

Changes in Scientific Articles Over Two Hundred Years: A Coh-matrix Analysis

Michell Bruss
Department of English
University of Memphis
Memphis TN 38152

Michael J. Albers
Department of English
University of Memphis
Memphis TN 38152
901-678-4776

Danielle McNamera
Department of Psychology
University of Memphis
Memphis TN 38152
901-678-2326

malbers@memphis.edu

d.mcnamara@mail.psync.memphis.edu

ABSTRACT

We analyzed texts from years 1800-2004 from the *Philosophical Transactions of the Royal Society of London*. Two-thousand-word sections from about 20 articles published at 25-year intervals (1800, 1825, 1850, etc.) for a total of 127 articles were analyzed by a new tool (Coh-matrix) developed by McNamera, Louwerse, and Graesser [9] at the University of Memphis' Institute for Intelligent Systems. The study discerned significant differences in four general measurement areas: word information, connectives, causal cohesion, and syntactic complexity. Specifically, there was a significant decrease in concreteness, imagability, number of causal verbs, number of causal particles, number of connectives (including total number of connectives, and positive temporal and causal connectives), and the mean number of higher-level constituents per sentence and per word. We also found a significant increase in age of acquisition, syntactic complexity (measured in mean number of modifiers per noun phrase), and indicators of analytical and logical difficulty.

Categories and Subject Descriptors

A.m MISCELLANEOUS

General Terms

Documentation, Measurement

Keywords

historical research, journal articles, computational linguistics

1. INTRODUCTION

Recent studies have just begun to apply quantitative analysis to scientific texts. Most famous of these studies is perhaps Bazerman's study of articles within a particular discipline of physics[1], in which he used both quantitative and qualitative analyses to examine the development of spectroscopic articles.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGDOC'04, October 10-13, 2004, Memphis, Tennessee, USA.
Copyright 2004 ACM 1-58113-809-1/04/0010...\$5.00.

Rhetorical analyses of these texts and those from other disciplines include works on biology articles [10], tax documents [5], and even citations [4]. These studies have been limited in their application of quantitative methods, in part by a resistance on the part of technical communication researchers to using such methods, and in part because the logistics involved in attempting such analyses were practically insurmountable. With recent developments in technology, however, these logistical difficulties are no longer such a formidable issue.

Objections to quantitative analysis of texts have rested both on technological and semantic grounds; anyone familiar with the whims of the grammar checker is well aware that current technology leaves much to be desired, while purists object that any content analysis must necessarily be descriptive and not exploratory in nature. Studies, such as the analysis of a grammar checker by Wei and Davies [16] statistically emphasize the shortcomings of the most commonly applied parsers, while theoreticians such as Kracauer [11] elucidate what they see as a flawed method fraught with data that is misleading.

These detractors have a point. Quantitative analysis of texts has been traditionally flawed, as computers and databases have previously been unable to appreciate the variety and subtlety of the written language. However, simply because a computer cannot apprehend precise methodologies in writing, and can misapprehend phraseology is no reason to discard the method entirely. Guerin-Pace [7] proposed that textual statistics be utilized by researchers relying on interviews; the computer-generated statistics could readily be defined by researchers to approximate their reasoning. That is, researchers who use interviews or examine through large amounts of texts can use these tools to simplify their work. Charney [3] provided an excellent defense of textual empiricism, assailing questions about both the ethical responsibility and the social veracity of the practice; he contends that, while researchers need assess the impact of their work, empiricism itself should not be confused with ideology, and that the practice of quantitative evaluations of texts has given us a wealth of knowledge while suffering from no more shortcomings than any other method. Conversely, Mitchell [13] asserted that content analysis was not only a viable descriptive method, but should be utilized for exploratory studies as well.

Bazerman was one of several researchers that took up Mitchell's challenge. In several of Bazerman's texts[2], we find a researcher utilizing content analysis and even computers in order to obtain

statistics that might illuminate large-scale trends in the writings he examines. For example, Bazerman [1] analyzed the rhetorical content and design of texts from the *Philosophical Transactions of the Royal Society* in order to deepen his understanding of social influences and the emergence of the scientific article. While this experiment did not specifically employ common methods of quantitative analysis, the method of rhetorical analysis Bazerman employs is certainly empirical in nature, and sets the stage for subsequent articles employing less subjective methods.

Bazerman utilized both quantitative and qualitative methods in his examination of the development of the spectroscopic article: "Modern Evolution of the Experimental Report in Physics: Spectroscopic Articles in *Physical Review*, 1893-1980" [1]. Computer-based analysis at that time left something to be desired, so Bazerman used it in combination with close analytical reading: "using statistics to indicate gross patterns or trends but using close analytical reading to explore the finer texture, the meaning and implication of those trends" (p. 63). In his paper, Bazerman examines full articles depicting first-hand accounts of spectroscopic experiments, focusing on changes in length, use of references and graphics, organization and mode of argument. From utilizing both quantitative and qualitative approaches, Bazerman notes a correlation between increase in article length in direct proportion to the increased use of references and graphic elements. Further, he notes the change in organization from that of merely observational records to planned experiments focusing on supporting a particular change.

Others have followed this successful tactic. Myers [10] employed content analysis in his "Stories: Styles in Two Molecular Biology Review Articles." Devitt [5] applied a similar methodology to tax accounting documents in "Intertextuality in Tax Accounting: Generic, Referential, and Functional."

There are several reasons why quantitative studies of texts are important to the field of technical communications, not the least of which is the necessity of producing a respectable body of empirical literature to aid in the legitimization of the profession. However, this is not the sole reason to undertake such studies. Descriptive empirical pursuits such as these help technical communicators and scientists to understand the ways in which their fields have changed. Combined with rhetorical and social analysis, studies such as these are invaluable for the improvement of the profession as a whole, and allow the individual communicator to better recognize trends in their respective disciplines and adapt accordingly.

The experiment hereafter recorded takes a different tact, employing a new tool in a quantitative analysis of texts. In this experiment, 127 texts from the pages of the *Philosophical Transactions of the Royal Society of London* were quantitatively analyzed by an emergent parser, the Coh-matrix, in the hopes of discovering trends in cohesion across the sample's period of 197 years. The aim was not to prescribe definite trends, or analyze social, political, and developmental factors influencing the writing on these pages, but to identify broad changes occurring across the texts presented in this journal. Validation of the tool was also not an experimental goal, although there is, of course, some element of evaluation in this process. This tool, like any other, is subject to multiple flaws, as much from its nature as from its infancy. Thus, the quantitative analysis here undertaken is merely a stepping point from which expanded analyses may emerge.

2. PROCEDURE

The texts analyzed for this research came from the *Philosophical Transactions of the Royal Society of London*, and were taken from journals printed between 1800 to 1997. This journal reprints articles that have been read before the Royal Society of London, the premier meeting ground for scientific study in Great Britain. According to the Society, these articles are chosen by "the importance and singularity of the subjects, or the advantageous manner of treating them; without pretending to answer for the certainty of the facts, or propriety of the reasonings, contained in the several papers so published, which must still rest on the credit or judgement of their respective authors." This choice was driven as much by availability as appropriateness: not only is it perhaps the only journal to have over two hundred years of articles stored on the JSTOR database, but it also has representations of scientific articles from numerous disciplines. The journal consists of articles of observational, theoretical, and experimental natures, which comprise the breadth of the scientific writing spectrum. Although the early articles are primarily observational in nature, the articles progress to include experimental and theoretical articles, mirroring the scientific trend towards empirical research and the evaluation of theories.

Out of these journals, we chose the first twenty-four articles from each year in twenty-five year increments, beginning in 1800 and ending in 1875. In 1887, the journal split into two publications, called Series A and B. Series A deals with papers concerning mathematics and physics, while Series B is concerned with biology. After this split, beginning in 1900, we took the first ten articles from each Series, initially omitting appendices, articles under 2500 words, and articles that consisted primarily of lists or mathematical equations. These chosen articles were downloaded in TIFF or PDF format and converted, using an OCR reader, into plain-text documents in order to accommodate the limitations of the Coh-matrix tool. Since OCR readers are notoriously unpredictable in their conversions, we then did a line-by-line edit to verify that the articles were properly converted.

A 2000 word sample was extracted from each text. All graphic elements, including tables, charts, and graphs were discarded. The first 500 words of each text were discarded to the closest paragraph in an effort to eliminate introductory language and set each article on equal stylistic footing. Abstracts and lists were not excluded from this initial count. Next, all paragraphs containing lists, separate formulae, and embedded tables were eliminated. Formulae that occurred within a paragraph and did not significantly affect comprehension of the paragraph were left. The paragraphs were removed to accommodate limitations of the parser in Coh-matrix. Then a second word count was done, and everything after the closest paragraph to 2000 words was eliminated. Articles that failed to measure 2000 words after the graphic elements and lists were removed and discarded from the experiment. Subsequent articles from the initial pool of twenty-four articles were added as substitutions. This gave us texts with a range from 2000 words to 2300 words.

A second cleansing of the texts was undertaken to ensure compatibility with the Coh-matrix. The tool is unable to distinguish between end-of-sentence punctuation and abbreviation markers. Therefore, all periods and decimals were removed from the texts, without disturbing the text surrounding them. All in-text citations in parentheses were removed, while phrases purely informational in nature were left, minus the punctuation. In this

way, we were able to sanitize the texts to accommodate the program while limiting textual manipulation as much as possible. Foreign words such as Latin phrases and French quotations were left intact, and pains were taken to ensure their proper spelling. British spellings were also left unchanged, as were scientific terminology and the names of scientists.

Next, each article was run through the Coh-matrix tool and analyzed. The tool returns over two hundred and fifty measures. The average measures for each year were then plotted on line graphs in order to recognize any significant trends in the measurements over time. Of these measurements, it was determined that only a minor number showed significant trends. This minority was then subjected to a two-tailed Pearson Correlation, and the results are provided in the data section.

3. COH-MATRIX MEASURES

Coh-matrix provides over 250 measures [9], divided into several disparate sections: word, sentence, and paragraph levels; word information, including familiarity and concreteness; word frequency; polysemy and hypernym values; concept clarity; connectives; syntactic complexity; readability; core cohesion; word and phrase density; causal cohesion; core cohesion; density of logical operators; latent syntactical analysis (LSA) space assessments; type-token ratios; and measures of parts of speech. Of these, significant correlations were found in only causal cohesion values, incidence of connectives, syntactic complexity, and word information. Of these segments, significant figures were found in only four groups:

- Word information
- Causal cohesion
- Connectives
- Syntactic complexity

Word information values measure word familiarity, concreteness, imagability, Colorado meaningfulness, Pavo meaningfulness [14], and age of acquisition. Causal cohesion values measure the number of causal particles in the text, the number of causal verbs in the text, and the ratio of causal particles to causal verbs. Connectives values measure the incidence of clarification connectives, additive connectives (positive and negative), temporal connectives (positive and negative), causal connectives (positive and negative), and the total number of all connectives. Syntactic complexity values measure noun-phrase density, high-level constituents per sentence and per word, and incidence of word classes that signal logical or analytical difficulty.

4. RESULTS

There were 127 articles included in this statistical analysis. The values were analyzed for mean, minimum and maximum values, and Pearson correlation.

4.1. Word Information

Significance was found in the measure of 5 values out of 40 different word information measures.

Measure (Coh-matrix variable)	Mean (SD)	Min (year) Max (year)	Pearson correlation
Age of Acquisition, mean for content words (WORDAacw)	382.4 (22.1)	326 (1900) 436 (1997)	0.341 (p < .001)
Imagability, mean for content words (WORDIacw)	387.4 (17)	338 (1997) 430 (1825)	-0.283 (p = .001)
Age of Acquisition, mean for all words (WORDAaw)	382.2 (22.1)	325 (1900) 436 (1997)	0.344 (p < .001)
Concreteness, maximum in sentence for content words (WORDCxcw)	528.1 (40.3)	446 (1925) 611 (1800)	-0.540 (p < .001)
Concreteness, maximum for all words (WORDCxes)	529.4 (39.4)	455 (1900 and 1950) 612 (1800)	-0.568 (p < .001)

4.2. Causal Cohesion

Significance was found in the measure of 1 value out of 3 different causal cohesion measures.

Measure (Coh-matrix variable)	Mean (SD)	Min (year) Max (year)	Pearson correlation
number of causal particles in the text (CAUSP)	25.2 (7.3)	12.6 (1997) 58.2 (1850)	0.372 (p < .001)

Included in the analysis but not found to be statistically significant were the number of causal verbs in the text and the ratio of causal particles to causal verbs. The number of causal verbs in the text shows a slight, but statistically insignificant, negative correlation with the year. Oddly, in light of the stronger negative correlation for the number of causal particles in the text, the ratio between causal particles and causal verbs also showed a statistically insignificant negative correlation.

4.3. Connectives

Significance was found in the measure of three value out of eight different connectives measures.

Measure (Coh-matrix variable)	Mean (SD)	Min (year) Max (year)	Pearson correlation
Incidence of all of the connectives (CONI)	79.1 (12)	43.66 (1997) 119.70 (1850)	-.237 (p = .007)
Incidence of positive temporal connectives (CONTPPI)	10.4 (4.2)	2.04 (1975) 22.16 (1825)	-.338 (p < .001)

Incidence of positive causal connectives (CONCSPI)	20.2 (6.4)	8.90 (1950) 44.35 (1850)	-.375 ($p < .001$)
--	---------------	-----------------------------	-------------------------

4.4. Syntactic Complexity

Significance was found in the measure of all four values of syntactic complexity.

Measure (Coh-matrix variable)	Mean (SD)	Min (year) Max (year)	Pearson correlation
Mean number of modifiers per noun phrase (SYNNP)	0.683 (0.1)	0.7 (1800) 1.4 (1997)	.683 ($p < .001$)
Mean number of higher-level constituents per sentence (SYNHs)	5.4 (1.7)	2.97 (1975) 10.68 (1825)	-.716 ($p < .001$)
Mean number of higher-level constituents per word (SYNHw)	0.15 (.017)	.197 (1800) .110 (1925)	-.300 ($p = .001$)
Incidence score (per 1000 words) of indicators of analytical and logical difficulty (SYNLOGIC)	36.2 (7.7)	24.54 (1997) 63.92 (1850)	0.216 ($p = .015$)

5. DISCUSSION

These results indicate that the texts mirror some of the trends identified by Bazerman [1] in his review of spectroscopic articles. That is, the texts have become increasingly abstract, using more complicated subjects and less sentence constituents, while decreasing the overall use of connectives. These results indicate that as the texts have become more theoretically focused and the style of writing has changed towards a less literary and more persuasive bent. The statistics below are identified in order, along with an explanation of how each supports this hypothesis.

5.1. Word Information

Concreteness is based on human ratings of whether words are more or less abstract or concrete. These values were also obtained from the MRC Psycholinguistics Database. The strongest negative correlation values are found in the number of higher-level constituents per sentence, the maximum concreteness in sentences for content words, and the maximum concreteness for all words. The strong decrease in maximum concreteness for content words (WORDCxc) indicates that the main sentence components have become more abstract over the 197-year period measured. The strong decrease in maximum concreteness for all words supports this conclusion.

The age of acquisition values reflect the determined age that some words appear in the speech of children based on the work of Gilhooly and Logie [8]. The strongest positive correlation values are found in the age of acquisition and number of noun-phrase modifiers. This indicates that the people learned the words used in

texts at a later ages in the recent past than in the 1800s. The positive correlation values of the mean age of acquisition for both content words and all words indicate a rise in the difficulty of words chosen for the texts. This may be a reflection of the increased use of jargon and the advancement of a particular scientific language; an evolution that has shaped the modern scientific article. This finding is consistent with the increased specialization of the modern researcher. It is interesting to note that the statistical correlation for all words is slightly stronger than the correlation for content words.

The imagability ratings against which the words in the text were measured can be found in the work of Togliola and Battig [15], and Gilhooly and Logie [8]. To measure imagability, subjects were asked to rate how easily they could construct a mental image of the word. There was a slight decrease in the mean imagability for content words—statistically significant but not strong. This indicates that the words used in later texts are more difficult to visually conceived than those in earlier texts. This may be related to the focus of earlier texts on observation—especially the conveyance of observed phenomena to a possibly skeptical audience in a manner that portrayed the episode as eminently believable. Later articles are more theoretically inclined, attempting to persuade an audience well-versed in theory and scientific jargon that the attributes of a particular theory or methodology are superior or more theoretically probable than its predecessors or contemporaries. Such an audience desires not linguistic simplicity, but in-depth argumentation.

5.2. Causal Cohesion

Causal cohesion measures apply to texts in which events are related causally, that is, one event is said to cause another. WordNet, a lexicon based on the work of Fellbaum [6] and Miller et al. [12] was used as the source for defining verbs as causal.

The number of causal particles in the text (for example: since, so that, because, consequently) has a negative correlation with the publication year. The incidence of causal verbs has shown a slight and statistically insignificant decrease, while the ratio between causal particles and causal verbs has remained steady. This could indicate that the incidence of both is so small that a relative reduction in one makes little overall difference. An interesting observation is that while there has been a decrease in causal particles, the number of causal verbs has not changed.

5.3. Connectives

Connectives include extra words that are placed in texts to clarify and illuminate concepts and to make logical connections both within and between sentences. These are therefore an important indicator of sentence complexity. Connectives are defined as clarifying, additive, or temporal, and further subdivided into positive and negative categories. The positive and negative categories represent whether the subsequent information will continue or contradict, respectively, the previous line of reasoning.

The negative correlation between the incidence of positive temporal connectives and the year indicates a decrease in the number of phrases similar to “from now on,” and “first...then” used, phrases that indicate a positive connection between two parts of the sentence based on space and time. Also in this category is a negative correlation between the incidence of positive causal connectives and publication year. Positive causal connectives include “as a consequence”, “as a result”, phrases that

indicate that the previous part of the sentence caused its successive sentence constituent.

The first question this trend leads one to ask is whether the opposite became the norm. That is, was there a corresponding increase in the use of negative connectives that would indicate a shift from a positive towards a negative emphasis? This question is disproved by the data. Negative temporal connectives showed an insignificant increase, while negative causal connectives showed an insignificant decrease. This indicates that, while the use of these two types of connectives has decreased, there has not been a corresponding shift towards greater use of their negative counterparts. Further support for this lies in the negative correlation for the incidence of all connectives. This result suggests that the overall use of connectives has decreased over the years, which implies the texts have become less cohesive.

5.4. Syntactic Complexity

The syntactic complexity measures determine how difficult it is to analyze the syntactic composition of a text. According to the Coh-matrix theorists, sentence structures that are difficult to analyze are “structurally dense, syntactically ambiguous, have many embedded constituents, or are ungrammatical.” The procedure for determining syntactic complexity is itself complicated, and relies on several syntactic parsers, including a SCOL parser that is integrated with the Brill part-of-speech tagger.

The number of higher-level constituents per sentence (SYNNHS) had a very strong negative correlation with publication year. This indicates a high recidivism rate for the use of multiple higher-level constituents per sentence.

There are two strong correlations in this category, one negative and one positive. The negative correlation between the mean number of higher-level constituents per sentence and the year indicates that later texts contain much simpler syntactic structures than earlier texts. This, coupled with the highly significant increase in noun-phrase modifiers per sentence, indicates that while more elaboration has been given to main nouns, the number of sentence constituents (clauses, etc.) has decreased steadily over time. The positive correlation between the mean number of noun-phrase modifiers and the year (SYNNP) is also strong, which indicates that the use of noun-phrase modifiers has risen steadily between 1800 and 1997.

There was also a negative correlation between the mean number of higher-level constituents per word. This correlation is less strong than the others, but is still significant.

Indicators of logical and analytical complexity increased slightly. This shows a small increase in the number of conditionals, negations, ands, ors, and if-thens. It seems, then, that there has been an important change in syntax over the measured time period in which the shift has been towards simpler sentences with more complex subject matter. Rather than combining multiple ideas into a single syntactic structure by means of multiple phrases, the writers of the latter half of the twentieth century convey complex meaning using expansive and detailed subject matter.

6. CONCLUSIONS

The study discerned significant differences in four general measurement areas: word information, connectives, causal cohesion, and syntactic complexity. Specific findings included:

- Decrease in word concreteness, indicating that the main sentence components have become more abstract.
- Decrease in word imaginability, indicating that the words used in later texts are more difficult to visually conceive than those in earlier texts.
- Increase in age of acquisition, indicating the words used in texts were learned at a later age in the recent past than in the 1800s
- Decrease in the number of causal verbs, indicating the texts have become less cohesive as they became more abstract.
- Decrease in the number of positive connectives, while the negative connectives remained unchanged.
- Negative correlation between the mean number of higher-level constituents per sentence and the year indicates that later texts contain much simpler syntactic structures than earlier texts.
- Steady increase in the use of noun-phrase modifiers.

This experiment represents only the first foray into the use of the Coh-matrix for quantitative analysis of scientific texts. Much further study should be undertaken in order to more accurately gauge both the semantic and cohesive changes in texts over time and the viability of this new resource for the advancement of both researchers' understanding of texts and writer's ability to best shape texts to maximize comprehension.

REFERENCES

- [1] Bazerman, C. (1984). Modern Evolution of the Experimental Report in Physics: Spectroscopic Articles in *Physical Review*. *Social Studies of Science*, 14, 163-96.
- [2] Bazerman, C. (2000). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison: University of Wisconsin Press.
- [3] Charney, D. (1996). Empiricism Is Not a Four-Letter Word. *College Composition and Communication*, 47(4), 567-593.
- [4] Cozzens, S. (1985). Comparing the sciences: Citation context analysis of papers from neuropharmacology and the sociology of science. *Social Studies of Science*, 15, 127-153.
- [5] Devitt, A. (1990) Intertextuality in Tax Accounting: Generic, Referential, and Functional, Bazerman, Charles, and James Paradis, eds. *Textual Dynamics of the Professions: Historical and Contemporary Studies of Writing in Professional Communities*. Madison: Univ. of Wisconsin Press.
- [6] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- [7] Geurin-Pace, F. (1998). Textual Statistics: An Exploratory Tool for the Social Sciences. *Population: An English Selection*, 10(1), 73-95.
- [8] Gilhooly, K.J. and Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation*, 12, 395-427.
- [9] Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Matrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36(2), 193-202

- [10] Myers, G. (1990). Stories and Styles in Two Molecular Biology Review Articles. Bazerman, Charles, and James Paradis, eds. *Textual Dynamics of the Professions: Historical and Contemporary Studies of Writing in Professional Communities*. Madison: Univ. of Wisconsin Press.
- [11] Kracauer, S. (1952-53). The Challenge of Qualitative Content Analysis. *Public Opinion Quarterly*, 16(4), 631-642.
- [12] Miller, G. et al. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography* 3(4). 235-244.
- [13] Mitchell, R. (1967). The Use of Content Analysis for Explanatory Studies. *Public Opinion Quarterly*, 31(2), 230-241.
- [14] Pavio, A., Yuille, J.C. & Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement*, 76 (3, part 2).
- [15] Toglia, M.P., & Battig, W.R. (1978). *Handbook of Semantic Word Norms*. New York: Erlbaum.
- [16] Wei, Yu Hong, & Davies, G. (1997). Do Grammar Checkers Work? A Report on the Research into the Effectiveness of Grammatik V Based on Samples of Authentic Essays by EFL Students. *New Horizons in CALL: Proceedings of EUROCALL 96*.