Language Use Reflects Scientific Methodology: A Corpus-Based Study of Peer-Reviewed Journal Articles

Shlomo Argamon[1]* **Jeff Dodick**[2] **Paul Chase**[1]
[1] Department of Computer Science, Illinois Institute of Technology
10 W. 31st Street, Chicago, IL 60616, USA
[2] Department of Science Teaching, The Hebrew University of Jerusalem
Givat Ram Campus, Jerusalem, ISRAEL

December 21, 2006

Abstract

Recently, philosophers of science have argued that the epistemological requirements of different scientific fields lead necessarily to differences in scientific method. In this paper, we examine possible variation in how language is used in peer-reviewed journal articles from various fields to see if features of such variation may help to elucidate and support claims of methodological variation among the sciences. We hypothesize that significant methodological differences will be reflected in related differences in scientists' language style.

This paper reports a corpus-based study of peer-reviewed articles from twelve separate journals in six fields of experimental and historical sciences. Machine learning methods were applied to compare the discourse styles of articles in different fields, based on easily-extracted linguistic features of the text. Features included function word frequencies, as used often in computational stylistics, as well as lexical features based on systemic functional linguistics, which affords rich resources for comparative textual analysis. We found that indeed the style of writing in the historical sciences is readily distinguishable from that of the experimental sciences. Furthermore, the most significant linguistic features of these distinctive styles are directly related to the methodological differences posited by philosophers of science between historical and experimental sciences, lending empirical weight to their contentions.

Topics: Communication, Discourse Analysis, Philosophy of Science, Corpus Linguistics

Keywords: Philosophy of science, Scientific method and reasoning, Communication, Discourse analysis, Systemic functional linguistics, Science education, Machine learning applications

---

[*] Corresponding Author

1 Introduction

Social interaction and communication among scientists is a key component of scientific practice (Dunbar, 2001; Latour & Woolgar, 1986). Bibliometric citation analysis of the scientific literaturehas been often used to explain large-scale factors of information flow between researchers, research topics, and disciplines (Myers, 1990; White & McCain, 1989, 1998; Cronin & Overfelt, 1994; MacRoberts & MacRoberts, 1996). Studying aspects of how scientists communicate directly with one another has also helped elucidate specific features of scientific problem solving, as “in-laboratory” studies of working scientists have shown (Dunbar, 2001; Dunbar & Blanchette, 2001). Through such analyses, scientists may be seen to use language to create a collaborative intellectual space (Cronin, 2005), whose shared linguistic worldview makes possible communication about complex observations and hypotheses (Goodwin, 1994), via various ‘explanatory activities’ (Okada & Simon, 1997).

Indeed, analyses of academic writing (Swales, 1990; Myers, 1990; Bazerman, 2004; Bazerman & Prior, 2005), have shown how consideration of large-scale patterns in textual organization (across many and varied scientific texts) can yield useful generalizations about the structure and function of scientific argumentation. A crucial finding has been the elucidation of the *socially situated* nature of scientific genres, in that their evolving structures reflect complex negotiations of claims and alliances among scientists (Hyland, 2000). Scientific fields may thus be profitably viewed as a sort of ‘discourse community’ each with its own way of using language to represent, evaluate, and decide competing knowledge claims (see, e.g., (Swales, 1990; Killingsworth & Gilbertson, 1992; Harris, 1989)).

Concurrent with these recent developments in our understanding of scientific communication and analysis of its genres, philosophers of science have been increasingly recognizing that the classical model of a single “Scientific Method” (based on experiment-based fields such as physics) improperly devalues sciences such as geology and paleontology, due to their use of non-experimental *historical* methods, despite the fact that their epistemological claims are also quite strong. Rather, some philosophers now argue that differences in method may stem directly from the types of phenomena under study (Cleland, 2002). In brief overview: *Experimental sciences* (such as physics) attempt to formulate general predictive laws, and so rely heavily on repeatable series of controlled experiments which test hypotheses. *Historical sciences*, on the other hand, deal with contingent phenomena, studying specific phenomena in the past in an attempt to find unifying explanations for effects caused by those phenomena (Mayr, 1976). Because of this, reasoning in historical sciences consists largely of reconstructive reasoning (*retrodiction*), as compared to the predictive reasoning from causes to possible effects characteristic of experimental science (Diamond, 2002; Gould, 1986; Whewell, 1837).

In this paper, we adduce the first quantitative empirical evidence for such methodological differences among different fields, by analyzing how scientists in different fields use language differently to construct scientific arguments in peer-reviewed journal articles. Our main method is to define a topic-independent set of language features which are used as input to a text classification learner which enables us to evaluate to what extent articles from one journal (in a given field) are similar to those in another journal (from the same or another field). We applied such discriminative analysis to articles in a number of journals in different experimental and historical sciences, seeking consistent constellations of linguistic/stylistic features that distinguish between these two types of science.

More specifically, we analyze linguistic variation between peer-reviewed articles in journals representing three historical sciences (geology, paleontology, and evolutionary biology) and three experimental sciences (physics, organic chemistry, and physical chemistry)[1]. Our hypothesis is that linguistic/rhetorical differences between articles in different fields will be found which are connected to differences between the experimental and historical methodologies. Experimental results using standard methods of computational stylistics support this hypothesis. Furthermore, we find that classification using a set of linguistically-motivated features, based on systemic functional principles, enables a more nuanced examination of rhetorical differences between texts in different kinds of science, allowing us to connect these differences in language use with methodological differences previously posited by philosophers of science. To be clear, these results imply no cognitive claims about how scientists in different fields may reason differently, rather they indicate certain differences in how historical and experimental scientists *collectively* represent and negotiate their claims.

2 Language in Science

We contend that gaining any kind of clear understanding of possible methodological differences among different scientific fields will require empirical investigation of how scientists actually *do science* in practice. A central component of any such investigation is to study how scientists *communicate* (Mulkay & Gilbert, 1983). All aspects of how scientists reason, make discoveries, and persuade their colleagues of their

---

[1] Fields were designated *a priori* as ‘experimental’ or ‘historical’ as classified by modern philosophers and historians of science (Cleland, 2002; Diamond, 2002).

conclusions are reflected to some extent in how they communicate with one another. Studying scientific communication in its natural context also avoids the bias inherent in introspective studies such as interviews or self-reports (Dunbar, 2001).

The study of communication as a way to investigate scientific cognition has begun during the past decade, focusing on the experimental sciences. Dunbar and colleagues have studied research activities in molecular biology and immunology laboratories from a cognitive perspective (Dunbar, 1995, 1999, 2001; Dunbar & Blanchette, 2001), which show subtle use of research heuristics and different kinds of analogy in laboratory discussions. An anthropological-sociological perspective was taken by Latour and Woolgar (1986) in their important study “Laboratory Life”, which investigated scientists’ production of “facts...on an assembly line” by analyzing the written products of research activities as “inscriptions” from an anthropological perspective. Related approaches, focusing on communication, rhetoric, and cognition, have been used to study how scientists teach, construct theories, negotiate claims, and interpret observations in physics (Ochs, Jacoby, & Gonzales, 1994; Ochs & Jacoby, 1997), archaeology (Goodwin, 1994), oceanography (Goodwin, 1995), medicine (Fujimura, 1987), organic chemistry (Bond-Robinson & Stucky, 2005), biology (Myers, 1990), and biomedical engineering (Nersessian, 2005).

The structure of scientific language has evolved to suit particular demands of scientific communication and practice, as shown, for example, in several studies by Halliday and Martin (1993) . One study examines a key feature of modern scientific writing, the use of complicated nominal groups to express events and processes, which often makes scientific writing difficult or impenetrable for the non-initiate. Halliday traces the evolution of this style in English from Newton’s *Treatise on Opticks* (1704) which is one of the earliest texts to use complex nominalizations to compress complex processes (e.g., “a diverging...of the heterogeneous Rays from one another”). Such nominalizations enable complex configurations of processes to be neatly packaged and subsequently referred to by name or pronominal reference; furthermore, reifying processes in this way (effectively turning them into ‘things’) enables causal, logical, and rhetorical relations to be made more explicit in the text through the use of verbs expressing causality and epistemic events (e.g., *causes, results [in], argues, proves, contradicts*, and so forth). This development in scientific language (as well as others documented by Halliday and Martin) is thus a direct outgrowth from the need of scientists to communicate precisely about complex causal and logical structures. We thus conjecture that scientific fields that study phenomena with largely differing causal structures or evidential requirements will consequently differ in their preferred use of linguistic resources.

Table 1: Types of communication between scientists and their audiences. This paper focuses on studying peer-reviewed, published articles.

		Communication with:	
Communication Mode	**Collaborators**	**Colleagues**	**Students**
Speech	Lab Meetings Field Research	Conference Talks	Lectures
Writing *Informal*	Research notes	Correspondence	Lecture notes
Formal	Lab reports	**Published articles**	Textbooks

The growing body of ethnographic, cognitive, and linguistic analyses of scientific communication, however, includes virtually no comparative work on how scientific communication may vary among different disciplines. Indeed, with the few exceptions of Goodwin's ethnographic study of oceanographers (Goodwin, 1995) and archaeologists (Goodwin, 1994), Abrams and Wandersee's (1995) study of life scientists (which had a small subset of historically based scientists), and Kelly and Bazerman's (2003) study of student writing in oceanography, we are unaware of any work specifically on how *historical* scientists conduct their research or communicate their findings. Moreover, there has been no research to date, as far as we know, on possible differences in communication modes in historical and experimental sciences. Our goal, therefore, is to show how the empirical study of such differences, via corpus-based text analysis, can help illuminate aspects of the characteristic communication and reasoning modes of both the historical sciences and the experimental sciences, with reference to their different epistemological needs.

2.1 Types of communication

There are several different types of communication between scientists (in any field), which can be roughly classified by the mode of communication (speech or formal/informal writing) and by the target of the communication (research collaborators, other colleagues, and students), as laid out in Table 1. The study of speaking patterns in scientists will require videotaping and transcription, as a prelude to linguistic discourse analysis. This is expensive and time-consuming, though, and generally precludes working with very large data sets. Written material, when available in electronic form, is often easier to work with, however, since automated tools can more easily be brought to bear. This is particularly true for formal written texts, such as journal articles or textbooks.

The question is then how analyzing the style of formal, edited texts can enable a better understanding of scientific methodology. The main idea is that many aspects of language style directly reflect how the author organizes ideas, argues for or against a position, or explains new concepts. Specifically, in scientific communication, we contend that different scientific methodologies should imply concomitant differences in how scientists organize their arguments, and that these differences in turn will be reflected in their use of language. Thus, by a detailed investigation of scientific speaking and writing we may obtain empirical evidence for and against different proposals for describing scientific methodology in different fields.

3 Analyzing Textual Style

This section describes our linguistic framework for analyzing variation in language style based on corpus analysis. Our principal theoretical tool is a collection of linguistically-motivated taxonomies of topic-independent keywords and phrases, which enables us to use machine learning methods to find meaningful features of stylistic variation between texts. The taxonomies we use are derived from Systemic Functional Linguistics (SFL), relevant notions of which are described immediately below. We then illustrate these concepts via hand-analysis of two representative scientific texts, the introductory sections of two journal articles in our corpus.

3.1 Systemic Functional Linguistics (SFL)

The linguistic framework we assume is that of systemic functional linguistics (Halliday, 1994). Systemic functional linguistics (SFL) construes language as a set of interlocking choices for expressing meanings: "either this, that, or the other", with more general choices constraining the possible specific choices. As a simple intuitive example (cf. Fig. 1): "A clause is either a command (imperative) or about some fact (indicative); if indicative, it may be declarative or interrogative; if interrogative, it may be a Wh-question (*when*, *where*, *how*, etc.) or a yes/no question," and so on. A **system**, then, is a set of **options** for meanings to be expressed, with **entry conditions** specifying when that choice is possible — for example, if a message is not about doing, then there is no possible choice between expressing standalone action or action on something. Each option has also a realization specification, which gives constraints (lexical, featural, or structural) on statements that express the given option. Options often serve as entry conditions for more detailed systems, which we will call **subsystems** in this paper.

By structuring language as a complex of choices between mutually exclusive options, the systemic approach is particularly appropriate for examining variation in language use. A systemic specification allows

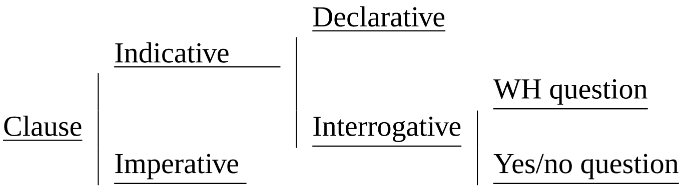

Figure 1: Example tree diagram showing a system and options for several message types.

us to ask the question: In places where a meaning of general type A is to be expressed in a text (e.g, "a message about action"), what sorts of more specific meanings (e.g., "standalone action" or "action on a thing") are most likely to be expressed by different types of people or in different contexts? While much of the meaning potential of language is determined by the sort of ideas being expressed, the specific form of an utterance is underdetermined by its purely representational meaning. Other layers of meaning in terms of interpersonal relations, attitude towards propositions, and intratextual logical or rhetorical connections (cohesion) are also present, as well as subtle choices of focus.

As an example of a cohesive system in English, when expanding the meaning of one clause by another clause (also see Sec. 5.3.2 and Fig. 4), one may choose between three possibilities: elaboration (deepening by restatement, comment, or exemplification: "He left, which was good", commenting on the event), extension (adding new information: "He left, and I felt better", adding on a related event), and enhancement (qualification by reference to circumstance, cause, manner, or result: "He left, so I rejoiced", creating a short causal chain). Note that all three examples have similar representational meanings, though more subtle distinctions are drawn. A general preference for one or another option is thus largely a question of style or of attitude, in which individual and social/contextual factors come to bear. Such preferences can be measured by evaluating the relative probabilities of different options by tagging their realizations in a corpus of texts. By comparing how probabilities vary between individuals or situations with different characteristics, we may determine how those characteristics affect linguistic behavior.

Previous work has investigated the relationship between choice probabilities and contextual factors. For example, Plum and Cowling (1987) demonstrate a relation between speaker social class and choice of verb tense (past/present) in face-to-face interviews. Similarly, Hasan (1988) has shown, in mother-child interactions, that the sex of the child and the family's social class together have a strong influence on several kinds of semantic choice in speech. These previous studies involved hand-coding a corpus for systemic-functional and contextual variables and then comparing how systemic choice probabilities vary with contextual factors via multivariate analysis. Our work primarily applies automated computational methods to analysis of choice probabilities (in addition to the manual analyses in Section 3.2 below), giving us a potentially broader (though somewhat shallower) view, by enabling effective analysis of larger numbers of documents than can be analyzed by hand. This enables more general conclusions than analysis of only a few hand-analyzed examples.

3.2 Cohesion and modal assessment

The stylistic features considered in the current study are drawn from Matthiessen's (1992) SFL grammar of English. We focus on two large-scale aspects of a text related to its rhetorical organization that enable it to realize complex conceptual structures[2]:

Cohesion refers to linguistic resources that enable language to connect to its larger context, both textual and extratextual (Halliday & Hasan, 1976). Such resources include a wide variety of referential modalities (pronominal reference, deictic expressions, ellipsis, and more), as well as lexical repetition and variation, and different ways of linking clauses together (Mann & Thompson, 1988). How an author uses these various cohesive resources is an indication of how the author organizes concepts and relates them to each other. Within cohesion, our computational work considers just conjunctions, for feasibility of automated extraction; in the illustrative hand analyses below we also consider the thematic organization of the text.

Clausal conjunctions are organized via the system of EXPANSION, which describes features that serve to link clauses together causally or logically. The first of the three main types of EXPANSION is EXTENSION, which links clauses giving different information together, realized by words such as "and", "but", and "furthermore". The second type is ENHANCEMENT, which qualifies information in one clause by another (e.g., "similarly..." or "therefore..."). The third is ELABORATION, which deepens a clause by clarification or exemplification (e.g., "in other words..." or "more precisely").

Another important system that contributes to cohesion is THEME, in which each clause in a text indicates what it is 'about'; note the difference between "John ate an apple" (about John) and "An apple was eaten by John" (about an apple). The organization and flow of different themes in a text contributes to its overall cohesion and rhetorical structure. (Note that our computational study does

---

[2] More detail on these systems and features as used in our work are given in Section 5.3.

not use THEME, since it cannot be reliably extracted automatically. We do consider it in the hand-analyzed examples below, however, to enhance the clarity of the exposition.)

Assessment we define as qualification of events or assertions in the text according to their rhetorical or epistemic properties. One key system for assessment is MODALITY in which the likelihood, typicality, or necessity of an event are indicated, usually by a modal auxiliary verb or an adjunct adverbial group. There are two main types of modality: MODALIZATION, which quantifies levels of likelihood or frequency (e.g., "probably", "might", "usually", "seldom"), and MODULATION, which quantifies ability or necessity of performance (e.g., "ought to...", "should...", "allows...").

A second important system for assessment is COMMENT, which is one of modal assessment, comprising a variety of types of "comment" on a clause's function as a 'message', assessing the writer's attitude towards it, or its validity or evidentiality. Comments are generally realized as adjuncts in a clause (and may appear initially, medially, or finally).

3.3 Example passages

To better clarify the roles of the above systems in structuring scientific rhetoric, we consider now hand analyses of the introductory paragraphs of two articles (Figs. 2 and 3), one from *Physics Letters A* and the other from the *Journal of Metamorphic Geology*. Note that these passages are given for illustrative purposes only — we do not claim that any broad conclusions about scientific rhetoric and reasoning can be drawn from just two specific examples. However, these analyses do illustrate some of the features that the automated corpus analysis below examines on a larger scale.

For the analysis here, the texts were divided into their consituent clauses (embedded clauses are not segmented), and the above-described textual elements highlighted: themes[3] are underlined, conjunctions (realizing EXPANSION) are in **boldface**, realizations of MODALITY are in SMALL CAPS, and COMMENTs are in *italics*.

---

[3] For simplicitly, we only consider topical themes, those that relate to the event or fact being described (Halliday, 1994, Ch. 6), in our illustrations here. The corpus-based study described below does not consider theme at all, for computational reasons.

1. Quantum entanglement is the novel phenomenon in quantum mechanics.
2. **Because of** Enhancement its unique properties presented in the Einstein-Podolsky-Rosen paradox,
3. quantum entanglement plays a more and more important role in quantum information science and has been widely used in quantum teleportation, superdense-coding, quantum error correction, etc.
4. In recent years, entangled states have been prepared by various physical schemes, which provided more entanglement resources for quantum information processes.

¶

5. In the branch of quantum teleportation, ones pay more attention on the physical realization of quantum teleportation.
6. **Although,** Elaboration in the originally theoretical scheme proposed by Bennett et al. the fidelity of the teleported unknown state CAN Modulation reach unity,
7. in practice, it is quite difficult to prepare the maximally entangled states (EPR states).
8. Studies on the quantum channel supported by non-maximally entangled states are of practical significance.
9. **Hitherto,** Elaboration H. Lu et al. have presented their theoretical schemes of probabilistic teleportation of two- and three-particle entangled states where two pairs of non-maximally entangled particles were used as quantum channel.
10. Shi et al. have shown two different probabilistic teleportation schemes of a two particle entangled state by pure three-particle states which are not maximally entangled.

¶

11. **But** Extension the teleportation COULD NOT Modulation be widely used in practice if just based on the sheer theoretical proposals.
12. The promising potential of teleportation SHOULD BE Modulation realized with the help of the experimental schemes.
13. **So,** Elaboration in 1997, the teleportation has been firstly realized in laboratory by Bouwmeester et al.
14. Some other schemes have been proposed by using cavity QED, ion traps, nuclear magnetic resonance and so on.
15. In Ref. [14], Zheng et al. proposed a strategy for the teleportation of atomic states within cavities.
16. In this Letter, we will present a more efficient scheme with fewer operations than that in Ref. [14], which will be discussed in Section 2.

Figure 2: Example Passage A: Introductory paragraphs of Z.-L. Cao, M. Yang, G.-C. Guo, “The scheme for realizing probabilistic teleportation of atomic states and purifying the quantum channel on cavity QED”, *Physics Letters A* **308** (2003) 349–354, marked up as indicated in the text with paragraph breaks indicated by ¶.

1. Crustal granulite facies rocks differ from those of lower metamorphic grades in that they contain predominantly anhydrous minerals, lack a free H_2O-rich fluid phase during metamorphism and involve partial melting.
2. These main features of granulite facies rocks are a consequence of dehydration melting in which H_2O is strongly partitioned into silicate melt.
3. This process is in contrast to progressive dehydration under subsolidus conditions in which the H_2O evolved in dehydration reactions is inferred to escape.
4. **Instead,** EXTENSION the partitioning of H_2O into silicate melt and the absence of a H_2O-rich fluid at higher grade conditions, means that the amount of H_2O in a rock will remain constant
5. **unless** ELABORATION melt is lost.
6. **Although** EXTENSION the idea that melt is lost from partially melted rocks is not a new one,
7. the amount of melt loss that occurred from many terranes is poorly constrained.
8. Estimates of the degree of melt loss from several terranes have been undertaken on the basis of geochemical and / or petrographic constraints.
9. **If** ELABORATION no melt is lost from granulite facies rocks,
10. high-grade anhydrous assemblages will be retrogressed on cooling via the reversal of the partial melting reactions to hydrous assemblages TYPICAL MODALIZATION of the upper amphibolite facies.
11. **Although** EXTENSION some degree of hydrous retrogression IS COMMON MODALIZATION in granulite facies terranes,
12. the effects of such retrogression are *generally* VALIDATION limited
13. **and** EXTENSION the high-grade anhydrous assemblages are preserved.
14. **Thus,** ELABORATION understanding the preservation of granulite facies mineral assemblages with little or no retrogression is important in the context of mid- to lower-crustal processes.
15. The origin and behaviour of melt-bearing metamorphic rocks (migmatites) has been studied widely through field observation, geochemical studies and experimental studies.
16. **However,** EXTENSION field and geochemical studies are hampered by the fact that only the inferred products of partial melting CAN BE MODULATION observed and not the melt itself.
17. **Furthermore,** EXTENSION most experimental studies give little information on the processes of melt migration and segregation.

    ¶

18. In this paper we model the effects of melt loss from granulite facies metapelitic rocks, investigating the effects of varying proportions of melt loss at different temperatures.
19. The effects of melt loss on the development and preservation of granulite facies mineral assemblages are illustrated using phase diagrams calculated for two model metapelitic compositions.

Figure 3: Example Passage B: Introductory paragraphs from R. W. White and R. Powell, “Melt loss and the preservation of granulite facies mineral assemblages”, *Journal of Metamorphic Geology*, 2002, **20**:621–632, marked up as in Figure 2

Analysis of Passage A

Cohesion: To analyze how the above passage is structured as a cohesive whole, we first examine the development of ideas as realized in the themes of successive clauses, related by particular kinds of conjunction. These include *phenomena*: “quantum entanglement” (clauses 1, 3), “its unique properties” (2), and “quantum teleportation” (5, 11, 12); *scientific positions*: theoretical schemes denoted by their proponents’ names (6, 9, 10, 14, 15), experimental studies (8), general practical experience (7), and “this Letter” (16); and *periods of time*: “recent years” (4) and “1997” (13). The passage opens, in its topic sentence, with “quantum entanglement”, explicitly mentioning its importance (3), enhanced by connection to its “unique properties” (2), and the fact that “recent years” have seen more attention to it. The second paragraph focuses in on one of the applications mentioned in (3), “quantum teleportation”, and various schemes to implement it. Clauses 6 and 7 form an opposition between a “theoretical scheme” and “practice”, followed by more theoretical schemes introduced by clause 8’s “studies” and elaborated on by mention of two eponymous “schemes” in clauses 9 and 10. This gives a ‘bracketed’ structure: Theory-Practice-Theory($\times 3$). This opposition between theory and practice is then seen at a higher level when moving now to the third paragraph, which begins with the adversative expansion of **but**, contrasting what is to come with the previous paragraph. The initial theme of the last paragraph, as of the previous one, is “teleportation”, but the next move is to “potential”, then temporally located recently in 1997 (13), recalling the “recent years” of the opening paragraph. The structure continues to mirror that of the second paragraph, where clauses 14 and 15 catalog other practical schemes for realizing teleportation, which then are contrasted with the scheme “[i]n this Letter” (16), giving a structure of Promising Potential–Realization($\times 3$)–This Letter. If seen as a rough parallel of the bracketing in the preceding paragraph, the scheme to be presented in “this letter” may be seen as a realization of the “promising potential” of clause 12.

Note in particular how conjunctions make explicit certain rhetorical relations between clauses or paragraphs in the passage. Most of the conjunctions mark one clause as subservient to another (elaborating on it, or enhancing our understanding of the context): **Because** (2 to 3), **Although** (6 to 7), **Hitherto** (9 to 10), and **So** (13 to 12). The sole exception is **But** in clause 11, which marks the entire third paragraph as contrasting with the second. Overall, the conjunctions make explicit the tightly-coupled storyline of each paragraph and the passage as a whole, moving from “stating the importance of quantum entanglement” to “establishing the practical difficulties in purely theoretical schemes for quantum teleportation” to “more recent practical schemes, including the more efficient one given in this Letter”.

Assessment: There are only a few realizations of modal assessment in the passage, which presents its ‘story’ as certain and uncontroversial. The three modal finite auxiliary verbs that do appear are all realizations of Modulation, i.e., they assess ability or necessity of actions or events, as opposed to the frequency or usuality of occurrence of events in reality (termed Modalization). In clause 6 ‘can’ allows fidelity to theoretically reach unity, which ability is then contrasted with the difficulty of preparing such a state in practice (7). Clause 11 asserts the utter inability of achieving teleportation in practice if only theoretical proposals are considered, and clause 12 consequently asserts the necessity for researchers to work on realizing experimental schemes.

Analysis of Passage B

Cohesion: By contrast to the previous passage, whose development created a storyline of different theories and studies, the current passage first (clauses 1–13) builds a picture whose thematic elements are largely physical entities and phenomena. Each of the three thematic entities “granulite facies”, “melt”, and “anhydrous assemblages” is associated with a process, “dehydration melting”, “melt loss”, and “hydrous retrogression”, respectively. Clauses 1–3 move from the main entity of “granulite facies” to the process of dehydration melting; clauses 4–7 introduce the “melt” as a product of melting and move to “melt loss”, noting in clause 8 that melt loss has been estimated, implying that this estimation is important. This point is then justified by moving from the condition “no melt” to the process of retrogression of anhydrous assemblages, focusing attention on the case where such retrogression is limited and assemblages are preserved (clauses 12, 13). The remainder of the first paragraph establishes a gap in the current research (see Lewin, Fine, and Young (1986) for a discussion of gap-establishing strategies), establishing first the problem of “understanding the preservation...” then asserting that only the “origin and behavior” of such rocks have been examined by “field and geochemical studies” and “most experimental studies”. The gap thus created is then filled (via foreshadowing) in the second paragraph by “this paper” (clause 18) which will show “effects of melt loss on...granulite facies mineral assemblages”.

Thematic analysis thus shows an overall organization of the passage around three main foci (facies, melt, and assemblages), along with the final shift from describing phenomena to discussing studies (including the current paper). Examination of conjunctions in the passage bears out this impression as well—six out of nine are EXTENSION, which links two independent information units together (rather than merely elaborating or enhancing a single information unit). For example, in clause 4 "partitioning of H_2O" is contrasted to the "progressive dehydration" mentioned in clause 3. Similarly, the **and** in clause 13 creates an aggregate idea comprising the facts that "the effects of ... retrogression are ... limited" and that "the ... assemblages are preserved." Note the linking of two independent ideas, as opposed to the elaboration in clause 14, where the notion that "understanding the preservation...is important" is presented as a clear conclusion of the paragraph up to that point, and not as an independent idea being introduced. So we see that in terms of both THEME and EXPANSION, the passage exhibits a multifocal structure, placing the main topic of the paper (melt loss) at the center of a complex of entities and phenomena: granulate facies, dehydration processes, anhydrous assemblages, and hydrous retrogression, whose causal relationships are quite complicated.

Assessment: Modal assessments in clauses 10, 11, and 12 show how explicit expression of modal strength implies the possibility of exceptions. First, we see in clauses 10 and 11 a pair of similar assessments of MODALIZATION/USUALITY: clause 10 expresses (*inter alia*) that hydrous assemblages are a *usual* feature of the upper amphibolite facies and clause 11 that hydrous retrogression is *frequent* in granulite facies. Clause 12 contrasts this frequency with the limited effect of such retrogression, but then qualifies this limitation as applying "generally". In all three of these cases, omission of the assessing words would have strengthened the statements; their inclusion thus nuances the text's claims (and makes it more resistant to possible counter-examples). One example of MODULATION is present in the text, in clause 16, referring to the inherent limitations of observational methodology.

Summary of passage analyses

In sum, we first note the strong interaction between thematic structure and the types of conjunctions used, in creating cohesion. In (experimental) Passage A, we saw that the strong use of elaborative and enhancing conjunctions, together with tightly-coupled themes in neighboring clauses, produces a highly focused effect. In (historical) Passage B, on the other hand, the use of extensive conjunction works with a more dynamic thematic development to create a multi-focal effect, keeping distinct entities and processes in simultaneous view.

Second, we also see how differences in kinds of modal assessment produce a different picture of how and why assertions in the text are so qualified. Passage A has very little modal qualification, and what is there relates to the mere possibility of certain events; this gives it a highly authoritative and definitive tone. Passage B, by contrast, qualifies statements by their usuality, making it clear that while general conclusions can be drawn, certainty cannot be claimed regarding any particular instance of the phenomenon under study.

4 Scientific Methods and Linguistic Variation

The dominant view in the philosophy of science, during the 19th and early 20th centuries, was that a single "experimental" scientific method was the key to all true science (Rudolph & Stewart, 1998). Based on the success of Newtonian physics, philosophers took physics as the model of how proper science should be done. Indeed the views of 19th century philosophers on science reflected a strong bias towards direct observations made during controlled experimental manipulation of nature (Mayr, 1985; Kitcher, 1993; Rudolph & Stewart, 1998). It was believed at the time that by using experimental methods, that is, by manipulating independent variables and measuring changes in dependent variables, scientific conclusions could be established with certainty.

Indeed, Gould (1986) has noted that many in the British scientific community of the time who accepted Darwin's conclusions concerning Evolution by Natural Selection still had deep misgivings about the historical-comparative method he used to arrive at his conclusions. As Rudolph and Stewart (1998) note, such misgivings partially derive from the broader philosophical backdrop of the time. Earlier in the century, philosophers including John Herschel in his *Preliminary Discourses on the Study of Natural Philosophy* (1830) and John Stuart Mill's *System of Logic* (1843) had established supposedly firm foundations for 'scientific methodology' (Hull, 1973). In their view, true science was firmly associated with the empirical, inductive methods of Newton and other physicists.

This restricted view of the scientific method, however, ignored the epistemic basis of evolutionary biology up to the time of Darwin. Naturalists of the time had described and classified a great many phenomena of the past and present which could explain general patterns of development. In contrast to the complex diversity of life that naturalists tried to explain, the methods of Newton attempted to reduce the world to its simplest forces. The success that Newton had with such reductionist methods provided the model for science that many in the philosophical community of the 19th century used as their standard for how to produce (supposedly) certain knowledge. Perhaps due to this bias, it was not until Darwin's development of a historical scientific methodology that a physical theory of biological evolution could be provided that was generally accepted by the scientific public.

The bias of philosophers towards Newtonian physics as the model of science continued well into the 20th century, and was reinforced by the development of the school of thought known as ‘logical empiricism’. In its view, the goal of inquiry in science is the establishment of laws or universal mathematical statements (Sober, 1993). Such sciences then produce theories that lend themselves well to experimental confirmation (Rudolph & Stewart, 1998). While such a methodology can be, and is, generally applied in the physical sciences, attempts at reconfiguring historical sciences such as evolutionary biology and geology towards this methodology have largely failed. As an illustration, consider Charles Lyell’s attempt to base geology on a Newtonian model by requiring geologists to assume a priori that actual causes observed in the present were wholly adequate to explain the geological past, not only in kind but also in degree (Rudwick, 1998). This framework greatly influenced many geologists’ practice in the modern age. However, recent work on large scale problems such as glacial flooding and dinosaur extinction has shown that this framework is inappropriate for investigating many such geological problems (Dodick & Orion, 2003).

More generally, as noted by Frodeman (1995), some philosophers of science have specifically labeled geology as a “derivative science”, viewing its reasoning and methodology as reduceable to elements supplied by physics and chemistry. Moreover, in contrast to experimental sciences in which all variables could be controlled, geology has been said to have a number of fundamental problems (including the great expanse of geological time, and gaps in the stratigraphic record) that undercut its claim to knowledge.

Recently, however, a growing group of philosophers, scientists, and educators have become skeptical about the existence or even possibility of a single method for all of science (Cleland, 2001, 2002; Cooper, 2002, 2004; Diamond, 2002; Gould, 1986; Mayr, 1985; Rudolph & Stewart, 1998). This work recognizes that experimental and observational methods and various kinds of evidential reasoning can play different roles in different sciences. Nonetheless as Cleland (2001, 2002) notes, most philosophers have been reluctant to make large scale comparisons concerning methodology across different disciplines. The ultimate result is that many outside the philosophy of science are often left with the impression that historical science is lacking when compared to experimental science.

While approaches differ among different researchers, a broad outline of the differences that have been proposed between the experimental and historical sciences may be given. Based on Diamond (2002), we consider here four (interrelated) fundamental dimensions categorizing this difference, summarized in Table 2.

Table 2: Summary of dimensions of difference between experimental and historical scientific methodologies (after Diamond (2002)).

	Experimental	**Historical**
Research goal	General laws of behavior	Explanations for ultimate and contingent causes
Evidence gathered by	Controlled manipulation of nature	Observing pre-existing entities and phenomena
Hypotheses are tested for	Predictive accuracy	Explanatory adequacy
Objects of study	Uniform and interchangeable entities	Complex and unique entities

Experimental science, as is well known, gathers knowledge by controlled experimentation, in which natural phenomena are manipulated in order to test a theory. The quality of such a theory is measured by the consistency of its predictions with such experiments, and ideally, such a theory expresses a general causal law or universal statement that is applicable to a wide variety of phenomena in many contexts. Finally, the form of such research is dictated in large part by the study of uniform entities, either singly or in assemblage; the fact that such entities (atoms, molecules, genes) are identical, or nearly so, makes the formulation of general laws possible in principle, and experimental reproducibility a reasonable requirement in practice.

Historical science, on the other hand, investigates ultimate causes which often lie very deep in the past, and whose effects are observed only after very long and complex causal chains of intervening events. Consequently, evidence is gathered by observation of naturally occurring traces of phenomena, since manipulation is impossible (e.g., we cannot wait millions of years for the results of a hypothetical geological experiment!). Comparison of similar observed phenomena has thus developed as a key analytical technique for the historical sciences, typified by a kind of ‘natural experiment’ (Diamond, 2002) in which systems differing in a key variable are compared in order to estimate the effect of that variable on other system characteristics.

This focus on investigating past causation further implies that the ultimate test of quality in historical science is explanatory adequacy, because the phenomena under investigation are unique and contingent, and so cannot be expected to repeat in the historical record. The methodology of such explanatory reasoning derives from what Cleland (2002) calls the ‘assymmetry of causation’, in that effects of a unique event in the past tend to diffuse over time, with many effects being lost and others muddled by other intervening factors. Making sense of this complexity requires, therefore, ‘synthetic thinking’ a lá Baker (1996), in which one fits together complex combinations of many pieces of evidence to form arguments for and against competing hypotheses. In addition to sorting through this great causal complexity, historical scientists must also deal with the complexity of the individuals under study. Unlike (atomic) particles (for example), which are all uniform, entities that are studied by historical scientists — people, species, strata — are all unique (though often similar) individuals, whose precise details of configuration and function are not always recoverable, in practice or even in theory. This removes the possibility of formulating universal laws of behavior, allowing only statistical statements of relative likelihoods at best — it is very difficult to rule a specific possibility out entirely, but rather arguments for and against hypotheses must be made on the preponderance of the best evidence. Thus reasoning about the relative likelihood of different assertions is endemic to historical science's synthetic thinking.

4.1 Hypotheses

What sorts of linguistic variation between experimental and historical science writing would be expected, based on the methodological differences posited above (Section 4)? We clearly must only consider topic-*independent* (or 'stylistic') features, since each field will be trivially distinguished by its own jargon (such as 'fossil' in paleontology, or 'quantum entanglement' in physics). Thus our main hypothesis is that the two kinds of science will be stylistically distinctive:

> $\mathbf{H_1}$: Stylistic features of the texts will distinguish more strongly between articles from different kinds (historical or experimental) of science than between articles from different journals[4] in the same kind of science.

We also formulate more detailed hypotheses regarding the kinds of rhetorical features we expect to be most significant in distinguishing articles in the different fields, based on the above-discussed methodological differences between historical and experimental sciences, as follows.

Recall from Section 4 that the goal of historical reasoning is to find ultimate causes, rather than the general laws sought by experimental science. This leads to the related issue of how research quality is measured, whether by adequacy of reconstructive explanation (historical science) or of prediction (experimental science). Thus, a key element of historical reasoning is the need to differentially weigh the evidence. Since any given trace of a past event is typically ambiguous as to its possible causes, many pieces of evidence must be combined in complex ways in order to form a confirming or disconfirming argument for a hypothesis. Such thinking is, as Cleland (2002) argues, a necessary commitment of historical science (as

---

[4] We compare *different* journals to partially control for possible influence of editorial rules on article style.

opposed to experimental science), due to the fundamental asymmetry of causation. A single cause will often have a great many disparate effects, which if taken together would specify the cause with virtual certainty. Since all the effects cannot actually be known (as some are lost in the historical/geological record), evidence must be carefully weighed to decide between competing hypotheses (the methodology sometimes known as "multiple working hypotheses"). Experimental sciences tend, on the other hand, to adhere more or less to a "predict and test" methodology, in which manipulative experiments are used to confirm, refute, or revise specific hypotheses (Cleland, 2002). We therefore hypothesize:

> $\mathbf{H_2}_a$: Writing in historical science has more features expressing the weight, validity, likelihood, or typicality of different assertions or pieces of evidence.
>
> $\mathbf{H_2}_b$: Writing in experimental science has more features typical of explicit reasoning about predictions and expectations, including the necessity or possibility of different events under different circumstances.

Next, as also discussed in Section 4, historical science primarily studies complex, ultimately unique, entities via observation and comparison, while experimental science studies uniform and interchangable entities via manipulative experimentation. The consequent manipulative nature of information-gathering in the experimental sciences will tend towards highly focused and controlled experimentation (indeed, this is considered a key quality of good methodology). Historical scientific methodology differs here, for two reasons. First, the limitation of evidence to that which is observed requires the broader context of any given observation to be explicitly noted and taken into account (as that context cannot be controlled or preserved). Second, since the objects of study are highly complex and individualized entities, rather than comparatively uniform entities or phenomena, true repeatability is not possible, and hence detailed comparison among similar and dissimilar observations is required to generalize across the data. Argumentation in the historical sciences thus may be expected to require a higher density of separate 'information units', than the more focused argumentation we expect in the experimental sciences. We therefore hypothesize:

> $\mathbf{H_3}_a$: Writing in historical science has more features indicating a greater number and variety of information units in the text, reflecting both 'synthetic thinking' about multiple hypotheses, comparing and contrasting evidence for all of them in parallel, as well as a concern with contextualizing information about varied and unique objects of study.

$\mathbf{H_{3}}_{b}$: Writing in experimental science has more features indicating more focused attention to a single (or small number of) ‘storylines’, reflecting a more focused approach of controlled manipulation of nature to test specific hypotheses about non-unique objects.

It is certainly clear that the mere presence or absence of a few linguistic features that can be linked to reasoning of a particular type is not by itself evidence of such reasoning. However, a consistent pattern of many of these features (as shown below) together aligned with the dichotomies proposed above strongly argues for such differences, which future research will attempt to elucidate in greater detail. The next section discusses in more detail methods of textual analysis that enable us to access relevant features of the text.

5 Methodology

5.1 The Corpus

The study reported here analyzes a corpus of recent (2003) articles drawn from twelve peer-reviewed journals in both historical and experimental sciences. The historical scientific journals in geology, evolutionary biology, and paleontology that we used are:

Journal of Geology includes research on the full range of geological principles including geophysics, geochemistry, sedimentology, geomorphology, petrology, plate tectonics, volcanology, structural geology, mineralogy, and planetary sciences.

Journal of Metamorphic Geology focuses on metamorphic studies[5], from the scale of individual crystals to that of lithospheric plates.

Biological Journal of the Linnean Society publishes work on organic evolution in a broad sense, particularly research unifying concepts of evolutionary biology with evidence from genetics, systematics, biogeography, or ecology.

Journal of Human Evolution covers all aspects of human evolution, including both work on human/primate fossils and comparative studies of living species.

---

[5] Metamorphism refers to changes in mineral assemblage and texture in rocks that have been subjected to temperatures and pressures different from those under which they originally formed.

Palaeontologia Electronica publishes papers in all branches of paleontology as well as related biological or paleontologically-related disciplines.

Quaternary Research published research in diverse areas in the earth and biological sciences which examine the Quaternary period of the Earth's history (from roughly 1.6 million years ago to the present).

The experimental scientific fields considered are physics, physical chemistry, and organic chemistry; the journals used in our corpus are:

Physics Letters A publishes research in a wide range of areas, including : condensed matter physics, theoretical physics, nonlinear science, statistical physics, mathematical and computational physics, atomic, molecular and cluster physics, plasma and fluid physics, optics, biological physics and nanoscience.

Physical Review Letters also covers a wide range of physics research, including: gravitation and astrophysics, elementary particles and fields, nuclear physics, atomic, molecular, and optical physics, nonlinear dynamics, fluid dynamics, plasma and beam physics, and condensed matter physics.

Journal of Physical Chemistry A publishes chemical research at the level of molecules (including dynamics, spectroscopy, gaseous clusters, molecular beams, kinetics, atmospheric and environmental physical chemistry, molecular structure, bonding, quantum chemistry, and general theory).

Journal of Physical Chemistry B publishes research on materials (including nanostructures, micelles, macromolecules, statistical mechanics and thermodynamics of condensed matter, biophysical chemistry, and general physical chemistry), as well as studies on the structure and properties of surfaces and interfaces.

Heterocycles publishes research in the areas of organic, pharmaceutical, analytical, and medicinal chemistry of heterocyclic compounds.

Tetrahedron publishes general experimental and theoretical research results in the field of organic chemistry and applications in related disciplines especially bio-organic chemistry.

The numbers of articles used from each journal and their average (preprocessed) lengths in words are given in Table 3.

Table 3: Journals used in the studies with number of articles and average length (in words) per article.

	Journal	# Art.	Avg. Len.
H_1	Journal of Geology	93	4891
H_2	Journal of Metamorphic Geology	108	5024
H_3	Biological Journal of the Linnean Society	191	4895
H_4	Human Evolution	169	4223
H_5	Palaeontologia Electronica	111	4132
H_6	Quaternary Research	113	2939
E_1	Physics Letters A	132	2339
E_2	Physical Review Let.	114	2545
E_3	Journal of Physical Chemistry A	121	4865
E_4	Journal of Physical Chemistry B	71	5269
E_5	Heterocycles	231	3580
E_6	Tetrahedron	151	5057

5.2 Computational methods

Research in *computational stylistics* seeks effective models of language style by applying machine learning algorithms to stylometrically meaningful features. In stylometrics, researchers have traditionally sought simple statistically valid models of stylistic distinctions, based on a small number (dozens, at most) of easily-computed textual statistics, such as word-frequencies (Mosteller & Wallace, 1964), phrase-type frequencies (Baayen, Halteren, & Tweedie, 1996), or sentence-complexity (Yule, 1938). Research on machine-learning techniques for text classification, on the other hand, has applied more sophisticated learning algorithms which can use combinations of many thousands of features to classify documents according to topic (see Sebastiani's (2002) excellent survey). Working systems that have been developed use a variety of modern machine learning techniques such as Naïve Bayes (Lang, 1995; Lewis, 1998), Winnow (Dagan, Karov, & Roth, 1997), and Support Vector Machines (Joachims, 1998). Recent work on applying machine learning and statistical methods to stylometric features for style analysis has achieved useful techniques for authorship attribution (Argamon, Koppel, & Avneri, 1998a; Stamatatos, Fakotakis, & Kokkinakis, 2000; Argamon, Šarić, & Stein, 2003), genre analysis (Biber, 1995; Matthews & Merriam, 1997; Argamon, Koppel, & Avneri, 1998b; Dimitrova, Finn, Kushmerick, & Smyth, 2002), and other applications (Holmes, 1998; Graham & Hirst, 2003; Koppel, Argamon, & Shimoni, 2003).

Linear separator learning algorithms, which do not require feature independence and are robust to presence of irrelevant features (such as Winnow and SVMs), tend to work quite well for stylistic categorization problems. The main current research issue in the field is the question of what kinds of textual features are good style discriminators, especially with the use of algorithms that can effectively deal with very large numbers of such features. Features for stylistic discrimination must be invariant as to topic but vary with the specific stylistic dimension under study. Our results and those of others (Finn, Kushmerick, & Smyth, 2002; Argamon, Koppel, Fine, & Shimony, 2003; Argamon, Šarić, & Stein, 2003; Graham & Hirst, 2003) have shown that using just relative frequencies of several hundred function words often gives excellent results, while adding syntactic or complexity-based features can sometimes be advantageous as well.

More recently, we and others (Herke-Couchman, Whitelaw, & Patrick, 2004; Whitelaw & Argamon, 2004; Whitelaw, Garg, & Argamon, 2005; Argamon et al., 2007) have successfully applied features based on SFL to stylistic text classification problems. The main idea (described in more detail in Section 5.3 below) is to build a lexicon of words and phrases associated with various options in a system network (this is only possible for those systems, such as EXPANSION, which are highly lexicalized). Counts of such 'indicators' are then used to estimate the relative contribution of various options within each system; those estimates are used as features for computing a classification model using machine learning.

5.3 Stylometric features

To apply machine learning over the corpus for text classification, each document was represented as a point in a high-dimensional space, that is, as a vector of numeric feature values. Our classification analysis was done using two feature sets (described in detail below): *function words*, a standard choice for stylistic text classification, and *systemic contribution* in three systems related to cohesion and modal assessment (as introduced in Section 3.2 above).

5.3.1 Function Words (FW)

As a standard stylistic baseline, we used a set of 546 function words (such as "the", "as", "of", "if", and "have") taken en masse from the stop-word list of the popular research information retrieval system AIRE (Grossman & Frieder, 2004); this procedure ensures task and theory neutrality. The set of function words used are similar to those used in many previous studies, such as Mosteller and Wallace's (1964) seminal stylometric work studying the authorship of the disputed Federalist Paper, for example. Under FW, each article is represented as a vector of 546 numbers between 0 and 1, each the relative frequency of one of the function words in the article.

5.3.2 System Contribution (SC)

The systemic features we used are based on options within three main systems related to cohesion and assessment (see Sec. 3.2), following Matthiessen's (1992) grammar of English, a standard SFL reference. Indicator lists were constructed by starting with the lists of typical words and phrases given by Matthiessen, and expanding them to related words and phrases taken from Roget's Interactive Thesaurus[6] (manually filtered for relevance). Keyword lists were constructed entirely independently of the target corpus, ensuring that we didn't 'cook the books'.

As discussed above in Section 3.2, we used systems and options within: EXPANSION (Fig. 4), linking clauses together (either within or across sentences); COMMENT (Fig. 5), expressing modal assessments of attitude or applicability; and MODALITY (Figs. 6 and 7), giving judgments regarding probability, usuality, inclination, and the like. MODALITY and COMMENT relate directly to how propositions are assessed in evidential reasoning (e.g., for likelihood, typicality, consistency with predictions, etc.), while EXPANSION is a primary system by which texts are constructed out of smaller pieces, and so will reflect choices of overall rhetorical structure[7].

We use the relative frequencies of sets of keywords and phrases which indicate that a particular part of the text realizes a certain system in the language. For example, an occurrence of the word 'certainly' usually indicates that the author is making a high-probability modal assessment of an assertion. Such a keyword-based approach has obvious practical advantages in the current absence of a reliable general systemic parser. The primary downside is the fact that a number of these keywords are ambiguous, since a keyword's precise interpretation can depend crucially on its context. By using as complete a set of such systemic indicators as possible for each system we represent, and then by using only measures of comparative frequency between such aggregated features, we hope to reduce the effect of ambiguity. In addition, since we use very large sets of indicators for each system, it is unlikely that such ambiguity would introduce a systematic bias, and so such noise is more likely to just reduce the clarity of any results instead of biasing them. The details of the systems we use in this study are as follows (refer to Section 3.3 for real-world examples).

EXPANSION On the discourse level, the system of EXPANSION serves to link a clause with its textual context, by denoting how the given clause expands on some aspect of its preceding context. Similar systems

---

[6] `http://www.thesaurus.com`

[7] Other textual/cohesive systems, such as PROJECTION, TAXIS, THEME, and INFORMATION (Halliday, 1994) are relevant, but could not be addressed in this work, due to the difficulty of automatically extracting them using current methods.

also operate at the lower levels of noun and verbal groups, ‘overloading’ the same lexical resources which, however, generally denote similar types of logico-semantic relationships, e.g., ‘and’ usually denotes “additive extension”. The three options within EXPANSION are Elaboration, Extension, and Enhancement. Each of these options has its own options, also used as features. (Note that the system network can be deepened further (Matthiessen, 1992, p. 521), but our keyword-based method allows only a relatively coarse analysis.) The EXPANSION system, with examples of systemic indicators we use, is shown in Figure 4.

Note that the actual features by which we represent an article are the frequencies of each option’s indicator features, each measured relative to its siblings. So, for example, one feature is Appositive Elaboration, whose value is the total number of occurrences of Appositive indicators divided by the total number of occurrences of Elaboration indicators (Appositive + Clarifying). The relative frequencies of Elaboration, Extension, and Enhancement within EXPANSION are also used as features.

COMMENT The system of Comment is one of modal assessment, comprising a variety of types of “comment” on a message, assessing the writer’s attitude towards it, or its validity or evidentiality. Comments are generally realized as adjuncts in a clause (and may appear initially, medially, or finally). Matthiessen (1992) lists eight types of Comment, given in Figure 5 along with representative indicators for each option.

MODALITY The features for interpersonal modal assessment that we consider here are based on Halliday’s analysis of the Modality system, as formulated by Matthiessen (1992). In this scheme, modal assessment is realized by a simultaneous choice of options within four systems as shown in Figure 6. The cross-product of all of these systems and subsystems creates a large number of potential modality assessment features, each of which is realized through a particular set of indicators (see Figure 7). We consider as simple features, each option in each system (e.g., the frequency of Modalization/Probability vs. Modalization/Usuality) as well as complex features made up of all the pairwise combinations of simple features from different systems (such as “Median Probability”). The indicator set for each such feature is the intersection of the indicator sets for the two component features. Frequencies were normalized by the total set of occurrences of both primary systems (Modalization and Value in the previous example).

- Elaboration: Deepening the content of the context
  - Appositive: Restatement or exemplification
    in other words, for example, to wit
  - Clarifying: Correcting, summarizing, or refocusing
    to be more precise, in brief, incidentally
- Extension: Adding new related information
  - Additive: Adding new content to the context
    and, moreover, furthermore
  - Adversative: Contrasting new information with old
    but, yet, however, on the other hand
  - Verifying: Adjusting content by new information
    instead, except for, alternatively
- Enhancement: Qualifying the context
  - Matter: What are we talking about
    here, as to that, in other respects
  - Spatiotemporal: Relating context to space/time
    - Simple: Direct spatiotemporal sequencing
      then, now, previously, lastly
    - Complex: More complex relations
      soon, that day, meanwhile, immediately
  - Manner: How did something occur
    in the same way, similarly, likewise
  - Causal/Conditional:
    - Causal: Relations of cause and effect
      so, therefore, for this reason
    - Conditional: Logical conditional relations
      then, in that case, otherwise

Figure 4: The EXPANSION system, options, and examples of indicators (in italics).

- Admissive: Message is assessed as an admission
  frankly, to tell the truth, honestly
- Assertive: Emphasizing the reliability of the message
  really, actually, positively, we confirm that
- Presumptive: Dependence on other assumptions
  evidently, presumably, reportedly, we suspect that
- Desiderative: Desirability of some content
  fortunately, regrettably, it was nice that, hopefully
- Tentative: Assessing the message as tentative
  tentatively, initially, depending on, provisionally
- Validative: Assessing scope of validity
  broadly speaking, in general, strictly speaking
- Evaluative: Judgement of actors behind the content
  wisely, sensibly, foolishly, justifiably, by mistake
- Predictive: Coherence with predictions
  amazingly, fortuitously, as expected

Figure 5: COMMENT options, and examples of indicators (in italics).

Type: What kind of modality?

- Modalization: How 'typical' is it?
  - Probability: How likely is it?
  - Usuality: How frequent/common is it?
- Modulation: Will someone/something do it?
  - Readiness: How ready is it (are they/am I)?
  - Obligation: Must it (they/I)?

Value: What degree of the relevant modality scale?

- Median: In the middle of the normal range
- High: More than normal
- Low: Less than normal

Orientation: Is the modality expressed objectively or subjectively?

- Objective: As an attribute of the clause
- Subjective: As someone's attitude (the writer or another)

Manifestation: How is the modality structurally realized?

- Implicit: Realized by an adjunct or finite verb within the modality-modified clause
- Explicit: Realized by a projective verb (such as 'think') whose complement is the modality-modified clause

Figure 6: MODALITY systems and options; any expression of modality simultaneously realizes options from all four systems of Type, Value, Orientation, and Manifestation.

			Type:Modalization		Type:Modulation	
			Probability	Usuality	Readiness	Obligation
Objective	Explicit	Median	*is likely*	*are frequent*	—	*is preferable*
Objective	Explicit	High	*is undeniable*	—	—	*is required*
Objective	Explicit	Low	*is possible*	*are infrequent*	—	*is permissible*
Objective	Implicit	Median	*probably*	*usually*	*eager to*	*ought to*
Objective	Implicit	High	*certainly*	*always*	*decided to*	*obliged to*
Objective	Implicit	Low	*maybe*	*seldom*	*allowed to*	*able to*
Subjective	Explicit	Median	*we believe*	—	*we prefer*	—
Subjective	Explicit	High	*we know*	—	—	*we require*
Subjective	Explicit	Low	*we suspect*	—	—	*we permit*
Subjective	Implicit	Median	*will*	*will*	*would rather*	*should*
Subjective	Implicit	High	*must*	*must*	*must, has to*	*ought to*
Subjective	Implicit	Low	*can, may*	*can, may*	*can, will*	*can, could*

Figure 7: Examples of indicator features for various combinations of MODALITY options. Note that not all combinations are realized in the language; note also the ambiguity of some of the indicators.

5.4 Machine learning for text categorization

The method was to represent each document as a numerical vector, each of whose elements is the relative frequency (as above) of a particular lexical feature of the text. We then applied the SMO learning algorithm (Platt, 1998) as implemented in Weka (Witten & Frank, 2000), using a linear kernel, no feature normalization, and the default parameters. (Other options did not appear to improve classification accuracy in early tests, so we used the simplest option, which also enables easy interpretation of results.) SMO is a support vector machine (SVM) algorithm (Cristianini & Shaw-Taylor, 2000); as noted above, SVMs have been previously applied successfully to text categorization problems (Joachims, 1998).

6 Results

6.1 Distinctiveness

We first consider hypothesis $\mathbf{H_1}$ (refer to Section 4.1 above), that experimental science texts are strongly distinguished from historical science texts, by estimating learning accuracy (via 10-fold cross-validation) for distinguishing articles in different journals. For each pair of two different journals, we used SMO (as above) to learn classifiers distinguishing articles in one journal from those in the other; the estimated learning accuracy is a measure of how distinguishable the journals are from one another (in terms of the features

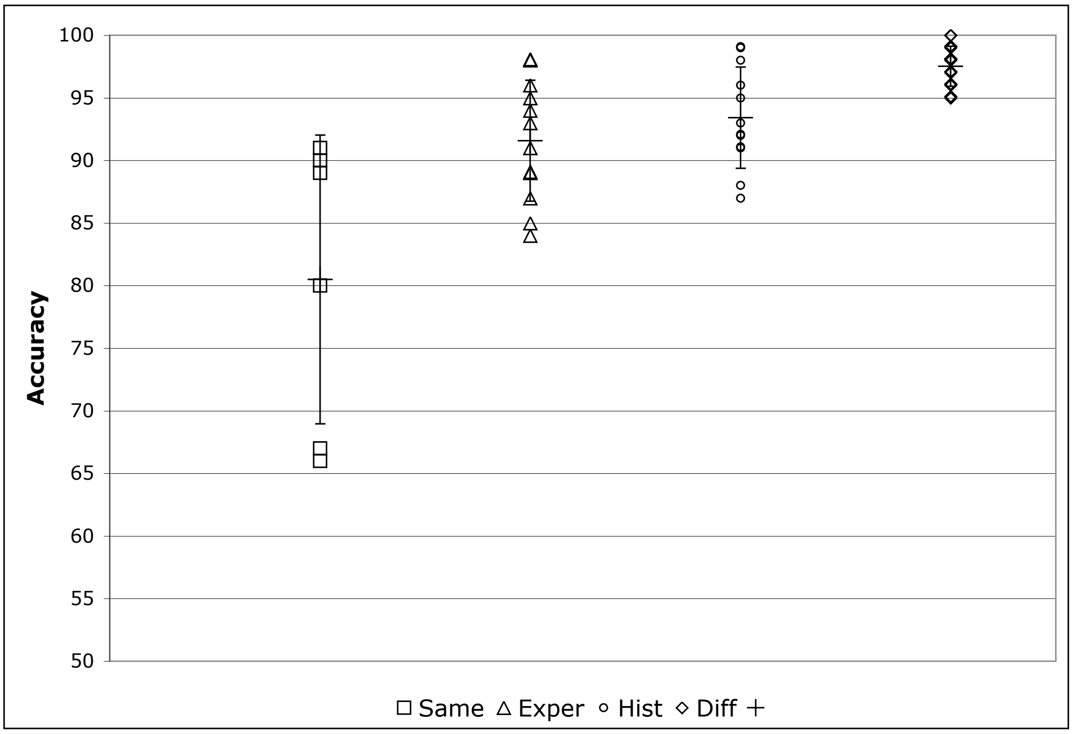

Figure 8: Learning accuracy for distinguishing articles in different pairs of journals using FW features. ‘Same’ are pairs where both journals are in the same field, ‘Exper’ pairs of journals in different experimental fields, ‘Hist’ pairs of journals in different historical fields, and ‘Diff’ pairs of journals where one is experimental and the other historical. Means and $\pm\sigma$ ranges are shown.

we have defined). These pairwise classification results are shown for feature sets FW and SC in Figures 8 and 9, respectively. Note that journal pairs from a single field are the least distinguishable, followed by those from two different (out of three) historical fields, then from two different experimental fields, and finally the most distinctive are journal pairs where one journal is historical and the other experimental. While there is some variance within each group, the difference between the “Diff” pairs and the pool of all “Similar” pairs (i.e., the union of “Same”, “Exper” and “Hist”), for both feature sets, is quite noticeable and statistically significant (FW Mann-Whitney $U = 940.5$, $p < 0.0001$; SC Mann-Whitney $U = 915.5$, $p < 0.0001$).

We further considered 10-fold cross-validation for learning over the entire corpus of articles from 12 journals. Average accuracy was 81.6%. To calibrate these results, we ran the same discrimination test for

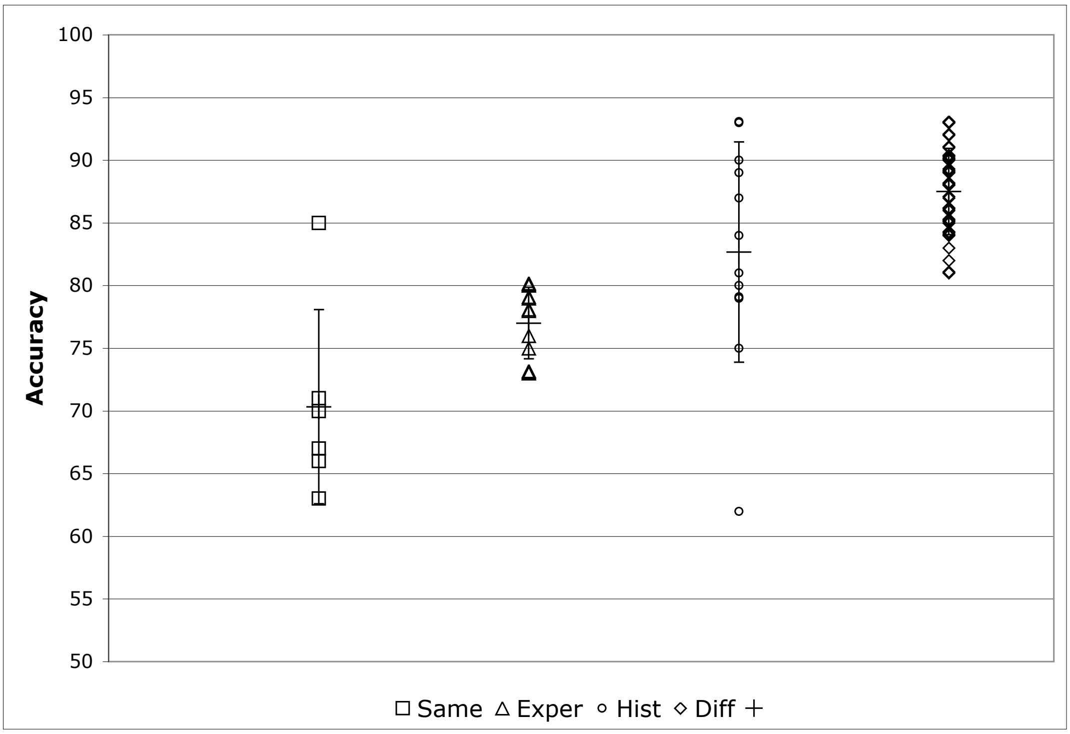

Figure 9: Learning accuracy (as in Fig. 8) using SC features.

all 462 different partitions of the twelve journals into two groups of six journals each, which gave a mean accuracy of 65.4% (s.d. 4.9%, range 55.4% to 81.6%). The highest accuracy is thus attained specifically for the partition between experimental and historical fields, which has $p < 0.0022$. The linguistic difference between experimental and historical science articles is thus highly unlikely to be due to chance.

Hypothesis $\mathbf{H_1}$ is thus confirmed, in that articles from different kinds of science are more easily distinguished from one another stylistically than are articles from a single kind of science (from different journals in one field or from totally different fields).

6.2 Distinguishing features

Given this result, we now consider what consistent pattern of distinguishing features, if any, emerges. That is, what system contribution (SC) features can be said to consistently indicate either historical or experimental science articles. To do this, we ran SMO (with a linear kernel) on all training data for each pair of a historical science journal with an experimental science journal (36 pairs in all), and ranked the features by their weight for one or the other journal in the weight vector computed by SMO. Tables 10, 11, and 12 show, for each SC feature f, for how many classification pairs (out of 36) f was weighted as an indicator feature for Experimental science, and for how many f was weighted as an indicator for Historical science. High values (≥ 25) are boldfaced, for ease of reference. Features which are indicators for one kind of science consistently over many pairwise comparisons have high values; we are mainly concerned with such features.

Examination of these results reveals some striking patterns, in the form of meaningful **oppositions**, where one option in a particular system is indicative of one article class (either experimental or historical science) and another option of that system is indicative of the other class. Such an opposition indicates a meaningful linguistic difference between the classes of articles, in that each prefers a distinctive way (each's preferred option) of expressing the same general meaning (the parent system). This notion derives from the key fact that even if a particular second-level option indicates a certain class (say Additive Extension for experimental articles) it does not necessarily mean that its parent (in this case, Extension) will be indicative of the same class. This is because options are only compared within a single system.

First, consider the system of EXPANSION. At the top level, we see an opposition between Extension, indicating historical science, and Enhancement, indicating experimental science. This implies that historical science articles generally have a higher density of discrete informational items, whereas experimental science articles tend to have fewer information items, which may have their meaning deepened or qualified by informationally-related clauses. This appears to reflect differing principles of rhetorical organization (as seen in the example passages in Section 3.3)—experimental scientists may prefer a single coherent 'story line' focused on enhancements of a small number of focal propositions, whereas historical scientists may prefer a multifocal 'landscape' of connected propositions. This supports hypotheses $\mathbf{H_3}_a$ and $\mathbf{H_3}_b$, respectively, comparing the contextual examination of various and highly unique entities by historical science with the more universalist, hence focused, examination of generic entities by experimental science.

Two other oppositions are visible in the 'subsystems' of Extension and Enhancement as well, which, while not bearing directly on our hypotheses, seem related. Within Extension, we see an experimental preference for Additive Extension, compared with a historical preference for Adversative Extension. This further supports the notion that historical scientists use a "multiple working hypotheses" method, in which comparison of alternatives is key (cf. $\mathbf{H_3}_a$). Further, we see that SpatioTemporal relationships form a core type

Expansion			Comment			Modality (simple)		
	Exper.	Hist.		Exper.	Hist.		Exper.	Hist.
Elaboration	12	24	**Admissive**	6	**30**	Type		
- **Apposition**	**28**	8	Assertive	18	18	- **Modalization**	0	**36**
- Clarification	18	18	Desiderative	13	23	- - Probability	16	20
Extension	10	**26**	Evaluative	22	14	- - Usuality	22	14
- **Additive**	**26**	10	**Predictive**	**30**	6	- **Modulation**	**35**	1
- **Adversative**	6	**30**	Presumptive	23	13	- - **Obligation**	7	**29**
- Verifying	18	18	Tentative	17	19	- - **Readiness**	**26**	10
Enhancement	**31**	5	**Validative**	4	**32**	Value		
- **Matter**	7	**29**				- Median	18	18
- **Spatiotemporal**	**26**	10				- **High**	**27**	9
- - Simple	22	14				- Low	12	24
- - **Complex**	10	**26**				Orientation		
- Manner	23	13				- **Objective**	5	**31**
- CausalConditional	24	12				- **Subjective**	**31**	5
- - **Causal**	**28**	8				Manifestation		
- - Conditional	13	23				- Implicit	21	15
						- Explicit	17	19

Figure 10: Number of times different systemic features were indicators for either Experimental or Historical journals in the 36 classification runs. For ease of reference, values ≥ 25 (a heuristic value) are boldfaced, to show the most salient oppositions found.

of Enhancement for experimental scientists (presumably describing forms and relationships of experimental subjects and results), while historical scientists do Enhancement more via Matter and so contextualize their statements more (cf. $\mathbf{H_{2}}_{a}$).

Posited methodological distinctions between the kinds of science are further supported by our results for Comment. Here we see preference for Validative and Admissive Comments by historical scientists compared to a very strong consistent preference for Predictive Comments by experimental scientists. The latter result is a straightforward consequence (hypothesis $\mathbf{H_{2}}_{b}$) of the experimental scientist's focus on predictive consistency. The historical scientist, on the other hand, evinces a rhetorical need (via Validative Comments) to explicitly delineate the scope of validity of different assertions (hypothesis $\mathbf{H_{2}}_{a}$), likely as a consequence of synthetic thinking (Baker, 1996) about complex and ambiguous webs of past causation (Cleland, 2002). An Admissive comment marks a clause as the opinion (perhaps strongly held) of the author; this too appears indicative of a more hedged and explicitly comparative argumentation style.

Finally, we consider the (more complicated) system of Modality, starting with the coarse level represented by the simple features. The primary opposition is in modality Type. The preference of experimental science writing for Modulation (assessing what 'must' or 'is able' to happen) is supportive of hypothesis

			Value			Orientation		Manifestation	
			Med.	Low	High	Obj.	Subj.	Impl.	Expl.
Type	Modalization		9	9	11	0	22	0	15
	Modalization	Probability	2	**29**	**27**	15	**34**	3	**32**
	Modalization	Usuality	14	**27**	20	4	**32**	**34**	0
Type	Modulation		**27**	**33**	22	**31**	24	**35**	22
	Modulation	Readiness	13	**29**	1	20	21	**26**	
	Modulation	Obligation	**32**	12	9	**33**	1	5	20
Manifestation		Implicit	15	22	**25**	12	**28**		
		Explicit	19	11	**27**	13	18		
Orientation		Objective	12	20	**29**				
		Subjective	24	19	12				

Figure 11: Feature weightings for paired MODALITY features indicative of experimental journal articles. Each cell represents a pair-wise feature for the conjunction of two independent attributes of MODALITY; the number in the cell shows in how many models the feature was indicative of the experimental journal (out of all 36 pairs of an experimental and a historical journal). For example, Medium Modulation values were indicative of experimental fields 27 times out of 36, while Medium values for Objective modality were indicative of Experimental fields just 12 times.

			Value			Orientation		Manifestation	
			Med.	Low	High	Obj.	Subj.	Impl.	Expl.
Type	Modalization		**27**	**27**	**25**	**36**	14	**36**	21
	Modalization	Probability	**34**	7	9	21	2	**33**	4
	Modalization	Usuality	22	9	16	**32**	4	2	**36**
Type	Modulation		9	3	14	5	12	1	14
	Modulation	Readiness	23	7	**35**	16	15	10	
	Modulation	Obligation	4	24	**27**	3	**35**	**31**	16
Manifestation		Implicit	21	14	11	24	8		
		Explicit	17	**25**	9	23	18		
Orientation		Objective	24	16	7				
		Subjective	12	17	24				

Figure 12: Feature weightings for complex MODALITY features indicative of historical journal articles, formatted as in Figure 12 above.

$\mathbf{H_2}_b$ and consistent with a focus on prediction and manipulation of nature. Concurrently, historical science writing shows a preference for Modalization (assessing 'likelihood' or 'usuality'), supporting hypothesis $\mathbf{H_2}_a$. Such a preference is consistent with the outlook of an observer who usually cannot directly manipulate or replicate outcomes, and therefore (i) cannot make unqualified statements of what must (or must not) happen, and (ii) uses therefore the method of "multiple working hypotheses".

Within Modulation, we see also a more delicate opposition, where experimental science articles prefer Readiness (possibility), and historical science articles prefer Obligation (necessity). To understand this, note first that either kind of assessment opens up a clause for value-based consideration; if a statement is to be simply accepted, on the other hand, modal assessment will typically be avoided (compare "John went to the store" with "John should have gone to the store"; the first is either true or false, whereas the second makes a value judgement that can be disputed). Thus experimental scientists open Readiness up for consideration and possible disputation by the reader, while the default is for events to be necessary (and so they need not be noted explicitly as such). For historical scientists, on the other hand, necessity cannot be assumed, rather Obligation is noted openly in the text, allowing dispute, while the possible and contingent is assumed as the default (and nuanced levels of probability are expressed by Modalization, as noted above).

Digging still deeper, related patterns are also seen within the paired features conjoining two independent dimensions of MODALITY (Figures 11 and 12). Some results here simply confirm the one-dimensional results discussed above, though others give a more detailed view. First, consider oppositions between Implicit and Explicit variants of different MODALITY Types. Implicit realizations are more likely to be used for options that are well-integrated into the expected rhetoric than corresponding Explicit variants, which have more attention drawn towards them and so are more likely to be used for less usual modal assessments. First, we see Implicit Modulation for experimental science writing, and Implicit Modalization for historical science writing. This implies that Modalization is integrated smoothly into the overall environment of historical scientific rhetoric, whereas Modulation is a normal part of the rhetorical environment of experimental science. This is further support for hypotheses $\mathbf{H_2}_a$ and $\mathbf{H_2}_b$, emphasizing the focus of experimental science on possibility/necessity and that of historical science on relative likelihood and typicality.

The evident preference for Implicit and Subjective Usuality in experimental science is probably explained by the fact that most keywords used to indicate these options are ambiguous, and may very well be indicating other options in MODALITY. This is likely connected to the historical preference for Objective Usuality, which may just indicate that unambiguous Usuality (as a clear option of Modalization) is an indicator for historical science writing. The historical preference for Implicit Probability, on the other hand, seems to indicate more directly the importance in historical methodology (noted above) of weighting propositions by likelihood ($\mathbf{H_2}_a$), by placing such assessment in the rhetorical background (as above).

Some other interesting oppositions may also be observed in these results. One is an opposition within Modalization, of Implicit Readiness (experimental) vs. Implicit Obligation (historical). This likely reflects the overall opposition (noted above) between Readiness and Obligation. Too, an opposition is observed between Objective Obligation (experimental) and Subjective Obligation (historical) paralleled by an opposition between Subjective Usuality (experimental) and Objective Usuality (historical). This appears to reflect the choice of Objective realization for types of modality which is rhetorically central (Modulation/Obligation for experimental science, Modalization/Usuality for historical science, as discussed above) and Subjective for less central types of assessment.

7 Conclusions

In this paper, we have provided the first empirical evidence (however indirect) for methodological variation among the sciences, by analysis of their main product, the peer-reviewed research article. Application of machine learning techniques to linguistically well-motivated textual features has provided clear empirical evidence for rhetorical differences between writing in different scientific fields, supporting recent claims by philosophers of science for a broad division of sciences into ‘experimental’ and ‘historical’ fields. Moreover, analysis of the models output by the learning procedure gives insight into which language features most consistently realize the differences in functional text type corresponding to different kinds of science. The main linguistic preferences identified can be directly linked with the particular modes of reasoning posited by philosophers for these different kinds of science. This study thus lends empirical weight to the argument for a multiplicity of methods in science, rather than a single monolithic “scientific method”.

In addition to the philosophical and linguistic implications of our results, we believe that this line of research may have meaningful application for improving science education. By precisely articulating some of the specific language skills necessary for effective communication in different styles of science, we may develop methods to improve teaching of such skills to scientific novices both at the high school and university levels. Such skills are a key part of a greater set of competencies needed for pursuing scientific research in inquiry-based environments (AAAS, 1990; NRC, 1996). It may also be possible to design novel assessment methods which evaluate how a student develops the conceptual structure of a field, by measuring use of the corresponding linguistic and rhetorical structures.

Future work

Immediate future work includes validating these results against a larger corpus of articles including more scientific fields. In particular, potentially ‘mixed’ fields, such as astrophysics, need to be studied, to examine more precisely the interface between historical and experimental scientific language. Another important area for future research is suggested by the fact that while the current study treats each article as a single indivisible whole, the rhetorical organization of an article varies strongly among the different sections of the text, as noted by Lewin et al. (1986) in their analysis of social science texts; Introductions differ from Materials and Methods sections, which differ from Results sections, and so on. We thus intend next to study linguistic variation across different sections of individual texts within and across fields. Techniques such as those of Teufel and Moens (1998, 2002) for determining rhetorical saliency and function of sentences in scientific texts may be applicable.

Too, as a complement to this work on written communication, we are beginning to study historical scientists *in vivo*, as they pursue their research (in real time) within the authentic setting of the field and the laboratory. Results of this work will be compared with those of other similar studies focusing on experimental scientists (Dunbar, 1995, 1999; Dunbar & Blanchette, 2001; Stucky & Bond-Robinson, 2004; Nersessian, 2005). Ultimately, this work will provide a fuller account of the relation between language and methodological reasoning in different fields of science. Such expert models of reasoning and communication are critical to current work in science education, as they will enable researchers and curriculum designers to better model the inquiry-based materials currently being proposed and implemented in science curriculum reform.

On the technical side, we are also working on refining automated methods for linguistic processing, first on refining identification of systemic indicators, and then linking individual indicators with their context for greater information. In this regard, lexical chains as used in text summarization (Barzilay & Elhadad, 1999; Harabagiu, 1999; Brunn, Chali, & Pinchak, 2001) may be useful, as may recent work on automated rhetorical structure analysis of texts (Hovy, 1993; Marcu, 2000). Methods for discovering important features of opinion expression might also help, such as subjectivity collocations (Wiebe, 2000; Riloff, Wiebe, & Wilson, 2003).

Acknowledgements

We would like to thank Martha Evens, Jonathan Fine, Peter Greene, and Don Swanson for their helpful comments on earlier drafts of this paper, as well as the anonymous reviewers. Especial thanks to the late Nambury Raju for his statistical advice and friendship.

References

Abrams, E., & Wandersee, J. H. (1995). How does biological knowledge grow? A study of life scientists' research practice. *Journal of Research in Science Teaching*, *32*(6), 643–63.

American Association for the Advancement of Science. (1990). *Benchmarks for science literacy*. New York: Oxford University Press.

Argamon, S., Koppel, M., & Avneri, G. (1998a). Routing documents according to style. In *Proc. int'l workshop on innovative internet information systems.* Pisa, Italy.

Argamon, S., Koppel, M., & Avneri, G. (1998b). Style-based text categorization: What newspaper am i reading? In *Proc. AAAI workshop on learning for text categorization* (pp. 1–4).

Argamon, S., Koppel, M., Fine, J., & Shimony, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, *23*(3).

Argamon, S., Šarić, M., & Stein, S. S. (2003). Style mining of electronic messages for multiple author discrimination. In *Proc. ACM conference on knowledge discovery and data mining.*

Argamon, S., Whitelaw, C., Chase, P., Dhawle, S., Garg, N., Hota, S. R., et al. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science*. ((In press))

Baayen, R. H., Halteren, H. van, & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *7*, 91-109.

Baker, V. R. (1996). The pragmatic routes of American Quaternary geology and geomorphology. *Geomorphology*, *16*, 197-215.

Barzilay, R., & Elhadad, M. (1999). Using Lexical Chains for Text Summarization. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 111–121). The MIT Press.

Bazerman, C. (2004). What activity systems are literary genres part of? *Journal of the Interdisciplinary Crossroads*, *1*(3).

Bazerman, C., & Prior, P. (2005). Participating in emergent socio-literate worlds: Genre, disciplinarity, interdisciplinarity. In Multidisciplinary perspectives on literacy research. Hampton Press.

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

Bond-Robinson, J., & Stucky, A. P. (2005). Grounding scientific inquiry and knowledge in situated cognition. In Proceedings of the 27th annual cognitive science society. Stresa, Italy.

Brunn, M., Chali, Y., & Pinchak, C. J. (2001). Text Summarization Using Lexical Chains. In Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. New Orleans, LA.

Cleland, C. E. (2001). Historical science, experimental science, and the scientific method. Geology, 29(11), 987-990.

Cleland, C. E. (2002). Methodological and epistemic differences between historical science and experimental science. Philosophy of Science.

Cooper, R. A. (2002). Scientific knowledge of the past is possible: Confronting myths about evolution and the nature of science. The American Biology Teacher, 64, 476-481.

Cooper, R. A. (2004). Teaching how scientists reconstruct history: Patterns and processes. The American Biology Teacher, 66(2), 101-108.

Cristianini, N., & Shaw-Taylor, J. (2000). An introduction to support vector machines. Cambridge Press.

Cronin, B. (2005). The Hand of Science: Academic Writing and Its Rewards. Scarecrow Press.

Cronin, B., & Overfelt, K. (1994). Citation-based auditing of academic performance. Journal of the American Society for Information Science, 45(2), 61–72.

Dagan, I., Karov, Y., & Roth, D. (1997). Mistake-driven learning in text categorization. In C. Cardie & R. Weischedel (Eds.), Proceedings of EMNLP-97, 2nd conference on empirical methods in natural language processing (pp. 55–63). Providence, US: Association for Computational Linguistics, Morristown, US.

Diamond, J. (2002). Guns, germs and steel: The fates of human societies. New York: W.W. Norton.

Dimitrova, M., Finn, A., Kushmerick, N., & Smyth, B. (2002). Web genre visualization. In Proc. conference on human factors in computing systems.

Dodick, J. T., & Orion, N. (2003). Geology as an historical science: Its perception within science and the education system. Science and Education, 12(2).

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In *Mechanisms of insight* (p. 365-395). Cambridge MA: MIT Press.

Dunbar, K. (1999). The scientist InVivo: How scientists think and reason in the laboratory. In *Model-based reasoning in scientific discovery.* Plenum Press.

Dunbar, K. (2001). What scientific thinking reveals about the nature of cognition. In *Designing for science* (pp. 115–140). Mahwah, NJ: Lawrence Erlbaum Associates.

Dunbar, K., & Blanchette, I. (2001). The invivo/invitro approach to cognition: The case of analogy. *Trends in Cognitive Sciences*, *5*, 334-339.

Finn, A., Kushmerick, N., & Smyth, B. (2002). Genre classification and domain transfer for information filtering. In F. Crestani, M. Girolami, & C. J. van Rijsbergen (Eds.), *Proceedings of ECIR-02, 24th european colloquium on information retrieval research.* Glasgow, UK: Springer Verlag, Heidelberg, DE.

Frodeman, R. (1995). Geological reasoning: Geology as an interpretive and historical science. *Geological Society of America Bulletin*, *107*, 960-968.

Fujimura, J. H. (1987). Constructing ‘do-able’ problems in cancer research: Articulating argument. *Social Studies of Science*, *17*, 257–293.

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96(3)*, 606-633.

Goodwin, C. (1995). Seeing in depth. *Social Studies of Science*, *25*, 237-74.

Gould, S. J. (1986). Evolution and the triumph of homology, or, why history matters. *American Scientist*, *Jan.-Feb.*, 60-69.

Graham, N., & Hirst, G. (2003). Segmenting a document by stylistic character. In *Workshop on computational approaches to style analysis and synthesis, 18th international joint conference on artificial intelligence.* Acapulco.

Grossman, D., & Frieder, O. (2004). *Information retrieval: Algorithms and heuristics* (Second ed.). Springer.

Halliday, M. A. K. (1994). *Introduction to functional grammar* (Second ed.). Edward Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. London: Longman.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. London: Falmer.

Harabagiu, S. (1999). From Lexical Cohesion to Textual Coherence: A Data Driven Perspective. *Journal of Pattern Recognition and Artificial Intelligence*, *13(2)*(4), 247–265.

Harris, J. (1989). The Idea of Community in the Study of Writing. College Composition and Communication, 40(1), 11–22.

Hasan, R. (1988). Language in the process of socialisation: Home and school. In Language and socialisation: Home and school. North Ryde, N.S.W.: Macquarie University.

Herke-Couchman, M., Whitelaw, C., & Patrick, J. (2004). Identifying interpersonal distance using systemic features. In Proc. AAAI spring symposium on exploring attitude and affect in text: Theories and applications.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing, 13(3), 111–117.

Hovy, E. (1993). In Defense of Syntax: Informational, Intentional, and Rhetorical Structures in Discourse. In O.Rambow (Ed.), Intentionality and structure in discourse relations (pp. 35–39). Ohio.

Hull, D. (1973). Darwin and his critics: The reception of Darwin's theory of evolution by the scientific community. Cambridge: Harvard University Press.

Hyland, K. (2000). Disciplinary Discourses: Social Interactions in Academic Writing. Longman.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Machine learning: ECML-98, tenth european conference on machine learning (p. 137-142).

Kelly, G. J., & Bazerman, C. (2003). How students argue scientific claims. Applied Linguistics, 24(1), 28–55.

Killingsworth, M., & Gilbertson, M. (1992). Signs, Genres, and Communities in Technical Communication. Baywood Pub. Co.

Kitcher, P. (1993). The advancement of science. New York: Oxford University Press.

Koppel, M., Argamon, S., & Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4).

Lang, K. (1995). NewsWeeder: learning to filter netnews. In Proceedings of the 12th international conference on machine learning (pp. 331–339). Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

Latour, B., & Woolgar, S. (1986). Laboratory life: The construction of scientific facts. Princeton, NJ: Princeton University Press.

Lewin, B. A., Fine, J., & Young, L. (1986). Expository discourse: A genre-based approach to social science research texts. Continuum.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), Proceedings of ECML-98, 10th european conference on machine learning (pp. 4–15). Chemnitz, DE: Springer Verlag, Heidelberg, DE.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. Scientometrics, 36(3).

Mann, W., & Thompson, S. (1988). Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. Text, 8(3), 243–281.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: A surface-based approach. Comp. Ling., 26(3), 395-448.

Matthews, R. A. J., & Merriam, T. V. N. (1997). Distinguishing literary styles using neural networks. In Handbook of neural computation (chap. 8). IOP publishing and Oxford University Press.

Matthiessen, C. (1992). Lexicogrammatical cartography: English systems. Tokyo, Taipei and Dallas: International Language Sciences Publishers.

Mayr, E. (1976). Evolution and the diversity of life. Cambridge: Harvard University Press.

Mayr, E. (1985). How biology differs from the physical sciences. In Evolution at the crossroads: The new biology and the new philosophy of science (p. 43-46). Cambridge: MIT Press.

Mosteller, F., & Wallace, D. L. (1964). Inference and disputed authorship: The federalist. Massachusetts: Addison-Wesley.

Mulkay, N., & Gilbert, G. N. (1983). Scientist's theory talk. Canadian Journal of Sociology, 8, 179–197.

Myers, G. (1990). Writing Biology: Texts in the Social Construction of Scientific Knowledge. University of Wisconsin Press.

National Research Council. (1996). National science education standards. Washington, DC: National Academy Press.

Nersessian, N. J. (2005). Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In M. Gorman, R. Tweney, D. Gooding, & A. Kincannon (Eds.), Scientific and technological thinking (pp. 17–56). New York: Erlbaum Press.

Ochs, E., & Jacoby, S. (1997). Down to the wire: The cultural clock of physicists and the discourse of consensus. Language in Society, 26(4), 479-506.

Ochs, E., Jacoby, S., & Gonzales, P. (1994). Interpretive journeys: How physicists talk and travel through graphic space. Configurations, 1, 151-171.

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. Cognitive Science, 21(2), 109–146.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Redmond, WA: Microsoft Research Technical Report MSR-TR-98-14.

Plum, G., & Cowling, A. (1987). Social constraints on grammatical variables: Tense choice in english. In Language topics (Vol. 2). Amsterdam: John Benjamins.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In Proceedings of conll-2003 (pp. 25–32). Edmonton, Canada.

Rudolph, J. L., & Stewart, J. (1998). Evolution and the nature of science: On the historical discord and its implication for education. Journal of Research in Science Teaching, 35, 1069-1089.

Rudwick, M. J. S. (1998). Lyell and the principles of geology. In D. Blundell & A. Scott (Eds.), Lyell: The past is the key to the present (pp. 3–15). Geological Society of London. (Special Publications 143)

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1).

Sober, E. (1993). Philosophy of biology. Boulder, CO: Westview Press.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. K. (2000). Automatic text categorization in terms of genre, author. Computational Linguistics, 26(4), 471-495.

Stucky, A. P., & Bond-Robinson, J. (2004). Empirical studies of scientists at work: Analysis of Authentic Inquiry experiences. In Proceedings of the national association of research in science teaching annual meeting. Vancouver, BC, Canada.

Swales, J. M. (1990). Genre analysis. Cambridge University Press.

Teufel, S., & Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. In Proc. AAAI spring symposium on intelligent text summarization.

Teufel, S., & Moens, M. (2002). Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status. Computational Linguistics, 28(4).

Whewell, W. (1837). History of the inductive sciences. London: John W. Parker.

White, H. D., & McCain, K. W. (1989). Bibliometrics. Annual Reviews of Information Science and Technology, 24, 119–186.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline. an author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 49, 327–355.

Whitelaw, C., & Argamon, S. (2004, October). Systemic functional features in stylistic text classification.

In Proc. AAAI fall symposim on style and meaning in language, art, music, and design. Washington, DC.

Whitelaw, C., Garg, N., & Argamon, S. (2005, November). Using appraisal taxonomies for sentiment analysis. In Proceedings of the ACM conference on information and knowledge management. Bremen, Germany.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In Proceedings of the seventeenth national conference on artificial intelligence and twelfth conference on innovative applications of artificial intelligence (pp. 735–740). AAAI Press / The MIT Press.

Witten, I. H., & Frank, E. (2000). Data mining: Practical machine learning tools with java implementations. San Francisco: Morgan Kaufmann.

Yule, G. (1938). On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. Biometrika, 30, 363-390.