drawback was the processing time required to produce and analyze the set of randomly generated partitions required for each pseudo-trial. The analysis part of this task was ported from the original implementation in an interpreted language (Maple) to a much faster compiled implementation (C++), but computational processes that generated the random partitions required for the pseudo-trials took over 150 hours to complete. For larger trials, the entire computational process must be implemented in a faster, compiled language or on a faster machine.

Another promising approach for future work would be to use another paradigm to measure similarity. The free-sort task was intuitive and easy for the subjects, but it only could produce binary values of similarity (a pair of writing samples could either be placed in the same pile or apart). This type of task does not permit graded relationships; graded relationships can only emerge through the aggregation of many sorting arrangements. As an improvement, the writing samples could be spatially organized in terms of their stylistic similarity (or more directly, by the presence of stylistic incongruities). This task produces useful similarity data, as established in (Goldstone, 1994), and should be considered for future work.

4.5 Conclusions

This exploratory study was motivated by the desire to learn more about the subjective nature of stylistic assessments made by readers of written texts. To accomplish this, a group of subjects was presented with a "free-sort" of a set of writing samples. A free-sort is a task in which a subject is given a set of items and is instructed to arrange the items into piles according to some criteria; for this exploratory study, the criteria was stylistic similarity. Past stylostatistical research has attempted to establish statistical correlates of superficial textual features to authorial style, mainly with the motivation of attributing authorship. The first goal of this exploratory study was to determine if readers find writing samples written by the same author more similar than those that are written by different authors. Interestingly, only approximately half of the subject's stylistic judgements were significantly influenced by the authorship of written samples. The other half of the subjects made stylistic assessments that were negatively correlated with the authorship of the samples. This suggests some limitations for using authorial stylostatistical tests to predict a reader's impression of a text's style.

A second goal of this study was to determine if there was agreement among a group of readers with respect to their judgements of stylistic similarity. Broadly speaking, our subjects' stylistic assessments were significantly similar. What is more surprising is that, in light of the different effects of authorship and the factor analysis subsequently conducted, there may be different patterns of stylistic judgement. This area will be the subject of further investigation. For the time being, sweeping predictive statements about a text's stylistic effect in a reader audience should be made cautiously, since a group of readers might not share homogeneous stylistic judgements.

Finally, the stylistic indicator of sentence length was explored as a possible explanation for the observed stylistic assessments. This study found that sentence length is an unsatisfactory predictor, and suggests that more sophisticated indicators of style will be required to explain perceived stylistic similarity.

Chapter 5

Conclusions

In this chapter, the conclusions of the thesis are presented, as well as areas for future research. The chapter is divided into three sections. First, we present a summary of the work in this thesis that addresses the methodological issues. The second section deals with the work to build a requirements analysis for an application for achieving stylistic incongruity, and the third deals with work done for the preliminary design stage.

5.1 Design Methodology

5.1.1 Contributions

This thesis has succeeded in identifying an appropriate methodology for the development of a solution to the problem of stylistic incongruity experienced by collaborative writers. The iterative software development model was chosen because of its flexibility to handle poor problem definition, the absence of a clear solution, and a problem area drawing from several different research communities. Although the emphasis of this thesis was to address research issues (complemented by the goal of developing software), I felt that using a software development model would provide a more useful decomposition for this thesis research than the intuitive decomposition that typically results from general investigative approaches. Using this development model would subsume one of the original motivating research goals, which was to identify a useful approach for studying the phenomenon of stylistic incongruity in text, especially collaboratively written text.

Applying this model has given this thesis three focuses: developing a clear problem definition, analyzing user requirements, and creating a usable and useful design.

The work done to develop a clear problem definition was presented in chapter 2. In this chapter, two types of stylistic inconsistencies were described and differentiated, and the term *stylistic incongruity* was put forth in order to accurately identify the type of stylistic inconsistency of interest to this thesis. With the problem area clarified, the task for this thesis was identified as developing a software solution addressing the problems in *achieving stylistic congruity* or *eliminating stylistic incongruity.*

This thesis also was successful in separating two important issues that have been intertwined previously: the abstraction of the problem, and the design and implementation of its solution. Past effort to develop a method of eliminating problematic stylistic inconsistency has assumed a particular design (e.g., (Glover and Hirst, 1995)). Other software implementations that were designed and implemented to perform stylistic assessment are limited, since they do not have a separate notion of stylistic incongruity; instead, they rely on a larger and poorly-defined notion of a *stylistic error.* In order to analyze these past efforts, a framework was developed to tease apart the implementation details, the design, and the purpose (which in turn indicates the problem definition). It was used to describe not only the traditional style-checkers, but other types of applications with capabilities for stylistic assessment (e.g., applications for natural language generation, machine translation, and language instruction).

5.2 Requirement Analysis

5.2.1 Contributions

In this thesis, the requirements were analyzed with a focus on the user of the tool. The intended user of this tool was defined to be any member of a collaborative writing group and not necessarily those doing the text transcription (the traditional sense of 'writer'). At the basis of the analysis was the belief that a stylistic congruity tool should support the needs of user's natural activities and not those imagined by the software designer. The goal of the analysis in this thesis was to provide a basis for the identification of *support strategies*, strategies of providing assistance by targeting an area of difficulty experienced by collaborative writers, at any possible stage of the collaborative writing process. Additionally, the hypothesis was that the effectiveness of a tool would be partially dependent on the particular area of difficulty targeted. For example, one past approach assumed a strategy of targeting difficulties in revising — a common and noticeable area of difficulty, but not acknowledged as just one possible area of many, nor shown to be the most effective target. The goal of the analysis was to uncover as many areas of difficulty as possible and to describe them with respect to their effect on stylistic incongruity.

This analysis has been completed and the results have been organized in a taxonomy based on a classification of the collaborative writer's activities. In order to perform this analysis, a model of the user's patterns of collaborative writing was developed by unifying three different models of collaborative writing as well as a model of the singular writing process.

A key observation from this analysis is that stylistic congruity can be characterized in terms of:

- properties of the artifact of the (collaborative) composition process — the text; or
- the properties of the composition process that caused the stylistic incongruity in the text.

5.2.2 Future Work

This thesis has taken the first step in the requirements analysis by identifying the need to select — rather than assume — support strategies based on a targeted area of difficulty. It has also taken the second step in identifying a set of such problem areas. The next step should be to evaluate; there must be some basis for the selection of a target problem area for a support strategy. These problem areas must be evaluated with respect to their weight of consequence for the text's stylistic congruity; with respect to their pervasiveness in collaborative writing practices; and with respect to their suitability as the target for computational assistance. Although it was an observation with respect to the last criterion — that targeting revision activity seems to be the obvious support strategy, but perhaps not the most effective support strategy — which helped focus the analysis in the first place, developing these criteria more fully remains an area for future work. These criteria should be used in not only the tool design stage, but also in evaluating the tool's effectiveness.

5.3 Design

5.3.1 Contributions

In the requirements analysis presented in chapter 2, some preliminary design directions emerged. More specifically, there were several areas of difficulty that were suggestive of a support strategy that entails computational stylistic assessment (e.g., possibly including assessment for stylistic incongruities directly, or more indirect assessments). This thesis describes results that establish a basis for the future design of stylistic assessment components. This basis is in the form of an abstract model of stylistic assessment, the *construct/indicator model* (chapter 3). This thesis described how the stylistic assessment capabilities in existing software tools all can be explained in terms of the components of this model, with their differences given in terms of the constructs and indicators chosen, as well as the sophistication and validity of the indicators.

As well, this thesis described an inventory of constructs and indicators that was created as a resource for future stylistic assessment designs (chapter 3). The inventory was compiled by identifying and extracting relevant pieces from a diverse range of research papers on stylistics.

This thesis also has succeeded in providing an answer — based on experimental evidence — to the question of the appropriateness of a deterministic indicator, rather than a probabilistic indicator, in the construct/indicator model. The indicator component in existing stylistic tools, both commercial and academic, is used deterministically to measure one or more qualitative stylistic constructs (which vary according to the particular application). Human stylistic assessment, however, involves a high degree of subjectivity, which often results in a lack of agreement on the outcome. This has consequences for the construct/indicator relationship; to be a valid measure, the indicator either must operate within the context of a particular type of reader audience (classified by some as-yet unknown criteria that distinguish the various types of subjective responses) or must reflect the variance in the corresponding human judgement. To address this issue, an experiment was conducted (chapter 4). The results of the experiment demonstrated the following:

- The correlation of human stylistic assessment with authorship is polarized; either authorship has a strong positive correlation with stylistic assessment or a strong negative correlation. This is an important result, since the relationship of stylistic incongruities and authorship boundaries has been previously assumed, and the use of stylostatistical indicators of authorship were proposed as having a potential application in the detection of stylistic incongruities.

- The stylistic assessments corresponding to a group of human judges are significantly similar, but a subsequent cluster analysis indicated distinct categories of stylistic assessments. This is an important result, since it not only confirms our intuition that, while stylistic assessment is not completely anarchic, there are different types of subjective experiences.
- A simple indicator of text length does not explain the similarities in stylistic assessments.

5.3.2 Future Work

This thesis has successfully completed the first cycle of the iterative software development model. In the next iteration, preliminary design directions should be developed, and refinements should be made to the requirements analysis. In this section, some directions are suggested for the design of an assistance tool.

Representation Facility for the Concept of the Text's Intended Communication

The planning done by a group of collaborative writers is very important to the rest of the composition process and has an impact on the eventual stylistic congruity of the resulting text. In section 2.5, a number of problem areas in the planning subprocess were identified. These included:

- difficulty in constructing a conceptualization of the text's intended communication;
- difficulty in ensuring that this concept is shared among all the collaborative writers; and
- discontinuity from the planning to the transcription stages.

In order to alleviate these problems, a facility within the collaborative writing environment can help support the planning process.

A representation facility could be added to the existing planning facilities. Many existing collaborative writing environments have planning facilities in the form of scratch pads and spaces for representing brainstorming sessions. These facilities could be modified so that a representation of the text's intended communication can be created and saved for future reference.

The nature of this representation is an area for future development. As a starting place, the framework for pragmatic information used by PAULINE can be used. Even making this pragmatic framework available for reference could be beneficial to a group of collaborative writers. It could encourage the collaborative writers to acknowledge, to discuss, and to negotiate issues that are important in avoiding stylistic incongruity.

The representation should be available at any stage during the composition process. For example, a reviewer may wish to refer to the initial text plan to verify the writer's original intentions. Similarly, the transcriber can refer to the initial text plan to verify their own goals in writing.

The facility should have the capability to handle modifications to the concept of the text's intended communication, as the text plan might evolve. One issue for this type of strategy is that the representation may not be kept up-to-date.

This facility subsumes a dictionary facility, such as that proposed by Mawby (1991) (see section 2.5.1). The dictionary would contain terms with assigned meanings for the text. The facility should be able to reuse and modify past dictionaries. For technical documents, this facility should be integrated with the thesaurus facility of the collaborative writing environment in order to provide verification for the group's word use. More specifically, the writers should prefer the use of the term with the assigned meaning rather than a synonym, in order to be stylistically congruous. This facility could also support an automated glossary production facility.

Facility for a Global View of the Text

In Chapter 2, the research of Severinson Eklundh (1992) was discussed, especially with respect to the importance of the author having a global perspective during the composing process. Her results showed than singular writers have difficulty constructing a global perspective of the text being composed. Collaborative writers have difficulties constructing a global perspective as well. They experience the same difficulties as singular writers, but also experience other difficulties, such as being "out-of-step," a situation where what members are writing is based on incorrect perceptions of what the others are writing. These difficulties underline the fact that members writing segments of a text must have access to global information.

To combat this, global information must be made available to all members of the collaborative writing group during the composing process, even those writing isolated segments. Through this global view, group members can monitor the writing done by other members. This promotes communication during the transcription process, which helps to repair divergences, to combat being "out-of-step," and to renegotiate and to promote a shared concept of the text's intended communication. The communication upon which the common conceptualization relies can possibly be indirectly achieved, through the inferences that an author can make of the other collaborators' conceptualizations (based on what they have written). Reducing or eliminating any of these problems will reduce stylistic incongruity.

One type of behaviour that the design of the facility must circumvent is the collaborative writer's need for control. Members are often reluctant to share their writing until it is polished. Document control gives members the perception of status (Posner, 1991), so a facility that makes all parts of a document available to everyone will be met with resistance. An allowance can be made for this, while at the same time satisfying the group's need for access. This lies in the identification of the features of the written segments that are most salient to stylistic congruity, and then sharing only that information. While viewing the actual text of a segment would yield the most information, an extraction of the salient parts could still be useful. This salient information would consist of the stylistic constructs relevant to stylistic congruity. A computational process could then perform stylistic assessments of all text segments and make the results available to all group members. An issue for this design is managing the global view of a text that is continually evolving.

Visualization Facility

In chapter 2, several areas of difficulty were based on the writer's inability to detect, diagnose, and repair stylistic incongruities. There are some indications that it will be difficult to create a computational application that emulates the skills required to perform these tasks: the subtlety in judgement; the use of the text's context; the real-world knowledge required; and the often intuitive knowledge about the audience that is also necessary. Instead of emulating these skills directly, I believe that a better approach would be to support these activities by providing the user with a modified form of the text. The text should be preprocessed so that the information that is salient to the tasks of detection, diagnosis, and repair are presented to the user. The presentation should be in the form of a visualization of the salient components of the text. This form of presentation, as opposed to detection and repair facilities analogous to a spelling-checker, has the advantage of being able to use the metaphors that writers often use to describe the style of text (e.g., acoustic, spatial, vocal). The visualization facility could also support many navigation and editing functions, as well as being integrated with related facilities, such as a facility for creating global views of a text, or a representation facility for the conceptualization of the text's intended communication.

Bibliography

C. W. Anderson and G. E. McMaster. Computer assisted modeling of affective tone in written documents. *Computers and the Humanities*, **16**(1):1–9, 1982.

C. W. Anderson and G. E. McMaster. Modelling emotional tone in stories using tension levels and categorical states. *Computers and the Humanities*, **20**(1):3–9, 1986.

J. L. Austin. *How to Do Things with Words.* Harvard University Press, 1962.

Eva L. Baker, Nancy K. Atwood, and Thomas M. Duffy. Cognitive approaches to assessing the readability of text. In Alice Davidson and Georgia M. Green, editors, *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, pages 55–85. Lawrence Erlbaum, Hillsdale, New Jersey, 1988.

John Charles Baker. Pace: A test of authorship based on the rate at which new words enter an author's text. *Literary and Languistic Computing*, **3**(1), 1988.

John Benjafield and Ron Muckenheim. Dates-of-entry and measures of imagery, concreteness, goodness, and familiarity for 1,046 words sampled from the Oxford English Dictionary. *Behavior Research Methods, Instruments and Computers*, **21**(1):31–52, 1989.

John Benjafield, Kris Frommhold, Tom Keenan, Ron Muckenheim, and Dierk Mueller. Imagery, concreteness, goodness, and familiarity ratings for 500 proverbs sampled from the Oxford Dictionary of English Proverbs. *Behavior Research Methods, Instruments and Computers*, **25**(1):27–40, 1993.

Douglas Biber. *Variation across speech and writing.* Cambridge University Press, Cambridge, England, 1988.

Philip Bolt. Grammar checking programs for learners of english as a foreign language. In Masoud Yazdani, editor, *Multilingual Multimedia*, pages 140–197. Intellect Ltd., 1993.

S.P. Borgatti, M.G. Everitt, and L.C. Freeman. *UCINET IV Version 1.00.* Analytic Technologies, Columbia, 1992.

Barron Brainerd. The computer in statistical studies of William Shakespeare. *Computer Studies in the Humanities and Verbal Behavior*, **4**(1), 1973.

Penelope Brown and Steven C. Levinson. *Politeness: some universals in language usage.* Cambridge University Press, New York, 1988.

B. Bruce, A. Rubin, and K. Starr. Why readability formulas fail. *Reading Education Report No. 28*, 1981.

Peter J. Cameron. *Combinatorics: Topics, Techniques, Algorithms.* Cambridge University Press, Great Britain, 1994.

John B. Carroll. Vectors of prose style. In Thomas A. Sebeok, editor, *Style in Language*, pages 283–292. M.I.T. Press, 1960.

Robert Cluett. *Prose Style and Critical Reading.* Teacher's College Press, Columbia University, 1976.

David Crystal and Derek Davy. *Investigating English Style.* Longmans, Green and Co. Ltd., 1969.

Nell Boylan Dale. Traditional and computational stylistics: A model of metaphor. *Association for Literary and Linguistic Computing Bulletin*, **5**(2):132–146, 1977.

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

Chrysanne DiMarco. *Computational Stylistics for Natural Language Translation.* PhD thesis, University of Toronto, 1990. Published as technical report CSRI-239, Department of Computer Science, University of Toronto.

Lubomir Doležel. A framework for the statistical analysis of style. In Lubomir Doležel and Richard W. Bailey, editors, *Statistics and Style*, pages 10–25. American Elsevier Publishing Co. Ltd., 1969.

Lisa S. Ede and Andrea A. Lunsford. *Singular Texts/Plural Authors: perspectives on collaborative writing*. Southern Illinois University Press, Carbondale, IL, 1990.

N. E. Enkvist. *Linguistic Stylistics*. Mouton, The Hague, 1973.

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

Linda S. Flower and John R. Hayes. The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg and E. R. Steinberg, editors, *Cognitive Processes in Writing*, pages 31–50. Erlbaum, Hillsdale; New Jersey, 1980.

L. Flower and J. R. Hayes. Plans that guide the composing process. In C. K. Frederiksen and J. F. Dominic, editors, *Writing: the nature, development, and teaching of written communication*, volume 2 'Writing: process, development and communication', pages 39–58. Lawrence Erlbaum Associates, 1981.

Alan Garnham and Jane Oakhill. Discourse processing and text representation from a mental models perspective. *Language and Cognitive Processes*, **7**(3–4):193–204, 1992.

Amy Herstein Gervasio, John Taylor, and Stuart Hirshfield. Modelling emotional tone in stories using tension levels and categorical states. *Behavior Research Methods, Instruments and Computers*, **24**(2):298–302, 1986.

Angela Glover and Graeme Hirst. Detecting stylistic inconsistencies in collaborative writing. In Thea van der Geest et al., editor, *Writers at work: Professional writing in the computerized environment*, Springer, London, 1995.

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

Robert Goldstone. An efficient methods for obtaining similarity data. *Behavior Research Methods, Instruments & Computers*, **26**(4):381–386, 1994.

Steven J. Green. *A functional theory of style for natural language generation.* Master's thesis, Department of Computer Science, University of Waterloo, 1992. Also published as Research Report CS-92-48, Faculty of Mathematics, University of Waterloo.

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

R. R. K. Hartmann. Style values: Linguistic appraoches and lexicographical practice. *Applied Linguistics*, **2**(3):262–273, 1981.

John R. Hayes and Linda S. Flower. Idenitifying the organization of writing processes. In L. W. Gregg and E. R. Steinberg, editors, *Cognitive Processes in Writing*, pages 3–30. Erlbaum, Hillsdale; New Jersey, 1980.

John R. Hayes and Linda S. Flower. Writing as problem solving. *Visible Language*, **14**(4):388–399, 1980.

John R. Hayes and Linda S. Flower. Writing research and the writer. *American Psychologist*, **41**(10):1106–1113, 1986.

Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet*, The MIT Press, 1995.

James E. Hoard, Richard Wojcik, and Katherina Holzhauser. An automated grammar and style checker for writers of Simplified English. In Patrik O' Brian Holt and Noel Williams, editors, *Computers and Writing: State of the Art*, pages 278–296. Kluwer Academic Publishers, 1992.

P. O'Brian Holt and N. Williams. Expert systems for report writing, Module F: Artificial intelligence/expert systems. In *Technology Enhanced Training Conference (tet'89)*. Royal Military College of Science, Shrivenham, Swindon, Wiltshire, July 1989.

Tony Honoré. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, **7**(2), 1979.

Eduard H. Hovy. *Generating Natural Language under Pragmatic Constraints.* Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 1988.

Eduard H. Hovy. Pragmatics and natural language generation. *Artificial Intelligence*, **43**:153–197, 1990.

Lawrence J. Hubert and Joel R. Levin. Evaluating object set partitions: Free-sort analysis and some generalizations. *Journal of Verbal Learning and Verbal Behavior*, **15**:459–470, 1976.

Lawrence J. Hubert. Generalized proximity function comparisons. *British Journal of Mathematical and Statistical Psychology*, **31**:179–192, 1978.

Estelle Irizarry. Exploring conscious imitation of style with ready made software. *Computers and the Humanities*, **24**(3):187–206, June 1989.

Steve Jones. Indentification and use of guidelines for the design of computer supported collaborative writing tools. *Computer Supported Collaborative Work*, **3**(3-4):379–404, 1995.

Erna Kelly and Donna Raleigh. Integrating word processing skills with revision skills. *Computers and the Humanities*, **24**:5–13, 1990.

Richard A. Lanham. *A Handlist of Rhetorical Terms*. University of California Press, 2 edition, 1991.

G. R. Ledger. A new approach to stylometry. *Association for Literary and Linguistic Computing Bulletin*, **13**(5), 1985.

Harry Levin and Margaretta Novak. Frequencies of Latinate and Germanic words in English as determinants of formality. *Discourse Processes*, **14**(3):389–398, 1991.

Keith Mah. *Comparative stylistics in an integrated machine translation system*. Master's thesis, University of Waterloo, 1991. Published as Technical Report CS-91-67, Department of Computer Science, University of Waterloo.

Robert A. Matthews and Thomas V. N. Merriam. Neural Computing in Stylometry I: An Application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, **8**(4), 1993.

Kelly L. Mawby. *Designing collaborative writing tools*. Master's thesis, University of Toronto, 1991.

Steve McGowan. Ruskin to McRuskin - degrees of interaction. In Patrik O' Brian Holt and Noel Williams, editors, *Computers and Writing: State of the Art*, pages 297–318. Kluwer Academic Publishers, 1992.

Alex Mitchell. *Communication and shared understanding in collaborative writing.* Master's thesis, University of Toronto, 1996.

Julian Newman and Rhona Newman. Three modes of collaborative authoring. In Patrik O' Brian Holt and Noel Williams, editors, *Computers and Writing: State of the Art*, pages 20–28. Kluwer Academic Publishers, 1992.

Julian Newman and Rhona Newman. Two failures in computer-mediated text communication. *Instructional Science*, **21**:29–44, 1992.

Ellen W. Nold. Revising. In C. K. Frederiksen and J. F. Dominic, editors, *Writing: the nature, development, and teaching of written communication*, volume 2. 'Writing: process, development and communication', pages 67–79. Lawrence Erlbaum Associates, 1981.

A. Pavio, J. Yuille, and S. Madigans. Concreteness, imagery and meaningfullness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement*, **76**(1, Part 2), 1968.

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

Ilona Rott Posner. *A study of collaborative writing.* Master's thesis, University of Toronto, 1991.

Rachel Rimmershaw. Collaborative writing practices and writing support technologies. *Instructional Science*, **21**:15–28, 1992.

Donald Ross Jr. and Robert H. Rasche. Eyeball: Descriptions of literary style. *Computers and the Humanities*, **6**:213–221, 1972.

Alan P. Rudell. Frequency of word usage and perceived word difficulty: Ratings of the Kučera and Francis words. *Behavior Research Methods, Instruments and Computers*, **25**(4):455–463, 1993.

Mark Ryan, Chrysanne DiMarco, and Graeme Hirst. Focus shifts as indicators of style in paragraphs. Research Report CS-92-35, Department of Computer Science, University of Waterloo, June 1992.

Karen A. Schriver. Teaching writers to anticipate readers' needs: A classroom-evaluated pedagogy. *Written Communication*, **9**(2):179–208, April 1992.

J. Selzer. What constitutes a "readable" technical style? In P. V. Anderson, R. J. Brockman, and C. R. Miller, editors, *New Essays in Technical and Scientific Communication: Research, Theory and Practice*, pages 71–89. Baywood Publishing Co. Inc, New York, 1988.

Kerstin Severinson Eklundh. Problems in achieving global perspective of the text in computer-based writing. *Instructional Science*, **21**:73–84, 1992.

M. Sharples, J.S. Goodlet, E.E. Beck, C.C. Wood, S.M. Easterbrook, and L. Plowman. Research issues in the study of computer supported collaborative writing. In Mike Sharples, editor, *Computer Supported Collaborative Writing*, pages 9–28. Ablex Publishing Corporation, 1993.

M. P. Shaughnessy. *Errors and Expectations: A Guide for the Teacher of Basic Writing.* Oxford, 1977.

John M. Smith and Maxwell E. McCombs. The graphics of prose. *Visible Language*, **5**(4), 1971.

M. W. A. Smith. Recent experience and the new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, **11**(3):73–82, 1983.

Teresa Snelgrove. A method for the ananlysis of the structure of narrative texts. *Literary and Languistic Computing*, **5**(3):221–225, 1990.

John Walter Spencer, Michael J. Gregory, and N. E. Enkvist. *Linguistics and Style.* Oxford University Press, London, 1964.

Manfred Stede. Lexical choice criteria in language generation. In *Proceedings of the Sixth conference of the European Chapter of the Association for Computational Linguistics*, pages 454–459, 1993.

William Jr. Strunk and E. B. White. *The Elements of Style*. Macmillan, third edition, 1979.

Kenneth Teshiba and Mark Chignell. Development of a user model evaluation technique for hypermedia based interfaces. Working Paper 88-15, Department of Industrial and Systems Engineering, University of Southern California, 1988.

Louis Ule. Recent progress in computer methods of authorship determination. *Association for Literary and Linguistic Computing Bulletin*, **10**(3):73–89, 1982.

Robert J. Valenza. Are the Thisted-Efron authorship tests valid? *Computers and the Humanities*, **25**(1), Feb 1991.

J.-P Vinay and J. Darbelnet. *Stylistique comparée du français et de l'anglaise*. Didier, Paris, 1958.

Werner Winter. Styles as dialects. In Lubomir Doležel and Richard W. Bailey, editors, *Statistics and Style*, pages 3 – 9. American Elsevier Publishing Co. Ltd., 1969.