2782 - 7509

Heike Hänlein

Studies in authorship recognition – a corpus-based approach

PETER LANG
Europäischer Verlag der Wissenschaften

Contents

Introduction 13

1 A striking example of déjà vu 13
2 The broad theoretical framework: individual style, authorship identification and forensic linguistics 14
3 Objectives 19

Theoretical Part 23

4 Style and stylistics 23

4.1 Instead of yet another set of definitions: "fuzzy edges everywhere" 23
4.2 Individual style 30
4.2.1 The stylistician's perspective 30
4.2.1.1 Stylistic profiling and the "linguistic fingerprint" 30
4.2.1.2 The importance of reference corpora 32
4.2.1.3 Individual style and the extralinguistic context 34
4.2.1.4 Individual style and text type 36
4.2.2 The recipient's perspective: style and the notion of "prominence" ... 39
4.2.3 The author's perspective: style and "mind-style" 41

5 Essays and the *TIME Essay* – literary genre and text type 45

5.1 A brief outline of the genre: length, form, topic, "style" and related genres 45
5.2 A short history of the genre 50
5.2.1 Origins and forerunners 50
5.2.2 Montaigne, Bacon and the 16th century 50
5.2.3 From 17th to the 20th century Europe: the periodical essay 52
5.2.4 The American essay 53
5.3 The *TIME Essay*: a formal description of the text type 54
5.4 Editing 57
5.4.1 Editing news copy vs. editing opinion copy 57
5.4.2 Seven *TIME* magazines 60

6 Methodology 68

6.1 "Computer corpus linguistics": the corpus-based approach 68
6.2 Stylostatistics 71
6.2.1 Stylostatistics and authorship studies 71
6.2.2 The role of computational and statistical methods in this study 73
6.2.2.1 Chi-square testing 81
6.2.2.2 Three measures of central tendency: mean, mode and median 82
6.2.2.3 A measure of variability: standard deviation 83
6.3 The computer-assisted "cluster analysis" 83

7 The corpus 87

7.1 What's in the corpus and how it got there 87
7.1.1 Corpus size 87
7.1.2 Biographical notes on the authors 88
7.1.3 Corpus building 89
7.2 Corpus management and software 90

Empirical Part 93

8 **Empirical Part I:** A combined effort of déjà vu and statistics 93

8.1 Features present in all corpora 93

8.1.1 Subject matter – three approaches: word-frequency lists, keywords, proper nouns 93

8.1.2 The pragmatics of the essay: intention, line of argumentation and strategies of persuasion 112

8.1.3 We-groups 127

8.1.4 The use of inverted commas 139

8.1.5 Hedging: tentativeness vs. assertiveness 158

8.1.5.1 Questions 163
8.1.5.2 Modal verbs 170
8.1.5.3 Conjunctions 174
8.1.5.4 Other hedges 178
8.1.5.5 Hedging revisited 180

8.1.6 Detachment vs. involvement 182

8.1.6.1 Word length 184
8.1.6.2 Sentence length 186
8.1.6.3 Type-token ratio 189
8.1.6.4 Presence of the author (*I*): first person singular deictic pronouns and determiners 189
8.1.6.5 Reader inclusion (*WE* and *YOU*): first person plural and second person singular/plural deictic pronouns and determiners 193
8.1.6.6 Noun phrase complexity 197
8.1.6.7 Subordination vs. coordination 209
8.1.6.8 Contrastive linking 210
8.1.6.9 Frequency of quotations 211
8.1.6.10 Frequency of passive voice 212

8.1.6.11 Lexical peculiarities: hedges, fillers, generalised vocabulary, allness terms, indefinite quantifying words .. 213
8.1.6.12 Contractions 218
8.1.6.13 Summary: a brief evaluation of detachment and involvement features 218

8.2 Features prominent in one subcorpus 221

8.2.1 **Charles Krauthammer** 221

8.2.1.1 Repetition and short sentences as means of persuasion .. 221
8.2.1.2 Repetition and short sentences as text-building devices .. 236
8.2.1.3 Proper noun-common noun combinations: *the Clinton presidency* 241

8.2.2 **Barbara Ehrenreich** 246

8.2.2.1 Hyphenated words 246
8.2.2.2 Immediacy and vividness 259

8.2.3 **Michael Kinsley** 267

8.2.3.1 Repetition 267
8.2.3.2 Political keywords: *Democrat*, Republican** 271
8.2.3.3 Prediction 277
8.2.3.4 Authority, open criticism and candour 282
8.2.3.5 Recurrent expressions 285

8.2.4 **Lance Morrow** 290

8.2.4.1 Exclamatory style 290
8.2.4.2 Parenthetical and appositive style: dashes and brackets .. 292
8.2.4.3 Nominal clauses introduced by *what* 301
8.2.4.4 Comparative-contrastive style 304
8.2.4.5 Enumerative style 310

9 **Empirical Part II:** The statistical approach 315

9.1 Paragraphs per essay 315
9.2 The colon 317
9.3 The semicolon 323
9.4 Italics 327
9.5 Sentence types 329
9.6 Parts of speech 330

10 **Empirical Part III:** "Random shots" ... 335

10.1 *If*-clauses ... 335
10.2 Recurrent adjective-noun combinations (JJ * N*) ... 338
10.3 Collocational and valency analyses: *CASE, PROBLEM* ... 345
10.4 Numerals ... 350

Conclusion 361

11 Summary: four individual styles (Krauthammer, Ehrenreich, Kinsley, Morrow) ... 361

12 The *TIME Essay:* approaching a definition ... 381

13 Style as choice revisited: individual style as decisions and choices ... 383

14 Appraisal: statistics vs. intuition? ... 386

References 393

Appendix 413

A1 List of figures ... 413
A2 List of tables ... 416
A3 List of abbreviations ... 418
A4 Corpus statistics and outline ... 419
A5 Theoretical stylistic profile ... 423
A6 *COBUILD* taglist ... 425
A7 Chi-square distribution table and arithmetical example ... 426