TEXT TECHNOLOGY

❑ **Introduction**
Simon Shurville and Lyn Pemberton

❑ **The Importance of Subjectivity in Computational Stylistic Assessment**
Melanie Baljko and Graeme Hirst

❑ **The Future of Text Design on the World Wide Web: Developing the Reader User Interface**
Stephen Marcus

❑ **Crossing Cultures: Netting a Classroom Community**
Pamela Gay

❑ **Technology Rhetorics: Taking a Critical Perspective**
Ilana Snyder

VOLUME 9, NO. 1
SPRING 1999

WRIGHT STATE UNIVERSITY
LAKE CAMPUS

The Importance of Subjectivity in Computational Stylistic Assessment

Melanie Baljko and Graeme Hirst

Introduction

Computational processes with the ability to evaluate the style of text have a variety of applications, including the detection of plagiarism, forensic linguistics, and authorship determination. This research is motivated by the need to help users of collaborative writing environments maintain a consistent style in their documents. While the matter of whether the style of a text is correct or not is open to debate (as well as whether the question is valid to begin with), it is generally agreed that an inconsistent mélange of styles, as often found in collaboratively written texts, is undesirable. However, due to the subjective nature of style, the stylistic problems of a text differ from problems of grammar or spelling, for which a variety of natural language processing techniques have already been developed. Before techniques can be developed to assist in the elimination of problems of stylistic inconsistency, several underlying issues dealing with the subjectivity of style must first be addressed.

Computational Stylistics

In written language, one finds great variation in the style of different texts. Within different genres of written text, syntactic correlates can be used to distinguish between the respective writing styles (Biber, 1988). But even within a single genre of writing, such as newspaper, expository, or literary, there is still a wide range of stylistic variation. Some of this range is apparent within a single text; a skilled writer uses stylistic variation to engage the reader, to emphasize, and to accomplish many other goals. An objective characterization of stylistic variation would be extremely useful in the creation of stylistically aware natural language processing systems. Examples of such applications include machine translation systems that could understand the stylistic nuances of a source text and produce the corresponding stylistic effects in the target text, perhaps even by a different stylistic mechanism (Edmonds, 1998; DiMarco and Hirst, 1993) and natural language generation systems that could produce texts that say essentially the same thing but differ stylistically to be appropriate for different audiences and contexts (Hirst et al., 1997). Prototype applications that exhibit these kinds of stylistic awareness have been developed. For instance, the generation system PAULINE incorporated a set of heuristics that guided the style of the text produced (Hovy, 1990). As a refinement to this heuristic-based approach, a more precise formulation of style was developed in the form of a grammar of style (Green, 1992; DiMarco and Hirst, 1993; DiMarco and Mah, 1994). But in all of this research, it is implicitly assumed that it is valid for computational stylistic assessments to be made deterministically—that is, for any input text, there is one single corresponding stylistic assessment. For human readers, however, the perception and interpretation of the style of a text is a subjective matter. The meaning that a text has for a particular reader— which includes pragmatic features that are often conveyed stylistically—is not merely decoded from the text but rather is a matter of interpretation and thus dependent on the particular individual's experience and present state of mind. This kind of subjectivity, however, is simply not a factor in the stylistic assessments made by computational processes, and the validity of such assessments has not yet been established due to the paucity of empirical evidence.

Eliminating Stylistic Incongruities in Collaboratively Written Text

Characterizing stylistic variation in texts is important not only for the natural language processing applications described in the previous section, but perhaps even more so for the application that will be described in this section. While stylistic variation can be a positive quality of text, not all stylistic variation achieves some advantage. One such type of stylistic variation is seen in many texts that have been written collaboratively. Such texts often contain many writing styles, each corresponding to one contributor to the text. These styles are not merely inconsistent within the text but also incongruous: they are inconsistent to the point of detracting from the quality of the text. Readers often describe the problem of stylistic incongruity as "the text doesn't sound quite right" or "the text doesn't speak with a single voice."

Our long-term goal is to develop a collaborative writing tool that can be used to alleviate the problem of stylistic incongruity. Such a tool has potential use in the business and academic communities where the amount of collaborative writing is increasing and high text quality is expected (Ede and Lunsford, 1990).

In initial research (Glover, 1996; Glover and Hirst, 1995), it was assumed that in a process analogous to a spelling checker such a tool would employ a strategy of eliminating stylistic incongruities that are already present in a text. Two key components of this process were identified:

- Detection—the tool would discover stylistic inconsistencies.
- Diagnosis—it would present its findings in a manner that would enable authors to correct the inconsistencies, which implies being able to articulate the problems, and any suggestions for changes, in terms comprehensible to the average user.

This detection-and-diagnosis strategy targets revision activity within the writing process by both initiating revision and then supporting it. If the reliable detection of stylistic inconsistencies were possible, this type of strategy potentially could be quite useful. However, the computational detection of stylistic problems which initiates the revision activity is quite difficult. First, there are no clues that can be obtained from the progress of the writing process itself. The process of composing text includes many interleaved subprocesses, but revision—the process by which a piece of writing is successively improved—is not merely one of these subprocesses. Rather, it is another subactivity of composition, itself consisting of subprocesses and also interleaved within the processes of composition. Second, it is unlikely that an effective computational process can be designed when even humans do not have effective techniques or, if they do, can't articulate them. It is well known that the detection of stylistic problems is difficult even for human writers to master (Nold, 1981; Hartley, 1991; Schriver, 1992; Kelly and Raleigh, 1990). There might be stylistic incongruities in a text, but the author might fail to notice them. Indeed, authors might be dissatisfied with the document yet not know why (Glover and Hirst, 1995). The problem of detection is further compounded in collaboratively written texts. Stylistic incongruities are especially likely to occur in these texts, not just because the collaborative writing process is more complex than singular writing but also because the detection of problems in the text as a subprocess is often distributed among the collaborators. If all authors revise their own texts, then the authors will fail to detect incongruities arising between the separately authored components of the text. If the revision is performed at a later stage, writers might have difficulty diagnosing the problems and repairing them.

An additional obstacle is the role of subjectivity in stylistic assessment. The potential subjectivity in stylistic assessment could interfere with the detection of stylistic incongruities by the reader of a text. A text can conceivably contain a stylistic incongruity that is perceived by some and not by others. In general, the style of a text conveys part of the author's intended meaning, and it is known that what the audience brings to the reading process is important to the overall interpretation of the text's meaning. Since style is qualitative and its assessment is subjective, it is reasonable then to predict that some readers may interpret the intended content, which could be partly or mainly conveyed through style, differently from what the author intended. We know there is some agreement because an audience generally agrees on what a text means—i.e., there is not "interpretive anarchy," as Taylor (1992) points out—but neither the amount of subjectivity nor the role it plays is known.

The author, who emulates the reader during the revision process in an attempt to simulate the audience's response to a text, faces a difficult decision in whether to repair a perceived stylistic inconsistency or not. On one hand, the author may belong to the group that doesn't find the inconsistency deleterious, but in fact it might bother a sufficient proportion of the audience to warrant a change. On the other hand, the author may detect an incongruity but be unsure whether the expense of a change is warranted if only a small proportion of the audience would be bothered.

The possibility that some readers might perceive a stylistic incongruity while others do not is a very fundamental issue for an application designed to assist writers in eliminating stylistic incongruities. A tool that automatically detects stylistic incongruities must differentiate between those that need to be fixed and those that do not. In order to do so, more information is needed about how the audience as a whole perceives stylistic incongruities. As a first step, we have designed and conducted an experiment that explores subjectivity with respect to general stylistic perceptions among readers. We view this practice as a reasonable starting point since the detection of stylistic incongruities is a sub-skill of the more general skill of stylistic awareness.

Subjectivity in Stylistic Assessments

Our experiment was designed to answer the following questions:

- Do the readers of a given text indeed perceive its style in a similar way?
- Do readers perceive texts by the same author as stylistically more similar to each other than texts by different authors?

In order to answer these questions, the following methodological issues must first be addressed:

- What is a valid measurement of a reader's assessment of the style of a text?
- Given a set of such stylistic assessments, what is a valid measurement of the degree of their similarity?

We will describe our solution to the methodological issues first and then present the results that answer the experimental questions.

Methodological Issues

Our first step was to establish a valid procedure for representing and interpreting readers' stylistic assessments.

We wanted the representation of a particular subject's stylistic assessment to be as flexible as possible. We wanted to give all subjects the greatest possible amount of latitude in expressing their judgement about the style of a text. We did not want to evaluate the assessments against an a priori standard since as the validity of postulating such a standard is itself being investigated. This assumption would be implicit if the subjects were to be given scales against which their stylistic assessments would be registered (e.g., a scale for formality, concreteness, or any other stylistic quality). For this reason, a free-sort task was considered as an alternative. 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. Since the set of writing samples was controlled as much as possible so that only the style of the samples varied (e.g., they did not have great semantic differences), it was thought that the subjects' sorting arrangement would be a reflection of their perception of style in a text (measured by similarity with respect to other texts). We give the name *sorting arrangement* or *partition* to a subject's arrangement of the given writing samples.

The next task was to devise a way in which a set of subjects' sorting arrangements could be interpreted. It should be stressed that the outcome of a subject's stylistic assessment is a partition of the set of samples, not a numerical value. So, two additional issues must be addressed:

- How does one determine a value that describes how similar two partitions are to each other?
- How does one interpret that value?

Each of these issues is described in turn in this section. First, a metric for measuring the similarity between partitions is described and, second, a procedure is described for evaluating the significance of the resulting value.

A Similarity Measure for Partitions

An ideal measure of similarity between two partitions would reflect the number of primitive steps required to transform one partition into another, where primitive steps are operations such as moving or exchanging elements. The number and type of such steps correspond to a distance. The smaller the distance between two partitions, the more similar the subjective opinions that the partitions represent. A distance of zero between two partitions implies that they are identical and thus so are the underlying subjective assessments that they represent. The problem of easily determining the number of steps required to transform one partition into another has an analogous form in graph theory, where it remains an open research problem. So while this measure of similarity is ideal, it is not yet implementable. Alternative measures must be used instead. Baljko and Hirst (1999) discuss three such alternatives. None of the heuristic measures can capture the subtle nuances of differences between very similar partitions. In the general case, however, the *gamma* measure of proximity is an acceptable emulation of the ideal measure. This *gamma* measure is well established in mathematical psychology as a measurement of similarity between sorting arrangements (Hubert and Levin, 1976); it has been used in the analysis of data that is analogous to ours (Teshiba and Chignell, 1988), and therefore it was used in this experiment to measure the similarity between our subjects' sorting arrangements.

Now we wish to extend the idea of similarity between two partitions to similarity among a larger set of partitions. Measurement of a group's agreement is a function of the similarity between each of the subjects' sorting arrangements, or the inter-subject distances (ISDs). Since we are basing our measure of similarity on the proximity measure *gamma*, we can calculate the ISD_{gamma} between every pair of subjects. The mean of these ISD_{gamma}s was used as an indicator of overall agreement in a group.

Interpreting the Similarity Measures

The measure of the similarity between two partitions (or, in other words, the similarity of two subjects' stylistic assessments) can now be quantified, but it is still not known how the observed quantities can be interpreted. Assessing the significance of these ISD values is not straightforward, however. There is no direct procedure for evaluating the significance of the value. Rather, the significance of ISD values depends on their frequency distribution.

The theoretical frequency distribution of these dependent variables, the inter-subject distances, was not known, and it was not valid to assume a normal distribution due to the blatant lack of independence among the variables. To combat this problem, we developed a procedure for constructing a representative frequency distribution. (See Baljko and Hirst, 1999, for further details.) The frequency distribution, produced by performing several Monte Carlo simulations, provided a standard against which the observed ISD values could be evaluated. Each of these simulations produced thousands of analogous, meaningful, yet random sets of partitions. In order for these artificially generated simulations to be useful comparators, it was important that they be random, but only to the degree that a human would be with respect to the particular task. In other words, they should resemble the free-sorts that a human subject would make (e.g., it was observed that human subjects tended to make six plus or minus two groups, and the group sizes tended to be similar). It is not desirable to randomly select a sorting arrangement from the entire space of possibilities, which is very unlike the range of possibilities facing a human subject, given their cognitive constraints. Several simulations were required in order to account for a representative set of scenarios.

Experiment: Task Selection

As described previously, each subject was given a free-sort task. An initial pilot study was conducted to determine the length of time required to complete the task.

Subjects

Subjects were solicited by e-mail within the University of Toronto computer science research community. 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.

Materials

One set of materials was prepared, containing 24 writing samples, consisting of eight subgroups of three samples each. For each subgroup, the three writing samples consisted of a paragraph extracted from an academic paper on the mind/body problem written by a single author. 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. Due to the length of time that subjects required to complete the task in the pilot study, the testing materials were altered to contain shorter writing samples.

The subject matter of the writing samples was chosen deliberately to be opaque to a lay reader. We did not want the sorting to be based on the semantic content of the samples, and we reasoned that the semantic clues could not be easily found to assist in the sorting procedure. The pilot study confirmed the expectation that the writing samples were sufficiently difficult with respect to the required background domain knowledge.

Procedure

The subjects were each given a stack of small slips of paper, each containing a writing sample. They were instructed to sort the writing samples into piles according to their different writing style. They were told to use their own intuitive sense of writing style. They were assured that any response, ranging from one pile of 24 samples to 24 piles, each containing one writing sample, was acceptable. The likelihood of a subject producing either of the extremes is very low; 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. The subjects were allowed to take up to an hour to complete the free-sort. The experimenter then recorded the contents of each pile.

Preparation of Data

For each subject's free-sort (or sorting arrangement), the experimenter prepared a mathematical representation of the corresponding partition in a format appropriate for the software that was used in the evaluation procedures.

Experimental Results: Inter-Subject Similarity

The first goal of the experiment was to determine whether there is a high degree of similarity in the subjects' stylistic assessments. By the *gamma* measure of proximity, the mean inter-subject distance calculated for our data was 11.9273, which was found to be in the 99th percentile (i.e., no more than 1 % of the generated partitions were found to be more similar).

Experimental Results: Effect of Authorship

The second goal was to discover the extent of the influence of authorship in the subjects' stylistic opinions. We also wanted to see whether 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. We suspected that perhaps subjects would not agree with each other in terms of exactly which writing samples are similar, but those they did choose as similar would largely be by the same author.

For each subject, we determined whether the authorship of the writing samples had a significant effect on the stylistic assessment. To do this, we considered the partition based on true authorship. The influence of authorship was then determined by calculating the *gamma* proximity distance between each subject's partition and the authorship partition: the "distance-to-author" measure. Each distance-to-author measure was assessed for significance by comparison to the data from the Monte Carlo simulations. More specifically, we wanted to determine whether a subject placed writing samples in piles that were by the same author significantly more or less frequently than was attributable to chance. We were able to determine, for each subject, whether the distance-to-author distance was significantly closer than the typical distance between two subject partitions (at a .95 significance level). We found that although a subset of the subjects' stylistic assessments correlated to the authorship of the writing samples, another portion of the subjects (of roughly equivalent size) negatively correlated with the authorship of the writing samples. From this, we can conclude that the authorship of a set of writing samples is not necessarily an indicator of stylistic similarity or at least one that is consistently perceived by a diverse set of readers.

Discussion and Future Work

This exploratory study has revealed a number of interesting results. The authorship of the writing samples was found to have a significant effect on the stylistic assessments of those writing samples, but the effect was not always positive since roughly half of the subjects made stylistic judgements that were significantly similar to the authorship of the writing samples and the other half made significantly dissimilar stylistic judgements. Additionally, it was found that there was a significant amount of agreement among the subjects with respect to their stylistic assessments. One area for future analysis is to identify the subgroups of subjects with especially similar stylistic assessments. Preliminary work with cluster analysis has suggested the existence of at least two basic yet undescribed factors which may provide the basis for further classification of the data.

Future study is required, especially with respect to the syntax-based indicators of the stylistic constructs of clarity/obscurity, abstraction/concreteness, and staticness/dynamism used by DiMarco and Hirst (1993). This work should be straightforward since the stylistic constructs already have established syntactic indicators.

This exploratory study was motivated by the desire to learn about the subjectivity in the reader audience with respect to general stylistic judgement and with respect to the detection of stylistic incongruities. Although examining the subjective assessment of style was a good starting place, this work should be extended to include also assessment of stylistic problems in texts. The framework developed by Schriver (1992) for classifying problems in texts according to locus and granularity could be used initially.

Conclusion

The work described in this paper addresses issues related to computational stylistics and their application to the design of writing tools intended to eliminate problems with style in collaboratively written texts. The results of the experiment with respect to the effect of authorship suggest that there are some limitations in using authorial stylo-statistical tests to predict a reader's impression of a text's style. Additionally, sweeping predictive statements about a text's stylistic effect on a reader audience should be made cautiously since the stylistic judgements of a group of readers might not be homogeneous. Although the stylistic assessments of our subjects were found to be similar, they varied enough to show that subjectivity does exist. The subjectivity itself is an interesting property of both the texts and the readers and is deserving of further investigation.

It is hoped that this and similar future research will contribute towards the design of effective writing tools and environments, especially those targeted to collaborative writers. The results from this investigation confirm the intuition that stylistic problems cannot be treated as the pragmatic equivalents of spelling errors. A consequence of this observation should be a rethinking of the paradigm under which stylistic problems are handled. It is hoped that this work will encourage researchers in the future to consider the role of style-improvement facilities or tools within the context of the entire writing process, and, rather than developing a stylistic equivalent of a spelling corrector, to create alternative paradigms in which natural language processing techniques can be made useful to the general users of text technologies.

References

Baljko M. (1997) "Ensuring Stylistic Congruity in Collaboratively Written Text: Requirements Analysis and Design Issues," Master's thesis, University of Toronto, 1997. Also published as technical report *CSRI-365*, Department of Computer Science, University of Toronto, ftp://ftp.csri.toronto.edu/csri-technical-reports/365.

Baljko M. and Hirst G. (1999) "Determinism in Computational Stylistics," in submission.

Biber D. (1988) *Variation Across Speech and Writing*, Cambridge, England: Cambridge University Press.

DiMarco C. and Hirst G. (1993) "A Computational Theory of Goal-Directed Style in Syntax," *Computational Linguistics*, 19 (3), pp. 451–499.

DiMarco C. and Mah K. (1994) "A Model of Comparative Stylistic for Machine Translation," *Machine Translation*, 9 (1), pp. 21–59.

Ede L. and Lunsford A. (1990) *Singular Texts/Plural Authors: Perspectives on Collaborative Writing*, Carbondale, IL: Southern Illinois Press.

Edmonds P. (1998) "Translating Near-synonyms: Possibilities and Preferences in the Interlingua," in Proceedings of the AMTA/SIG-IL Second Workshop in Interlinguas, Langhorn, PA, pp. 23–30. Published in technical report *MCCS-98-316*, Computing Research Laboratory, New Mexico State University.

Glover A. (1996) "Automatically Detecting Stylistic Inconsistencies in Computer-Supported Collaborative Writing," Master's thesis, University of Toronto, 1996. Also published as technical report *CSRI-340*, Department of Computer Science, University of Toronto, ftp://ftp.csri.toronto.edu/csri-technical-reports/340.

Glover A. and Hirst G. (1995) "Detecting Stylistic Inconsistencies in Collaborative Writing," in van der Geest T. et al., (eds.), *Writers at Work: Professional Writing in the Computerized Environment*, pp. 147–168, London: Springer.

Green S. (1992) "A Functional Theory of Style for Natural Language Generation," Master's thesis, University of Waterloo, 1992. Also published as *Research Report CS-92-48*, Faculty of Mathematics, University of Waterloo.

Hartley J. (1991) "Psychology, Writing and Computers: A Review of Research," *Visible Language*, 25 (4), pp. 339–375.

Hirst G., DiMarco C., Hovy E. and Parsons K. (1997) "Authoring and Generating Health-education Documents That Are Tailored to the Needs of the Individual Patient," in Jameson A, Paris C. and Tasso C. (eds.), *User Modeling: Proceedings of the Sixth International Conference, UM97* (Chia Laguna, Sardinia, Italy), Vienna and New York: Springer Wien, New York, June 1997, pp. 107–118. [Available at http://um.org]

Hovy E. (1990) Pragmatics and Natural Language Generation, Artificial Intelligence, 43, pp. 153–197.

Hubert L. J. and Levin J. R. (1976) "Evaluating Object Set Partitions: Free-sort Analysis and Some Generalizations," *Journal of Verbal Learning and Verbal Behavior*, 15, pp. 459–470.

Kelly E. and Raleigh D. (1990) "Integrating Word Processing Skills with Revision Skills," *Computers and the Humanities*, 24, pp. 5–13.

Nold E. W. (1981) "Revising," in Frederiksen C. K. and Dominic J. F. (eds.), *Writing: The Nature, Development, and Teaching of Written Communication*, volume 2, *Writing Process, Development and Communication*, pp. 67–79: Lawrence Erlbaum Associates.

Schriver K. A. (1992) "Teaching Writers to Anticipate Readers' Needs: A Classroom-evaluated Pedagogy," *Written Communication*, 9 (2): April 1992, pp. 179–208.

Taylor T. J. (1992) *Mutual Misunderstanding: Scepticism and the Theorizing of Language and Interpretations*: Duke University Press.

Teshiba K. and Chignell M. (1988) "Development of a User Model Evaluation Technique for Hypermedia Based Interfaces," Working Paper, pp. 88–15, Department of Industrial and Systems Engineering, University of Southern California.

Melanie Baljko is a doctoral student in computer science at the University of Toronto, Toronto, Canada. In her recent master's thesis, she analyzed and formalized the causes of stylistic incongruity in collaboratively written text and is currently using these findings for the development of a prototype tool. Baljko may be reached at melanie@cs.toronto.edu and http://www.cs.toronto.edu/~melanie.

Graeme Hirst is the author of *Semantic Interpretation and the Resolution of Ambiguity* (Cambridge University Press, 1987) and many papers on topics in and related to computational linguistics. His current position is with the Department of Computer Science, University of Toronto, Toronto, Canada. Hirst may be reached at gh@cs.utoronto.ca and http://www.cs.toronto.edu/DCS/People/Faculty/gh.html.