Subjectivity in Stylistic Assessment

Melanie Baljko and Graeme Hirst

1 Introduction

1.1 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 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 this stylistic variation would be extremely useful in order to create "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 achieved by a different stylistic mechanism; and natural language generation systems that can produce texts that essentially 'say' the same thing, but are altered stylistically to be appropriate for different audiences and contexts. 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* (DiMarco and Hirst, 1993), (Green, 1992), (DiMarco and Mah, 1994). But in all of these views of computational stylistics, the stylistic assessments or descriptions are made as deterministic processes. The possible subjectivity in making stylistic assessments, such as those done by human readers, simply is not a factor in the performance of the application. Furthermore, the validity of stylistic assessments that are performed, albeit deterministically, has not yet been established due to the paucity of empirical evidence.

1.2 Eliminating Stylistic Incongruities in Collaboratively Written Text

Characterizing stylistic variation in texts is important not only for the NLP 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 an important 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 are often composed of a collection of writing styles, each corresponding to one contributor to the text; these styles are not merely inconsistent within the text, but also *incongruous* — that is, they are inconsistent to the point of detracting from the quality of the text. Readers who are familiar with collaboratively written texts often describe this problem of *stylistic incongruity* as “the text doesn’t *sound* quite right” or “the text doesn’t *speak* with a single voice.”

Our 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 and Hirst, 1995), (Glover, 1996), it was assumed that such a tool would employ a strategy of *eliminating* stylistic incongruities that are already present in a text, in a process analogous to a spelling checker. Two areas were identified as key components for a successful computational tool that would assist writers in eliminating stylistic incongruity (Glover and Hirst, 1995):

- Detection – the tool would need to discover stylistic inconsistencies; and
- Diagnosis – the tool would have to present its findings in a manner that would enable authors to correct the inconsistencies, which implies being able to articulate the problems, and any suggestion for change, in terms comprehensible to the average user.

This strategy targets *revision activity* 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. 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 sub-activity of composition, consisting of subprocesses and also interleaved within the process of composition. But in order to initiate this important activity, however, a problem within the text first must be detected. Certainly this is one area in which a computational tool could provide assistance. Stylistic incongruities are especially likely to occur in collaboratively written texts, not just because the collaborative writing process is more complex than singular writing, but because the detection of text problems as a sub-process is often distributed among the collaborators. If each author revises her own text, then they will fail to detect incongruities arising between the separately authored components of the text. Moreover, if the revision is performed later, it might not be initiated at an optimal time. At this time, writers might have difficulty diagnosing the problems and repairing them.

But the detection of stylistic problems is difficult even for human writers to master (Nold, 1981), (Schriver, 1992), (Kelly and Raleigh, 1990), (Hartley, 1991). 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).

Furthermore, 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. We know that what the audience brings to the reading process is important in terms of 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 intends. We know there is some agreement, because an audience generally agrees on what a text means (i.e., there is not "interpretive anarchy" (Taylor, 1992)), but the amount of subjectivity, nor the role it plays, is not 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 whether to repair a perceived stylistic inconsistency or not. On one hand, the author herself may belong to the group that doesn't find the inconsistency deleterious, but in fact it might bother a sufficient amount of the audience to warrant a change; or, on the other, the author may detect an incongruity, but be unsure if the expense of a change is warranted, given the small proportion of the audience that would be bothered.

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

2 Evaluating Subjectivity in Stylistic Assessments

Our experiment was designed to answer the following experimental questions[1,2]:

- Do the readers of a given text perceive its style in a similar way?
- Do the 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:

- For a given text, what would be a valid measurement of a reader's assessment of its style?
- Given a set of such *stylistic assessments*, what would be a valid measurement of the degree of their similarity?

We will describe our solution to the methodological issues in the remainder of this section, and then, in the subsequent sections, present the results that answer the experimental questions.

2.1 Measuring a Reader's Perception of Style

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* standard (e.g., a scale for formality, concreteness, or any other stylistic quality), since the validity of postulating such an *a priori* standard itself is being investigated. 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.* If 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, or the semantic differences would not be detectable by the subjects), then a subject's sorting arrangement would be a reflection of a reader's perception of style in a text (measured by similarity with respect to other texts).

We give the name *sorting arrangement* or *partitioning* to a subject's arrangement of the given writing samples.

2.2 Measuring Similarity Among Stylistic Assessments

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. But, two additional issues must be addressed: how does one determine a value that describes how similar two partitions are to each other? and 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.

An ideal measure of similarity between two partitions would reflect the number of steps required to transform one partition into another. The number and type of steps (e.g., swapping and moving elements) corresponds to a distance. The smaller the distance between two sorting assessments, the more similar the subjective opinions. A distance of zero between two partitions implies that they are identical. However, the problem of 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, and alternative measures must be used instead. In comparing two partitions $\mathcal{A}$ and $\mathcal{B}$ of a set of K writing samples, most existing distance and proximity statistics depend on categorizing the $\binom{K}{2}$ writing sample pairs in the following terms (Hubert and Levin, 1976):

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 similarity[3]; however, the single quantity m_1 determines m_2, m_3, and m_4 (Hubert and Levin, 1976).

The *gamma measure of proximity* (Γ) — which calculates m_1 — is well established in mathematical psychology as a measurement of similarity between sorting arrangements (Hubert and Levin, 1976), has been used in the analysis of analogous data (Teshiba and Chignell, 1988), and, therefore, is used in this experiment to measure the similarity between our subject's sorting arrangements.

Now we wish to extend the idea of similarity between two subjects to similarity among a group of subjects. 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 Γ, we can calculate the ISD_Γ between every pair of subjects. For a study with N subjects, there are $\binom{N}{2}$ inter-subject distances upon which to base the measurement of agreement between the members of the audience sample. The mean of the ISD_Γ's ($\overline{\mathrm{ISD}_\Gamma}$) was used as an indicator of overall agreement in a group.

Assessing the significance of the ISD values is not straightforward, however. 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 $K \times K$ $(0,1)$-incidence matrix, to be described further in section 3) 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 this procedure, the ISD is evaluated only within the context of other sorting arrangements that 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. A Monte Carlo simulation is performed by repeatedly generating analogous, yet random, sets of data (each set is generated in a pseudo-trial). In our case, each pseudo-trial generated $n = 11$ (chosen to correspond to the number of subjects in our experiment) random free-sorts, on the basis of which a mean $\mathrm{ISD}_{\Gamma i}$ was calculated. Given a large set of pseudo-trials (say $i = 1 \ldots 5000$), a frequency distribution of the $\overline{\mathrm{ISD}}_{\Gamma}$ $(\forall i)$ can be constructed and against which the observed ISD can be evaluated.[4]

For the pseudo-trials, it is important to generate random free-sorts that resemble the free-sorts that a human subject would select. 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. For this experiment, Monte Carlo simulations were performed with respect to three relevant subsets of the entire space of sorting arrangements. These subsets were identified on the basis of their similarity to the data gathered from our subjects (e.g., with the exception of one outlier, all of the free-sorts had 6 ± 2 piles, the majority of free-sorts had 6 ± 1 piles, and, additionally, the free-sorts appeared to be *balanced* — that is, the sizes of the piles were roughly the same). In Table 1, the size of the subsets and the number of Monte Carlo simulations performed with respect to each are summarized.

	Restriction for Randomly Generate Subset	**Space of Possibilities**	**Number of 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*	$\approxeq 4.0 \times 10^{14}$	$2,010$

Table 1: Summary subsets of partition data used for Monte Carlo simulations I, II, and III.

3 Experiment

3.1 Design

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 ($n = 11$) were solicited by e-mail within the University of Toronto computer science graduate student 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. The 24 writing samples consisted of 8 subgroups of 3 samples each. For each subgroup, the three writing samples each consisted of a paragraph extracted from an academic paper on 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. 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.

Each of the chosen paragraphs consisted of between 1 and 8 sentences. The set of papers was chosen to represent the writing styles of the following 7 authors: Thomas Nagel, Sydney Shoemaker, Ned Block, Frank Jackson, Colin McGinn, Jerry Fodor (3 paragraphs from each of 2 different papers), and Michael Posner. 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). We didn't want the sorting to be based on the semantic content of the samples and the hope was that 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.

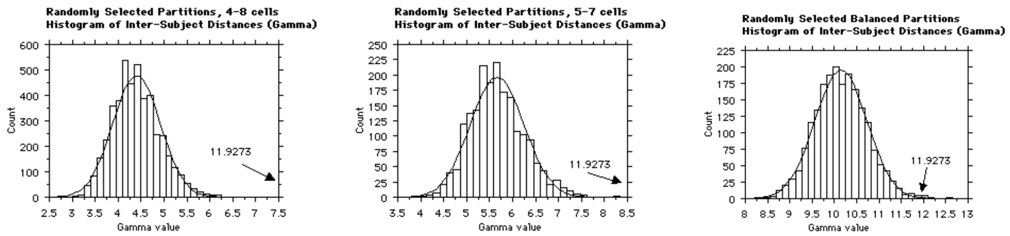
Procedure The subjects were each given a stack of small slips of paper, each slip containing a writing sample. They were instructed to sort the writing samples into piles so that each pile contained a different writing style. The subjects were told to use their own intuitive sense of writing style. They were assured that any permutation, ranging from one pile of 24 samples to 24 piles, each containing one writing sample, was acceptable.[5] 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 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 visualizing the data, the more intuitive matrix representation was used).

3.2 Results: Inter-Subject Similarity

The mean inter-subject distance calculated for our data was $\overline{\mathrm{ISD}}_\Gamma = 11.9273$. As described in section 2, a large number of pseudo-trials were generated in order to determine the significance of this number. Figure 1 presents a visualization of the distribution of the mean inter-subject distances observed in each of the pseudo-trials. As described earlier, the pseudo-trials were conducted with respect to three different scenarios of emulating a subject's free-sort. Histograms for each of the three conditions are presented. Our observed $\overline{\mathrm{ISD}}_\Gamma$ is in the 100^{th} percentile in simulations I and

II; it is in the 99^{th} percentile for simulation III. This is very convincing evidence that the amount of similarity observed among our subjects' stylistic assessments is statistically significant.



(a) Simulation I

(b) Simulation II

(c) Simulation III

Figure 1: Inter-subject agreement using a similarity measure based on the Γ distance statistic.

3.3 Results: Effect of Authorship

The second goal was to discover the extent of the influence of authorship in subjects' 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. 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.

For each subject, we determined whether the authorship of the writing samples had a significant effect on the stylistic assessment. To do this, we constructed an *authorship partition* by placing writing samples together if and only if they had the same author. The influence of authorship was then determined by calculating 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. More specifically, we wanted to evaluate whether a subject placed writing samples in piles that were by the same author significantly more or less 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, whether the subject-to-author partition distance was significantly closer than the typical distance between two subject partitions (at a .95 significance level). Although this doesn't assess the subject-to-author distance against a distribution of distances between the author partition and randomly generated partitions, it still provides a good comparison. Table 2 summarizes the results. We found that although a portion 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.

	Simulation Number		
Observed Effect	I	II	III
significant similarity	54.54%	45.45%	27.27%
none	18.18	18.18	9.09
significant dissimilarity	27.27	36.36	63.63

Table 2: A classification of the subjects' stylistic assessments (with respect to the effect of authorship)

4 Future Work

This exploratory study has revealed a number of interesting results. The authorship of 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. 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 can be explained in terms of a smaller number of factors. In this experiment, there were 276 dependent variables, each corresponding to one of the $\binom{24}{2}$ pairs of writing samples, indicating whether the elements were placed together in a pile (i.e., the value is 1) or were placed apart (i.e., the value is 0). Figure 2 shows that, with respect to two basic factors, our data fall into three clusters. This is suggestive of two fundamental, yet unknown, factors that possibly explain whether a subject would find two writing samples similar or not.

Future study is required, especially with respect to the syntax-based indicators of the stylistic constructs of clarity/obscurity, abstraction/concreteness and staticness/dynamism, created 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 not just about the subjectivity in

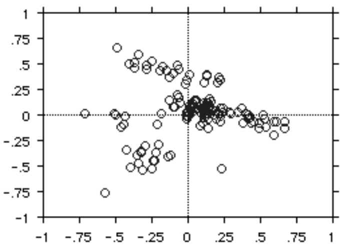


Figure 2: Factor analysis of the dependent variables describing the subject groups' stylistic assessments.

the reader audience with respect to general stylistic judgement, but 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 text. The framework for classifying problems in text (according to locus and granularity) developed by Schriver (1992) could be used initially.

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 in 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, which 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 presence of stylistic incongruities). This task produces useful similarity data, as established by Goldstone (1994), and should be considered for future work.

5 Conclusions

The work described in this paper was presented as addressing issues related to computational stylistics and its 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 stylostatistical tests to predict a reader's impression of a text's style. Additionally, 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. 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 text and the readers.

Notes

[1]Due to space limitations, not all aspects of the work can be described here; the interested reader is directed to Chapter 3 of Baljko (1997).

[2]In addition, the stylistic indicator of sentence length was explored as a possible explanation for the observed stylistic assessments, but was found to be an unsatisfactory predictor. This suggests that more sophisticated indicators of style will be required to explain perceived stylistic similarity.

[3]There are many variations of the similarity measure provided in Hubert and Levin (1976), but they are omitted here for brevity. Furthermore, a substantial discussion has been omitted here from the original work (Baljko, 1997) which describes the use of two other measures of similarity, linear correlation (ρ) and Euclidean distance (Δ). Additionally, this discussion included an analysis of the confounding present in a Δ-based measure of similarity. A data analysis was performed with respect to each of these measures, in order to compare their performances. It is not clear whether the ρ measure or the Γ measure is the better one, but they yielded very similar results and the results based on the latter are presented here, since it is an established measurement.

[4]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.

[5]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.

References

(Baljko, 1997) Melanie Baljko. *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.

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

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

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

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

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

(Glover and Hirst, 1995) 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.

(Glover, 1996) 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.

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

(Green, 1992) 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.

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

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

(Hubert and Levin, 1976) 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.

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

(Nold, 1981) 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.

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

(Taylor, 1992) Talbot J. Taylor. *Mutual Misunderstanding: Scepticism and the theorizing of language and interpretation*. Duke University Press, 1992.

(Teshiba and Chignell, 1988) 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.