# AUTOMATICALLY DETECTING STYLISTIC INCONSISTENCIES IN COMPUTER-SUPPORTED COLLABORATIVE WRITING

by

Angela Dorothy Glover

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

© Copyright by Angela Dorothy Glover 1996

AUTOMATICALLY DETECTING STYLISTIC INCONSISTENCIES IN COMPUTER-SUPPORTED COLLABORATIVE WRITING

Angela Dorothy Glover

Master of Arts 1996

Graduate Department of Education

University of Toronto

# Abstract

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

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

# Acknowledgements

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

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

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

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

To Tom Bellman for helping me run the experiment.

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

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

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

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

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

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

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

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

# Table of Contents

**Abstract**.....ii

**Acknowledgments**.....iii

**Table of Contents**.....iv

**Chapter 1. Introduction**.....1

**Chapter 2. Collaborative Writing**.....3

2.1 Why the interest in collaborative writing?.....3

2.2 Problems with computer-supported collaborative writing tools.....3

2.2.1 A matter of definition.....4

2.2.2 Towards a sociocognitive model of writing.....6

2.2.3 Collaborative writing research.....8

2.2.4 Testing collaborative writing software.....8

2.2.5 Conclusion.....9

2.3 A taxonomy of collaborative writing practices.....9

2.4 Support for the collaborative writer.....10

2.4.1 Communication.....11

2.4.2 Synchronous writing.....11

2.4.3 Roles.....11

2.4.4 Collaborator identity.....12

2.4.5 Annotation.....12

2.4.6 Version control.....12

2.4.7 Software and hardware compatibility.....12

2.4.8 A global perspective.....12

2.4.9 Format consistency.....13

2.4.10 Stylistic consistency.....13

2.5 Existing collaborative writing systems.....14

2.5.1 Aspects.....14

2.5.2 MILO.....15

2.5.3 SASSE .....15
2.6 Other computational approaches to consistency .....16
2.6.1 Consistency of format .....16
2.6.2 Consistency of writing style .....16
2.7 Conclusion .....20
**Chapter 3. Stylistics** .....21
3.1 The study of style .....21
3.1.1 History .....21
3.1.2 Models of style .....21
3.2 Stylistics .....22
3.2.1 Stylostatistics .....22
3.2.2 History .....23
3.2.3 Problems in stylostatistical research .....23
3.3 Stylostatistical research .....25
3.4 Authorial style .....25
3.4.1 Authorship attribution .....26
3.4.2 Forensic linguistics .....26
3.4.3 Authorship attribution studies .....27
3.4.4 Work-internal studies .....28
3.4.5 Conclusion .....28
**Chapter 4. Method** .....29
4.1 The detection of stylistic inconsistency .....29
4.1.1 My questions .....30
4.2 Stylistic tests .....30
4.3 Experiment .....33
4.3.1 Pilot study .....33
4.3.2 Subjects .....33
4.3.3 Procedure .....33

4.4 Tagging the data ....34
4.4.1 The taggers ....34
4.4.2 Performance of taggers ....36
4.5 Cleaning up the tagged data ....36
4.5.1 Tag manipulation ....36
4.5.2 Tagger errors ....37
4.5.3 Special problems ....38
4.5.4 Conclusion ....39
4.6 Statistical analysis ....39
4.6.1 Comparison of means ....39
4.6.2 Chi-square ....39
**Chapter 5. Analysis** ....43
5.1 Results: Matching pairs ....43
5.1.1 Two- and three-letter words ....43
5.1.2 Distribution of word length ....43
5.1.3 Tags overall ....44
5.1.4 Sentence-initial tags ....45
5.1.5 Sentence-final tags ....46
5.1.6 Comparison of means: Sentence length ....46
5.1.7 Differences between conditions ....47
5.2 Results: Non-matching pairs ....47
5.2.1 Two- and three-letter words ....47
5.2.2 Distribution of word length ....47
5.2.3 Tags overall ....50
5.2.4 Sentence-initial tags ....55
5.2.5 Sentence-final tags ....59
5.2.6 Comparison of means: Sentence length ....61

5.3 Discussion .....62
5.3.1 Summary of results .....62
5.3.2 Differences between taggers .....63
5.3.3 Conclusion .....64
**Chapter 6. Future Work** .....66
6.1 Correlation with human perception .....66
6.1.1 How can we ensure that stylometric tests identify only those documents that people perceive to be stylistically inconsistent? .....66
6.1.2 If people perceive a document as inconsistent, do stylostatistical tests identify it as inconsistent? .....67
6.1.3 Are previously-defined significance levels the best measure? .....67
6.1.4 Do salient stylistic factors need to be more consistent? .....67
6.2 Generalizability .....67
6.2.1 Does the text type affect the reliability of stylostatistical tests? .....67
6.2.2 To what extent does sample size affect the sensitivity of stylostatistical tests? .....68
6.2.3 Is my linguistic analysis valid for other writing situations? .....68
6.3 Applicability of tests .....68
6.3.1 Which tests are most useful? .....68
6.3.2 Which inconsistencies matter? .....68
6.4 Benefits to users .....69
6.4.1 Does the information about where the inconsistencies are located in the text improve people's speed or ability to merge different writing styles? .....69
6.4.2 What kind of stylistic advice is helpful to people? .....69
6.5 Technological advances .....69
6.5.1 What improvements in existing technology might provide better results? .....69

6.6 Collaborative writing strategies .....69
6.6.1 Do certain methods of collaboration result in more consistent text? .....69
6.7 Pre-defined styles .....70
6.7.1 Can stylostatistical information be incorporated into automatic style mergers? .....70
6.8 Conclusion .....70
**References** .....71
**Appendices** .....74
A Experiment documents .....78
B Tagsets .....80
C Writing samples .....82
D Tagged writing sample .....92