Analyzing Writing Styles with Coh-Metrix

Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, Danielle S. McNamara
Department of Psychology
Institute for Intelligent Systems
University of Memphis
Memphis. TN 38152
{pmmccrth, ddufty, glewis1, dsmcnamr} @ memphis.edu)

Abstract

Computer scientists, linguists, stylometricians, and cognitive scientists have successfully divided corpora into modes, domains, genres, registers, and authors. The limitations for these successes, however, often result from insufficient indices with which their corpora are analyzed. In this paper, we use Coh-Metrix, a computational tool that analyzes text on over 200 indices of cohesion and difficulty. We demonstrate how, with the benefit of statistical analysis, texts can be analyzed for subtle, yet meaningful differences. In this paper, we report evidence that authors within the same register can be computationally distinguished despite evidence that stylistic markers can also shift significantly over time.

Introduction

For many years, attempts to distinguish the subtle differences between written styles were limited by the paucity of stylistic markers available to analysts. For example, word length (Brinegar 1963), syllables per word (Fucks 1952), and sentence length (Mannion & Dixon 2004), are the kind of shallow indices that have all been thought sufficient to significantly distinguish texts. Such analysis has persisted despite the inadequacies of shallow metrics being known for over 100 years (Rudman 1998). This is not to say that simple metrics do not have merit or theoretical value. For example, word length correlates with word frequency (Zipf 1947); and less frequently used words tend to be processed more slowly (Just & Carpenter 1980). However, as Holms (1998) suggests, the lack of complex statistical analysis incorporating sophisticated textual indices largely stemmed from the inadequacy of available computational power.

Over the last 20 years, the situation has much changed. Recent research in text processing, and computational advances have facilitated significant progress at differentiating textual styles. Biber (1987, 1988), for example, highlighted significant differences between text types such as narrative/non-narrative and American/British Englishes—though he was unable to satisfy his main goal of differentiating between spoken and written modes. Karsgren and Cutting (1995) adapted Biber's approach and, using a modified version of much the same corpus and lexical features, computationally distinguished *informative* works from *imaginative* ones. Louwerse et al. (2004) used the same corpus as Biber (1988), but replaced the lexical features with cohesion and readability indices, showing that spoken and written modes could be successfully separated computationally. In this paper, we build on such research by presenting an approach to computationally distinguish the works of authors within the same register.

The problem

It seems logical to assume that if modes and genres can be identified through computational indices then those same indices can disambiguate authors equally well. Unlike genres and modes, however, an author's style can vary over time (Smith & Kelly 2002). This instability leads to a problem for computational identification through style markers if researchers are armed only with limited indices. The problem, as outlined by Laan (1995), is that researchers expect a textual feature to be both static enough to distinguish one author's works from another's (e.g., Louwerse 2004), but at the same time, variable enough to indicate where in the author's career an undated text may fit (e.g., Smith & Kelly 2002). Clearly, if an index is doing one job, then it cannot do the other.

The solution to this apparent paradox was identified by Rudman (1998). Rudman argued that thousands of stylistic markers have been identified and all are potentially useful for textual discrimination. Rudman further argued that casting a wide net of markers was the best approach to identifying authorship, rather than simply searching out an elusive set of indices and applying them equally to all analyses. In this paper, therefore, we avoid Laan's paradox by adopting Rudman's approach, and use the widest net of indices available on a single computational tool.

Introducing Coh-Metrix

Coh-Metrix is a computational tool that provides over 200 indices of cohesion, difficulty and readability (Graesser et al. 2004). Coh-Metrix is sensitive to a wide

range of deep levels of textual features that reflect cohesion relations, world knowledge, and language and discourse characteristics. Coh-Metrix accomplishes its task through a variety of modules, including: syntactic parsers (Charniak 2000); latent semantic analysis (LSA Landauer 1997); and many other features common in computational linguistics (Jurafsky & Martin 2000). Coh-Metrix also provides researchers with a range of traditional indices such as average word length, average sentence length, and the readability formulas of *Flesh Reading Ease* and *Flesch-Kincaid Grade Level* (Klare 1974-1975).

As spatial restrictions prevent us from a major discussion of all Coh-Metrix indices, we present only a summary of Coh-Metrix's key indices. An extensive overview and analysis is provided in Graesser et al. (2004).

Causal Cohesion. Coh-Metrix calculates causal cohesion as the ratio of causal verbs to causal particles. Causal verbs such as *kill*, *throw*, and *drop* are identified through WordNet (Fellbaum 1998; Miller et al. 1990). Causal particles are identified in a pre-defined set and include items such as *because*, *as a consequence*, and the semantically depleted verbs *make* and *cause*.

Coreferential Cohesion. Referential links aid textual comprehension, facilitating inferencing, and benefiting recall (Kintsch and van Dijk 1978; McNamara 2001). Coh-Metrix employs four forms of lexical coreference identification: noun overlap, argument overlap, stem overlap, and LSA-based semantic overlap. The overlap measures focus on comparing lexically based pairs such as *table*/*tables* and *run/running*. The LSA measures, on the other hand, employ singular value decomposition, a statistical technique, to analyze the semantic relationship between various textual elements. As such, LSA allows us to extend referential overlap beyond explicit relations such as *chair/chairs* into relative semantic similarities such as *chair/table*, *table/wood*, and *wood/grass*.

Connectives and Logicial Operators. Connectives form cohesive links between separated sentential ideas. Coh-Metrix reports the density of connectives in various ways. For example, there are scores for positive-additive connectives (e.g., *also moreover*), negative-additive connectives (e.g., *however, but*), positive-temporal connectives (e.g., *after, before*), and negative-temporal connectives (e.g., *until*). Connectives serve as an extremely important indication of cohesion in a text (Haliday & Hasan 1976; Louwerse 2002; Graesser et al. 2004). In addition, scores reflecting the density of logical operators such as *or*, *and*, and *not* are also reported. High densities of such items in a text place a high demand on the working memory of the reader.

Density of Major Parts of Speech. Coh-Metrix reports the incidence scores for various parts of speech (POS) as defined by the Brill (1995) POS tagger. These parts include pronouns, nouns, verbs, adjectives, adverbs, cardinal numbers, determiners, and possessives. Density scores help to detect textual difficulty with, for example, a higher proportion of pronouns generally leading to a greater cognitive strain on the reader caused by more referential bridging (Graesser et al. 2004).

Polysemy and Hypernymy. Coh-Metrix tracks the ambiguity and abstractness of a text by incorporating WordNet (Fellbaum 1998) to calculate values for lexical polysemy (number of senses) and hypernymy (number of levels in a conceptual, taxonomic hierarchy).

Syntactic Complexity. The measure of syntactic complexity assumes that sentences with embedded constituents are either structurally dense, syntactically ambiguous, or ungrammatical (Graesser et al. 2004).

Word Information and Frequency. Word information incorporates four matrices: familiarity, concreteness, imageability, and meaningfulness. Coh-Metrix derives scores for these aspects via the MRC Psycholinguistic database (Coltheart 1981). High frequency words are those that are used more often (either in speech or writing) and are therefore likely to be more easily understood and read faster (Haberlandt & Graesser 1985; Just & Carpenter 1980).

Other indices. Coh-Metrix includes a wide variety of shallow, traditional indices such as *syllable count*, *word length*, *sentence length*, *number of words per sentence*/*paragraph*/*text*, and various combinations of these such as Flesch Reading Ease (Klare 1974-1975). Since many of these measures have been previously used in stylistic analysis (e.g., Brinegar 1963; Fucks 1952), we opted to use only the new and theoretically more interesting indices available in Coh-Metrix. Traditional indices have been, and will remain, important indicators of mode, genre, and style, but our goal is to demonstrate that other features of language, such as cohesion, can also play a significant role in distinguishing authorial styles.

Methods

The traditional corpus for testing methods of authorial distinction is the 12 disputed Federalist Papers (e.g., Holmes & Forsyth 1995; Mosteller & Wallace 1964).

However, in this study, the Federalist papers would be inappropriate as such a corpus would only address half of Laan's paradox. That is, the Federalist Papers presents us with a synchronic problem (authorship attribution on the grounds of stable stylistic traits). The time line of such a corpus would not allow us to examine whether diachronic variation occurs. As such, we followed Stamatatos, Fakotakis, and Kokkinakis (2001) who understood that

demonstrating new approaches and new tools often demands piloting procedures on fairly mundane corpora. For our corpus, therefore, we compiled the freely available works of three well-known authors: Rudyard Kipling, Charles Dickens, and P.G. Wodehouse (see Table 1). The choice of authors was motivated by three considerations. First, the combined works were sufficiently diverse in terms of style that a reasonable system for stylistic differentiation (in this case, Coh-Metrix) should find differences. But second, the works were sufficiently similar that distinguishing between them would not be a trivial test. Third, the works by these authors covered a large time span (up to 50 years). This means that for each author, differences in certain style markers are likely to have developed, providing an opportunity to demonstrate Coh-Metrix in light of Laan's paradox: finding measures that are sensitive to differences between authors as well as stylistic changes for a single author.

Author	No. of texts	Years
Kipling	64	1886-1934
Wodehouse	21	1901-1923
Dickens	29	1833-1864

Table 1. Details of the texts used in this study.

For length of text to analyze, we followed the widespread practice of selecting random, continuous, sentence-beginning to sentence-end chunks of text, each of about 2000 words, from each work (e.g., Biber 1988; Louwerse et al. 2004; cf. Burrows 1987). All texts were then processed through Coh-Metrix.

Results

To determine stylistic differences, we followed many previous studies in the field by conducting a discriminant function analysis (Biber 1993; Karlgren & Cutting 1994; Ledger & Merriam 1994; Mealand 1995; Stamatatos et al. 1999). A discriminant function analysis produces a weighted linear equation for each category, in this case authorship. We estimated that with the current dataset, four indices would be the maximum number of predictors available before problems with overfitting occurred. However, there were many more than four indices available through Coh-Metrix. To fully explore the range of available variables, we allowed the data itself to select the most appropriate variables.

An analysis of variance was conducted on the Coh-Metrix indices to give a broad overview of which variables produced large difference between authors. Of these, 186 produced a significant effect at the .05 level or higher. All indices were then ranked by effect size. One of the assumptions of discriminant function analysis is that the predictor variables are not highly correlated. For this reason, if the correlation between any two variables was r=>.7, then the variable with the weaker univariate relationship was removed. The four predictors yielded from this process were (in decreasing order of univariate effect size): *number of higher level constituents per word*; *minimum word imageability per paragraph*; *wh-determiner incidence score*; and *incidence of conditionals.*

A discriminant function analysis was then conducted with author as the dependent variable. To provide an objective test of the analysis, we separated 24 texts, 8 from each author, from the main dataset (hereafter, the *test set*). The 90 texts remaining in the main dataset, we will hereafter refer to as the *training set.* The Structure matrix with the coefficients for each function for each variable is shown in Table 2.

Variable	AUTHOR		
	Kipling	Wodehouse	Dickens
HLCW	531.94	469.81	377.12
MWIP	1.90	1.72	1.78
IWD	-0.19	-0.36	0.49
IC	0.11	-1.27	0.20
(Constant)	-280.16	-223.61	-227.83

Table 2. Structure of the Discriminant Functions for Number of Higher Level Constituents per Word (HLCW), Minimum Word Imageability per Paragraph (MWIP), Incidence of wh-determiners (IWD), Incidence of Conditionals (IC), and Constant

Higher Level Constituents per Word. A word with more higher level constituents is a more specific word (e.g., *swallow* has four higher level constituents: *bird, animal, living entity, entity*). The results suggest that Kipling has a propensity for more specific words when the other variables in the analysis are also taken into account.

Minimum Word Imageability per Paragraph. A higher score for imageability indicates a more vivid use of words. In the context of all five variables, this result suggests that Wodehouse may tend to use words that are less easily pictured in the mind.

Incidence of wh-determiners. For wh-determiners (e.g., the *which* of sentences such as *Do you know which bus I should take?*), Dickens is weighted as being the most likely author to use such constructions, and Wodehouse the least likely. A cursory examination of some of Dickens' novels supports such a finding. For example, Dickens appears to write *which* ahead of *that* whenever a choice is available. Dickens also appears to write *which* in less than common phrases, e.g., in Chapter 2 of Oliver Twist, Dickens writes "any one of which cases."

Incidence of Conditionals. The structural equations show that texts authored by Wodehouse are least likely to have a higher incidence of conditionals. This may indicate that Wodehouse is less likely to use causal relations in his writing.

Accuracy

An estimation of the accuracy of analysis can be made by plotting the correspondence between the actual author and the predictions made by the discriminant analysis (see Table 3). The diagonal numbers represent the frequency with which an author was correctly identified. Conversely, the off-diagonals represent the number of incorrect attributions of authorship. The results show that the discriminant analysis correctly allocated 79 of the 90 texts, an average accuracy rate of 88%. This figure, however, may be slightly inflated by the data of training set. Using the test set data alone, the accuracy of author allocation was 79%. This second figure, although lower, is still a remarkably high level of accuracy, and demonstrates that the discriminant functions are robust.

Actual author	Predicted author		
Training set	Kipling	Wodehouse	Dickens
Kipling	**12**	1	0
Wodehouse	1	**47**	8
Dickens	0	1	**20**

Test set	Kipling	Wodehouse	Dickens
Kipling	**6**	0	2
Wodehouse	0	**7**	1
Dickens	0	2	**6**

Table 3. Predicted author versus actual author featuring results from both the training set and the test set.

The precision, recall, and F1 scores for each author further demonstrate the accuracy of the model (see Table 4). While the overall accuracy was lower for Dickens, an F1 score of .71 still offers evidence that reasonable accuracy was obtained in correctly identifying Dickens' texts.

Author	Precision	Recall	F1
Kipling	0.75	1.00	0.86
Wodehouse	0.88	0.78	0.82
Dickens	0.75	0.67	0.71

Table 4. Precision, recall and F1 measures for all three authors.

The success of the discriminant analysis lends support to the wide-net approach advocated by Rudman (1998). Specifically, some stylistic variables may not be indicative of a particular author because style over career may change. As such, one approach to distinguishing authorship is to use a very large number of variables and not assume that any one variable will always be successful. However, while the results so far support Rudman's (1998) claim, the issue as to whether chronological shifts in indices also occurs (as is presumed, for example, by Smith & Kelly 2002) has not yet been fully addressed. As such, we assessed possible chronological variable shift by reanalyzing the texts with bivariate correlations. Using year of publication as the dependent variable, we ran correlations using sets of indices for cohesion, parts of speech, and difficulty.

Cohesion Features. The correlations between measures of cohesion and year of publication are shown in Table 6. In terms of argument overlap, Kipling showed a significant increase in cohesion across years, Wodehouse showed a significant decrease in cohesion across years, and Dickens showed a moderate increase. Although LSA has previously indicated significant trends in cohesion (e.g., Foltz 1998), in this analysis, we found no significant correlations with year for any of the three authors. This may indicate that while an author's use of specific words may become more repetitive (Kipling) or more diverse (Wodehouse), there is no evidence for the emergence of a similar pattern in semantically related terms.

Measure	Kipling	Wodehouse	Dickens
AOallsens	0.52*	-0.36**	0.31
AOadjsens	0.55**	-0.37**	0.40*
LSAallsens	0.42	-0.15	0.10
LSAadjsens	0.39	-0.22	0.30

*significant at .05 level, ** significant at .01 level

Table 5. Correlations between cohesion measures and year for three authors: Argument overlap between all sentences (AOallsens), Argument overlap between adjacent sentences (AOadjsens), LSA similarity between all sentences (LSAallsens), and LSA similarity between adjacent sentences (LSAadjsens)

Lexical Features. The correlations for part of speech measures are shown in Table 6. Several indices capture part-of-speech information by measuring the rate of occurrence of a particular part of speech per thousand words. While Wodehouse appears to have been remarkably consistent in his use of these features, however, Dickens and Kipling both showed significant trends across years for

several part-of-speech measures. With *pronoun use*, for example, the works of both Dickens and Kipling see a significant increase in the frequency of this feature. Such a trend in the use of explicit referential links is consistent with the increase in cohesion of their works indicated in Table 5.

Measure	Kipling	Wodehouse	Dickens
LOIS	-0.27*	0.20	0.41*
PEIS	0.31*	-0.00	-0.52**
PIS	0.36**	-0.00	0.42*
PPIS	0.33*	0.10	0.49*

*significant at .05 level, ** significant at .01 level

Table 6. Correlation between year and parts of speech for the three authors: Logical operator incidence score (LOIS), Possessive ending incidence score (PEIS), Pronoun incidence score (PIS), and Personal pronoun incidence score (PPIS)

Difficulty Measures. One important and widely used measure of difficulty is the number of syllables per word. Longer words and lower frequency words are both known to be more difficult to process. Although these variables are correlated, they are known to have independent effects on reading difficulty (Haberlandt & Graesser 1985). The correlations for both of these measures with year are shown in Table 7 for each of the three authors. Kipling used lower frequency words as his career developed, whereas Dickens used higher frequency words (or fewer low frequency words). Wodehouse showed no significant change. These trends are mirrored in the average syllable length per word. Dicken's words became shorter, Kipling's became longer, whereas Wodehouse did not demonstrate a strong trend in either direction.

Measure	Kipling	Wodehouse	Dickens
CFpw0-6	-0.42**	0.10	0.41**
ASL	0.41**	0.10	-0.58**

*significant at .05 level, ** significant at .01 level

Table 7. Correlations between text difficulty measures and year for three authors. Celex frequency per word (range, 0-6) (CFpw0-6), Average Syllable Length (ASL).

Discussion

The results of this study offer support to the contention that various authorial style markers may shift significantly throughout the career of an author (Laan 1995; Smith & Kelly 2002). More specifically, these markers affect different authors, may shift in either direction, and can affect measures as different as cohesion, parts of speech, and difficulty. This is in keeping with Rudman's (1998) claim that each attempt to distinguish authors may require different markers. Consequently, Rudman advocated the approach to textual analysis that was adopted in this study: use a wide range of variables and allow significant distinguishing features to emerge. The large number and variety of indices made available by Coh-Metrix allowed the Rudman style of analysis to be conducted. The results suggest that distinguishing authorship by characteristic authorial style is achievable.

A computational approach to distinguishing texts offers researchers and educators a number of exciting avenues of interest: For example, it allows us the possibility of better estimating the creation of undated works. It allows us to better settle issues of authorship and cases of fraud. It allows computer text mining systems to predict text types so that parsers and taggers can make better predictions of syntax and parts of speech. It presents the possibility that student writers might be able to assess their works in progress so as to better understand the characteristics of the style they are developing. And it allows the possibility that the appropriateness of any given text to its audience may be more easily assessed.

Our future research will pursue these notions by analyzing texts of disputed authorship, looking for greater indication of characteristics of mode, genre, and register; and investigating characteristics of authorial style across grade levels. While more research is needed, this study contributes to the field by showing that a computational tool, Coh-Metrix, has the capacity to distinguish below the level of the register and into individual authorship characteristics.

Acknowledgements

This research was supported by the Institute for Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Biber, D. (1987). A textual comparison of British and American Writing. *American Speech, 62*, 99-119.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 1–15.

Brinegar, C.S. (1963). Mark Twain and the Quintius Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association, 58*, 85-96.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study

in part-of-speech tagging. *Computational Linguistics, 21*, 543–565.

Burrows, J. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing, 2*, 61–70.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine*, *18*, 33-44.

Charniak, E. (2000). A Maximum-Entropy-Inspired Parser. *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics.* (pp. 132-139). Seattle, WA.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Fucks W. (1952). On the mathematical analysis of style. *Biometrica*, *39*, 122-129.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*, 193-202.

Haberlandt, K., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology*, *114*, 357-374.

Haliday, M. A., & Hasan, R. (1976). *Cohesion in English*. London: Longman

Holms, D.I. & Forsyth, R.S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing, 10*, 111-127.

Jurafsky, D. S., & Martin, J. H. (2000). *Speech and language processing*. Englewood, NJ: Prentice Hall.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329-354.

Karlsgren J. & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. *International Conference on Computational Linguistics Proceedings of the 15th conference on Computational linguistics - Volume 2*. Kyoto, Japan: 1071-1075.

Kessler, B., Nunberg, G., & Schuetze, H. (1997). Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*. Madrid: Morgan Kaufmann Publishers, 32–38.

Klare, G. R. (1974-75). Assessing readability. *Reading Research Quarterly*, *10*, 62-102.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Ledger, G. R., & Merriam, T. V. N. (1994). Shakespeare, Fletcher, and the two noble kinsmen. *Literary and Linguistic Computing*, *9*, 235-248

Louwerse, M.M. (2002). An analytic and cognitiveparameterization of coherence relations. *Cognitive Linguistics*, *12*, 291-315.

Louwerse, M. M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities*, *38*, 207-221.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

McNamara, D. S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, *55*, 51-62.

Mealand, D. L. (1995). Correspondence analysis of Luke. *Literary and Linguistic Computing*, *10*, 171-182.

Morton, A. (1965). The authorship of Greek prose. *Journal of the Royal Statistical Society Series A*, *128*, 169–233.

Mosteller, F., & Wallace, D. (1984). Applied bayesian and classical inference: The case of the Federalist papers. MA: Addison-Wesley, Reading.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, *31*, 351-365.

Smith, J. A., & Kelly, C. (2002). Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works. *Computers and the Humanities*, *36*, 411–430.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (1999). Automatic extraction of rules for sentence boundary disambiguation. *Proceedings of the Workshop on Machine Learning in Human Language Technology, ECCAI Advanced Course on Artificial Intelligence (ACAI-99)*, 88–82.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*, 471-495.