Conjunction and Modal Assessment in Genre Classification:
A Corpus-Based Study of Historical and Experimental Science Writing

Shlomo Argamon
Dept. of Computer Science
Illinois Institute of Technology
Chicago, IL 60645
argamon@iit.edu

Jeff Dodick
Department of Learning Sciences
Northwestern University
Evanston, IL 60208
j-dodick@northwestern.edu

Abstract

We use textual features motivated by systemic functional linguistic theory for genre-based text categorization. We have developed feature sets representing different types of conjunctions and modal assessment, which together indicate (partially) how different genres structure texts and express attitudes towards propositions in the text. Using such features enables analysis of large-scale rhetorical differences between genres by examining which features are important for classification. The specific domain studied comprises scientific articles in historical and experimental sciences (paleontology and physical chemistry respectively). The SMO learning algorithm with our feature set achieved over 83% accuracy for classifying articles according to field, though no field-specific terms were used as features. The most highly-weighted features were consistent with hypothesized methodological differences between historical and experimental sciences, thus lending empirical evidence to the notion of multiple scientific methods.

Introduction

Research in automatically using features of a text to determine its stylistic character has a long history, going back to Mosteller and Wallace's landmark study on authorship attribution of the *Federalist Papers* (Mosteller & Wallace 1964). The kind of features typically used in such studies has not changed much over the years; the primary types of features are: frequencies of different function words (Mosteller & Wallace 1964), frequencies of different kinds of syntactic constituents (Baayen et al. 1996; Stamatatos et al. 2001), and various measures of sentence complexity (Yule 1938; Losee 1996). The implicit understanding is that such lexical, syntactic, or complexity-based features serve as useful proxies for the 'behind-the-scenes' pragmatic or contextual factors that determine stylistic variation, because such factors are realized in specific texts via word choice and syntactic structure. However, for the most part, computational stylistics research has not explicitly examined the semantics or pragmatics of such features. In particular, although it is clear that different textual genres use different rhetorical modes and different generic document structures (Martin 1992), the relation of low-level stylistic features to aspects of rhetorical structure remains obscure. Our general approach is to elucidate this relationship by exploring genre-based text categorization problems where particular rhetorical or cognitive communicative needs can be identified. In the future, this approach may be enhanced by incorporating rhetorical 'parsing' of the text (see the recent work of Marcu (2000)).

In this paper we outline an approach to defining linguistically-motivated features for genre classification based on systemic functional principles, and present an initial implementation of the approach. The specific domain we study here is that of scientific discourse, with the aim of gaining a better understanding of the nature and methods of science. We apply our system to a corpus-based study of genre variation between articles in a historical science (paleontology) and an experimental science (physical chemistry), where we expect to find significant rhetorical differences. Our results show how computational stylistic techniques can give a consistent picture of differences in reasoning between scientists in the two fields. This line of research also has the potential to contribute to the philosophy of science, by enabling empirical investigation of hypothesized methodological differences among fields. Also, since properly understanding different forms of scientific reasoning is critical for science education (Dodick & Orion 2003), we hope to contribute eventually to the development of better pedagogical practices.

System preference, genre, and register

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. For example: "A message is either about doing, thinking, or being; if about doing, it is either standalone action or action on something; if action on something it is either creating something or affecting something pre-existent," and so on. A *system*, then, is a set of options for meanings to be expressed, with *entry conditions*, i.e, 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 term here *subsystems*.

By structuring language as a structure involving a complex of choices between mutually exclusive options, the systemic approach is particularly appropriate to examining variation in language use. A systemic specification allows 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, 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 (Halliday 1991). By comparing how probabilities vary between individuals or situations with different characteristics, we may determine how those characteristics affect linguistic behavior.

By examining differences between systemic preferences across scientific genres, we are performing a quantitative analysis of register. *Register* denotes functional distinctions in language use related to the context of language use (Eggins & Martin 1997), and may be considered to comprise: *mode*, the communication channel of the discourse (interaction between producer and audience); *tenor*, the effect of the social relation between the producer and the audience; and *field*, the domain of discourse. This paper focuses on the field-related distinction between historical and experimental science, with mode and tenor held relatively constant, by using articles written by working scientists drawn from peer-reviewed journals. Our hypothesis, borne out by our results below, is that the difference in the types of reasoning needed by historical and experimental sciences leads to correlated differences in rhetorical preferences (perhaps best understood as 'functional tenor' (Gregory 1967)), which are realized by how the writer expresses attitudes towards assertions in the text (via modal assessment), as well as by what strategies they use for cohesion.

Previous work in this vein has investigated the relationship between choice probabilities and contextual factors. For example, Plum & Cowling (1987) demonstrate a relation between speaker social class and choice of verb tense (past/present) in face-to-face interviews. Similarly, Hasan 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 (Hasan 1988). The methodology that has been applied in these works involves first hand-coding a corpus for systemic-functional and contextual variables and then comparing how systemic choice probabilities vary with contextual factors, using correlation statistics or multivariate analysis techniques (such as principal components analysis). In this paper we present a first attempt at extending this idea to larger corpora using automated machine learning methods.

Scientific language

Our domain of interest in this study is scientific discourse, which we approach by examining peer-reviewed journal articles. Our goal in analyzing scientific communication is to study the nature and methods of science. Paradigmatic of this general approach is research on how scientists communicate with each other while doing science. It is clear that communication between the various scientists working in a laboratory is often crucial to scientific success (Dunbar 1995). The particular uses of language by scientists serve to create a sort of "collaborative space", whose background worldview makes possible communication about complex observations and hypotheses (Goodwin 1994). Analysis of specific linguistic features can help elucidate important features of the way discourse contributes to problem solving, as in the study by Ochs et al. (1994) of physicist's metaphoric talk of travel in a variety of graphical spaces.

Historical and experimental science

Increasingly, philosophers of science recognize that the classical model of a single "Scientific Method" (based on experimental sciences such as physics) does a disservice to sciences such as geology and paleontology, which are no less scientific by virtue of being historically oriented. Rather, differences in method may stem directly from the types of phenomena under study. Experimental science (such as physics) attempts to formulate general predictive laws, and so relies heavily on repeatable series of manipulative experiments to refine hypotheses about such laws (Latour & Woolgar 1986). Historical science, on the other hand, deals with *contingent* phenomena, involving study of specific individuals and events from the past, in an attempt to find unifying explanations for effects caused by those events (Mayr 1976). Because of this, reasoning in historical sciences is understood as a form of *explanatory* reasoning, as opposed to the reasoning from causes to

effects more characteristic of experimental science (Gould 1986; Diamond 1999).

An important element of historical reasoning is the need to differentially weight 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 (termed *synthetic* thinking by Baker (1996)). Such synthetic thinking is, as Cleland (2002) argues, a necessary commitment of historical science (as opposed to experimental science), due to the fundamental assymetry 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, the evidence must be carefully weighed to decide between competing hypotheses (the methodology sometimes known as "multiple working hypotheses").

In this paper, we take some first steps towards analysis of linguistic features of scientific writing in experimental and historical science, using several types of linguistically-motivated document features together with machine learning methods. Our goal is to show how linguistic features that are indicative of different classes of scientific articles may be usefully correlated with the rhetorical and methodological needs of historical and experimental sciences.

The Study

The Features

The features used in this study are 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 drawback, of course, is the possibility of ambiguity, in that the proper interpretation of such a keyword depends 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 significance of our results instead of biasing them.

We describe in this section the features we developed which are based on the options within three main systems, following Matthiessen's (1995) grammar of English, one of the standard SFL references. Keyword lists were constructed 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[1] (manually filtered for relevance to the given feature). The keyword lists were constructed entirely independently of the target corpus.

We use systems and subsystems within: CONJUNCTION, linking clauses together (either within or across sentences); MODALITY, giving judgements regarding probability, usuality, inclination, and the like; and COMMENT, expressing modal assessments of attitude or applicability. MODALITY and COMMENT relate directly to how propositions are assessed in evidential reasoning (e.g., for likelihood, typicality, consistency with predictions, etc.), while CONJUNCTION is a primary system by which texts are constructed out of smaller pieces[2].

CONJUNCTION

On the discourse level, the system of Conjunction 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 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 Conjunction are *Elaboration*, *Extension*, and *Enhancement*. Each of these options (subsystems) has its own options which we also use as features. (Note that the system network can be deepened further (Matthiessen 1995, p. 521), but our keyword-based method allows only a relatively coarse analysis.) The systems with their various subsystems, along with examples of indicators we use, are:

- 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*

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

[2] Other textual/cohesive systems, such as PROJECTION, TAXIS, THEME, and INFORMATION cannot be easily addressed, if at all, using a keyword-based approach.

Causal/Conditional:
Causal: Relations of cause and effect
so, therefore, for this reason
Conditional: Logical conditional relations
then, in that case, otherwise

Note that the actual features by which we represent an article are the frequencies of each subsystem's indicator features, each measured relative to its siblings. So, for example, one feature is Elaboration/*Appositive*, 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 CONJUNCTION 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 (1995), following Halliday (1994), lists eight types of Comment, which we give here along with representative indicators for each such subsystem.

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

MODALITY

The features for interpersonal modal assessment that we consider here are based on Halliday's (1994) analysis of the Modality system, as formulated by Matthiessen (1995). In this scheme, modal assessment is realized by a simultaneous choice of options within four systems[3]:

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

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 as an Objective attribute of the clause or as Subjective to the writer?

Manifestation: Is the assessment Implicitly realized by an adjunct or finite verb, or Explicitly by a projective clause?

The cross-product of all of these systems and subsystems creates a large number of modality assessment types, each of which is realized through a particular set of indicators[4]. We consider as *simple* features, each option in each system above (for example, Modalization/*Probability* opposed to Modalization/*Usuality*) as well as *complex* features made up of pairwise combinations[5] of simple features (such as Modalization/*Probability*:Value/*Median*). 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).

The Corpus

The initial study reported here was performed using a corpus of articles drawn from four peer-reviewed journals in two fields: *Palaios* and *Quaternary Research* in paleontology, and *Journal of Physical Chemistry A* and *Journal of Physical Chemistry B* in physical chemistry. (These particular journals were chosen initially in part for ease of access.) *Palaios* is a general paleontological journal, covering all areas of the field, whereas *Quaternary Research* focuses on work dealing with the quaternary period (from roughly 1.6 million years ago to the present). The two physical chemistry journals are published in tandem but have separate editorial boards and cover different subfields of physical chemistry, specifically: studies on molecules (*J. Phys Chem A*) and studies of materials, surfaces, and interfaces (*J. Phys Chem B*).

The numbers of articles used from each journal and their average (preprocessed) lengths in words are:

Journal	# Art.	Avg. Words	Total Size
Palaios	116	4584	3.4 Mb
Quaternary Res.	106	3136	2.0 Mb
J. Phys. Chem. A	169	2734	3.2 Mb
J. Phys. Chem. B	69	3301	1.6 Mb

Experimental Results

We took the preprocessed articles in our corpus and converted each of them into a vector of feature values

[3] Note that we did not consider here the system of POLARITY, since it cannot be properly addressed without more sophisticated parsing.

[4] Unfortunately, space precludes including a list of the MODALITY indicator features here.

[5] For simplicity, we did not consider 3- or 4-way combinations here. We may address this in future work.

(relative frequencies of system options), as described above. Throughout, classification models were constructed using the SMO learning algorithm (Platt 1998) as implemented in the Weka system (Witten & Frank 1999), using a linear kernel, no feature normalization, and the default parameters. (Using other kernels did not appear to improve classification accuracy, so we used the option that enabled us to determine easily the relevant features for the classification.) SMO is a support vector machine (SVM) algorithm; SVMs have been applied successfully to many text categorization problems (Joachims 1998). By using a linear kernel, we can easily evaluate which features contribute most to classification, by examining their weights (also, in preliminary tests, more complex kernels did not noticeably affect accuracy).

We first tested the hypothesis that paleontology articles are distinct from physical chemistry articles (along the field-independent linguistic dimensions we defined). Table 1 presents average classification accuracy using 20-fold cross-validation. In all four cross-disciplinary cases, classification accuracy is 83% and above, while in the two intra-disciplinary cases, accuracy is noticeably lower; *Palaios* and *Quat. Res.* are minimally distinguished at 74%, while *J. Phys. Chem. A* and *J. Phys. Chem. B* are entirely undistinguishable. This supports our main hypothesis, while pointing towards a possible more nuanced analysis of the difference between the two paleontology journals.

We now consider if a consistent linguistic picture of the difference between the two classes of scientific articles (paleontology and physical chemistry) emerges from the patterns of feature weights in the learned models. To do this, we ran SMO on all the training data for each of the four pairs of a paleontology with a physical chemistry journal, and ranked the features according to their weight for one or the other journal in the weight vector. Table 2 shows graphically which features were most indicative for each journal in its two trials. We restrict consideration to cases where a feature was strong (i.e., among the 30 with the highest weights) for a single class across all journal pairs (note that there were a total of 101 features). Even with this strong restriction, several striking patterns emerge. Space limits us to discussing the most important.

First, in the textual system of CONJUNCTION, we see a clear opposition between Extension, indicating paleontology, and Enhancement, indicating physical chemistry. This implies that paleontological text has a higher density of discrete informational items, linked together by extensive conjunctions, whereas in physical chemistry, while there may be fewer information items, each is more likely to have its meaning deepened or qualified by related clauses. This may be indicative that paleontological articles are more likely to be primarily descriptive in nature, requiring a higher information density, while physical chemists focus their attention deeply on a single phenomenon at a time. At the same time, this linguistic opposition may also reflect differing principles of rhetorical organization: perhaps physical chemists prefer a single coherent 'story line' focused on enhancements of a small number of focal propositions, whereas paleontologists may prefer a multifocal 'landscape' of connected propositions. Future work may include also interviews and surveys of the two types of scientists, regarding these points.

Next, in the system of COMMENT, the one clear opposition that emerges is between preference for Validative comments by paleontologists and for Predictive comments by physical chemists. This linguistic opposition can be directly related to methodological differences between the historical and experimental sciences. The (historical) paleontologist has a rhetorical need to explicitly delineate the scope of validity of different assertions, as part of synthetic thinking (Baker 1996) about complex and ambiguous webs of past causation (Cleland 2002). This is not a primary concern, however, of the (experimental) physical chemist; his/her main focus is prediction: the predictive strength of a theory and the consistency of evidence with theoretical predictions.

Finally, we consider the (complicated) system of MODALITY. At the coarse level represented by the simple features (in Table 2), we see a primary opposition in Type. The preference of the (experimental) physical chemist for Modulation (assessing what 'ought' or 'is able' to happen) is consistent with a focus on prediction and manipulation of nature. The (historical) paleontologist's preference for Modalization (assessing 'likelihood' or 'usuality') is consistent with the outlook of a "neutral observer" who cannot directly manipulate or replicate outcomes. This is supported also by patterns within the complex features crossing modality **Type** and **Manifestation** (not shown here due to space constraints). In Manifestation, we might say that Implicit variants are more likely for options that are well-integrated into the expected rhetorical structure, while Explicit realizations are more likely to draw attention to less characteristic types of modal assessment. We find that Modalization is preferably Implicit for paleontology but Explicit for physical chemistry; just the reverse holds true for Modulation. In this way, Modalization is integrated smoothly into the overall environment of paleontological rhetoric, and similarly Modulation is a part of the rhetorical environment of physical chemistry.

Examples

We now consider two short illustrative passages from articles in our corpus. These have been marked up (by hand) for the three main oppositions we identified above—we have marked realizations of each of the six features: **EXTENSION**, ***enhancement***, **validative comment**, **predictive comment**, **modalization**, and **modulation**.

Paleontology:

> Biologists agree that global warming is **likely** to produce changes in the diversity and distribution of species, **BUT** the magnitude, timing and nature of such responses remains

unclear. Animals **may** be affected directly by altered temperature and/or moisture regimes, for example, **OR** indirectly through associated vegetation changes. For herbivores in particular, direct effects are **likely** to be compounded by vegetation changes, especially in the case of animals with specialized habitat affinities or relatively small home ranges. Climatic change **may ALSO** occur too rapidly for animals to adapt, **OR** they **may** be unable to adapt because of physiological or phylogenetic constraints. **Under such circumstances**, species **may** become locally extinct. Estimating the potential range of adaptive response to climatic and vegetative shifts is **clearly** crucial to an understanding of the effects of global warming on terrestrial ecosystems, **YET it requires** a more thorough understanding of life history and ecosystem function than is often available.

(Smith & Betancourt 2003)

Physical chemistry:

In this experiment, the oxidation scan was run ***first and then followed*** by the reduction scan in the reverse direction. Above E=0.4V, the Cu(II) spin density reaches an average plateau value ***indicating that*** no spin coupling occurs between the neighboring copper centers. **HOWEVER**, in addition to this plateau, obtained in the oxidation scan, three local maxima in spin density are clearly observed at 0.48, 0.78, and 1.1 V, which very well correlate with the redox waves observed in the CV of poly[1,3,cu+]. Interestingly, in the reduction scan, the polymer film gives rise to only two spin density maxima at potential values close to those of the redox waves ***corresponding to*** the reduction of copper **AND** to the first redox wave of the polymer reduction. **Surprisingly**, the matrix spin density is very low **AND** reaches only ca. 5% of the copper(II) spin density, ***whereas*** both oxidizable components of poly[1,3,Cu+] show comparable electroactivity. In the simplest interpretation, such behavior **can be** regarded as a clear manifestation of the recombination of initially formed radical cations to spinless dications**. This is not unexpected**, ***since*** bipolarons are the dominant charge-storage configurations in essentially all thienylene-based conducting polymers. The onset of the spin appearance **can be** correlated with the onset of the first oxidation wave of the polymer oxidation. **HOWEVER**, contrary to the case of copper spin response, the spin response of the polymer is smooth and monotonic **AND** does not follow the current peaks recorded in the CV experiment. This underlines the efficiency of the polaron recombination process.

(Divisia-Blohorn et al. 2003)

Briefly, in the first passage above, from *Quaternary Research*, we see in a short space how frequent use of extension allows the construction of a complex of interrelated propositions, with no one focal point (though all are related to the basic theme of "global warming" and "climatic change"). We also see a clear preference for modalization, involving multiple levels of probabilistic assessment (e.g., "may", "likely", "clearly"), placing most propositions explicitly on a scale of variable likelihood. The use of a validative comment ("under such circumstances") also serves to circumscribe the validity of the assertions in the passage.

In the second passage, from *J. Phys. Chem. B*, on the other hand, we see the use of enhancement (primarily temporal and causal) in creating a narrative story-line which serves to organize presentation both of the experimental procedure but also of the interpretation of results. Extension is used mostly to construct small local structures which fit as a whole into the larger narrative line. Predictive comments are used ("surprisingly", "not unexpected") to emphasize certain results and also to place them into the larger context. Note also that the ambiguous modal assessment "can be" is used here to realize modulation (i.e., "it can be regarded…" = "we are able to regard it as…").

Conclusions

In this paper, we have shown how machine learning techniques together with linguistically-motivated features can be used to provide empirical evidence for rhetorical differences between writing in different scientific fields. Further, by analyzing the models output by the learning procedure, we can see what features realize the differences in register that are correlated with different fields. This method thus provides indirect empirical evidence for methodological variation between the sciences, insofar as rhetorical preferences can be identified which can be linked with particular modes of reasoning. This study thus lends support to those philosophers of science who argue against a monolithic "scientific method".

The current study is only a beginning, however. To make more general and stable conclusions, a much larger corpus of articles, from a wider variety of journals, will be needed. We are currently working on collecting and processing such a corpus. More fundamentally, there are serious limitations to using keyword/phrase counts as indicators for systemic options. Overcoming this limitation will require the construction of an accurate shallow systemic parser, which can enable a more general and more precise way to analyze the systemic functional options realized in a text. The rhetorical parsing methods developed by Marcu (2000) are an important step in this direction. Also, automatic methods for discovering rhetorically important features, similar to the subjectivity collocations of Wiebe et al. (2001) may be helpful.

It should also be noted that the current study treats each article as an indivisible whole. However, as noted by Lewin et al. (2001) in their analysis of social science texts, the rhetorical organization of an article varies in different sections of the text—future work will include studying how systemic preferences vary also across different sections of individual texts, by incorporating techniques such as those developed by Teufel and Moens (1998).

References

Argamon, S., M. Koppel, J. Fine and A. R. Shimoni (2003a). Gender, Genre, and Writing Style in Formal Written Texts. *Text*, **23**(3).

Argamon, S., M. Šari , S. S. Stein (2003b). Style mining of electronic messages for multiple authorship discrimination: First Results. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining 2003*.

Baayen, H., H. van Halteren, F. Tweedie (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, *Literary and Linguistic Computing*, **11**.

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

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

Diamond, J. (1999). *Guns, Germs, & Steel.* (New York: W. W. Norton and Company).

Divisia-Blohorn, B., F. Genoud, C. Borel, G. Bidan, J-M. Kern, and J-P. Sauvage (2003). Conjugated Polymetallorotaxanes: In-Situ ESR and Conductivity Investigations of Metal-Backbone Interactions, *J. Phys. Chem. B*, **107**, pp. 5126-5132.

Dodick, J. T., & N. Orion. (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 R.J. Sternberg, & J. Davidson (Eds.). *Mechanisms of Insight*. (Cambridge MA: MIT Press). pp. 365-395.

Eggins, S. & J. R. Martin, (1997). Genres and registers of discourse. In T. A. van Dijk, *Discourse as structure and process. A multidisciplinary introduction*. Discourse studies 1 (London: Sage), pp. 230–256.

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

Gould, S. J. (1986). Evolution and the Triumph of Homology, or, Why History Matters, *American Scientist* (Jan.-Feb. 1986): 60-69.

Gregory M., (1967). Aspects of varieties differentiation, *Journal of Linguistics* **3**, pp. 177-198.

Halliday, M.A.K. (1991). Corpus linguistics and probabilistic grammar. In Karin Aijmer & Bengt Altenberg (ed.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. (London: Longman), pp. 30-44.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. (London: Edward Arnold).

Hasan, R. (1988). Language in the process of socialisation: Home and school. In J. Oldenburg, Th. v Leeuwen, & L. Gerot (ed.), *Language and socialisation: Home and school* (Proceedings from the Working Conference on Language in Education, 17-21 November, 1986). North Ryde, N.S.W.: Macquarie University.

Holmes, D. I. and Forsyth, R. S. (1995). The federalist revisited: New directions in authorship attribution. . *Literary and Linguistic Computing*, 10(2):111-126

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*, pp. 137-142.

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

Latour, B. & S. Woolgar, (1986). *Laboratory Life: The Construction of Scientific Facts* (Princeton: Princeton University Press).

Lewin, B.A., J. Fine, & L. Young (2001). *Expository Discourse: A Genre-Based Approach to Social Science Research Texts* (Continuum).

Robert M. Losee (1996), Text Windows and Phrases Differing by Discipline, Location in Document, and Syntactic Structure. *Information Processing & Management,* **32**(6), 747-767.

Daniel Marcu (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, **26**(3), pages 395-448.

Martin, J. R. (1992). *English Text: System and Structure*. (Amsterdam: Benjamins).

Matthews, R. A. J. and Merriam, T. V. N. (1997). Distinguishing literary styles using neural networks. In Fiesler, E. and Beale, R., editors, *Handbook of Neural Computation*, chapter 8. (Oxford University Press).

Matthiessen, C. (1995). *Lexicogrammatical Cartography: English Systems*. (Tokyo, Taipei & Dallas: International Language Sciences Publishers).

Mayr, E. (1976). *Evolution and the Diversity of Life.* (Cambridge: Harvard University Press).

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist Papers,* Reading, Mass. : Addison Wesley.

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

Platt, J. (1998), *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,* Microsoft Research Technical Report MSR-TR-98-14.

Plum, G. A. & A. Cowling. (1987). Social constraints on grammatical variables: Tense choice in English. In Ross Steele & Terry Threadgold (ed.), *Language topics. Essays in honour of Michael Halliday*. (Amsterdam: Benjamins).

Sebastiani, F. (2002). Machine learning in automated text categorization, *ACM Computing Surveys*, **34**(1):1-47.

Smith, F. A. and J. L. Betancourt (2003). The effect of Holocene temperature fluctuations on the evolution and ecology of Neotoma (woodrats) in Idaho and northwestern Utah, *Quaternary Research* **59**, pp. 160 -171.

Stamatatos, E., N. Fakotakis & G. Kokkinakis, (2001). Computer-based authorship attribution without lexical measures, *Computers and the Humanities* **35**

Teufel, S., and Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. In *Proc. AAAI Spring Symposium on Intelligent Text Summarization*.

Wiebe, J., T. Wilson and M. Bell. (2001). Identifying Collocations for Recognizing Opinions. In *Proc. ACL/EACL '01 Workshop on Collocation, Toulouse, France, July 200.*

Witten, I.H. and Frank E. (1999). *Weka 3: Machine Learning Software in Java;* URL: http://www.cs.waikato.ac.nz/~ml/weka.

Yule, G.U. (1938). On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship, *Biometrika*, **30**, 363-390.

Table 1. Accuracy for linear SMO learning (with feature normalization) for pairs of journals, using 20-fold cross-validation.

	Historical		Experimental	
	Palaios	***Quat Res***	***Ph Ch A***	***Ph Ch B***
Palaios	--	74%	**91%**	**91%**
Quat Res	74%	--	**83%**	**86%**
Ph Ch A	**91%**	**83%**	--	68%
Ph Ch B	**91%**	**86%**	68%	--

Table 2. Significant simple features for each class in classification tests pairing each historical science journal in the study with each experimental journal. Features were sorted according to their weights learned in each two-class classification test (e.g., *Palaios* vs. *PCA*). Black squares represent features whose weights are in the top 15 for the main class of the column, and grey squares those with weights in the second 15.

Systemic Features				Historical				Experimental			
				Palaios		***QR***		***PCA***		***PCB***	
				PCA	*PCB*	*PCA*	*PCB*	*Pal*	*QR*	*Pal*	*QR*
CONJUNCTION	*Elaboration*										
		Appositive									
		Clarifying									
	Extension										
		Additive									
		Adversative									
		Verifying									
	Enhancement										
		Manner									
		Matter									
		S/T									
			Simple								
			Complex								
		C/C									
			Cause								
			Cond.								
COMMENT	*Admissive*										
	Assertive										
	Presumptive										
	Desiderative										
	Tentative										
	Validative										
	Evaluative										
	Predictive										
MODALITY	**Type**	*Modalization*									
			Probability								
			Usuality								
		Modulation									
			Obligation								
			Readiness								
	Value	*Median*									
		High									
		Low									
	Orientation	*Objective*									
		Subjective									
	Manifestation	*Implicit*									
		Explicit									