Methods for Genre Analysis Applied to Formal Scientific Writing

Paul CHASE
chaspau@iit.edu

Shlomo ARGAMON
argamon@iit.edu

Linguistic Cognition Laboratory Dept. of Computer Science Illinois Institute of Technology 10 W 31st Street Chicago, IL 60616, USA

1 Overview

Genre and its relation to textual style has long been studied, but only recently has it been a candidate for computational analysis. In this paper, we apply computational stylistics techniques to the study of genre, which allows us to analyze large amounts of text efficiently. Such techniques enable us to compare rhetorical styles between different genres; in particular, we are studying the communication of scientists through their publications in peer-reviewed journals. Our work examines possible genre/stylistic distinctions between articles in different fields of science, and seeks to relate them to methodological differences between the fields.

We follow Cleland's (2002) work in this area and divide the sciences broadly into Experimental and Historical sciences. According to this and other work in the philosophy of science, Experimental science attempts to formulate general predictive laws, and so relies on repeatable series of controlled experiments that test specific hypotheses (Diamond 2002), whereas Historical science deals more with contingent phenomena (Mayr 1976), studying unique events in the past in an attempt to find unifying explanations for their effects. We consider the four fundamental dimensions outlined by Diamond (2002, pp. 420-424):

1. Is the goal of the research to find general laws or statements or ultimate (and contingent) causes?
2. Is evidence gathered by manipulation or by observation?
3. Is research quality measured by accurate prediction or effective explanation?
4. Are the objects of study uniform entities (which are interchangeable) or are they complex entities (which are ultimately unique)?

The present experiment was designed to see if language features support these philosophical points. These linguistic features should be topic independent and representative of the underlying methodology; we are seeking textual clues to the actual techniques used by the writers of these scientific papers. This paper is partially based on our previously presented results (Argamon, Chase & Dodick, 2005).

2 Methodology

2.1 The Corpus

Our corpus for this study is a collection of recent (2003) articles drawn from twelve peer-reviewed journals in six fields, as given in Table 1. The journals were selected based both on their prominence in their respective fields as well as our ability to access them electronically, with two journals chosen per field and three fields chosen from each of Historical and Experimental sciences. Each article was prepared by automatically removing images, equations, titles, headings, captions, and references, converting each into a simple text file for further processing.

2.2 Systemic Functional Linguistics

We base our analysis on the theory of Systemic Functional Linguistics (SFL; Halliday 1994), which construes language as a set of interlocking choices or systems for expressing meanings, with general choices constraining the possible more specific choices. SFL presents a large number of systems, each representing a certain type of functional meaning for a potential utterance. Each system has conditions constraining its use and several options; once within a system we can choose but one option. Specific utterances are constrained by all the systemic options they realize. This approach to language allows the following types of questions to be asked: In places where a meaning of general type A is to be expressed in a text, what sorts of more specific meanings are more likely to be expressed in different contexts?

We focused on several systems for this study, chosen to correspond with the posited differences between the types of science we study: Expansion, Modality, and Comment (Matthiessen 1995). Expansion describes features linking clauses causally or logically, tying in to dimensions 1 and 4 above. Its three types are: Extension, linking different pieces of information; Elaboration, deepening a given meaning via clarification or exemplification; and Enhancement, qualifying previous information by spatial, temporal, or other circumstance. The second system, Modality, relates to how the likelihood, typicality, or necessity of an event is indicated, usually by a modal auxiliary verb or an adjunct adverbial group; as such it may serve to indicated differences on dimensions 2, 3, and 4. There are two main types of modality: Modalization, which quantifies levels of likelihood or frequency, and Modulation, which qualifies ability, possibility, obligation, or necessity of an action or event. Finally, the system of Comment is one of assessment, comprising a variety of types of ``comment'' on a message, assessing the writer's attitude towards it, its validity or its evidential status; this provides particular information related to dimensions 1 and 3.

In our analysis, it will be most helpful to look at oppositions, in which an option in a particular system is strongly indicative of one article class (either Experimental or Historical science) while a different option of that same 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 (its preferred option) of expressing the same general meaning.

2.3 Computational analysis

Because hand analysis is impractical on large document sets the first analyses were done via computer. We built a collection of keywords and phrases indicating each option in the aforementioned systems. Each document is first represented by a numerical vector corresponding to the relative frequencies of each option within each system. From here, machine learning was applied in the form of the SMO (Platt 1998) algorithm as implemented on the Weka machine learning toolkit (Witten & Frank 1999), using 10-fold cross-validation in order to evaluate classification effectiveness. This method was chosen in part because it generates weights for each feature; a feature has high weight (either positive or negative) if it is strongly indicative for one or the other class.

2.4 Human annotation

To measure the validity of our computational analysis, we are also performing hand tagging of systemic features on a subset of the corpus articles. Two articles from each journal have been chosen, each to be tagged by two trained raters. Part of the tagging process is to highlight key words or phrases indicating each option; we will compare these statistics to our previously generated feature lists in order to test and refine them. The tagging is currently under way; we will present results at the conference.

4 Results

To determine the distinctiveness of Historical and Experimental scientific writing, the machine learning techniques described above were applied to pairs of journals, giving for each pair a classification accuracy indicating how distinguishable one journal was from the other. These results are shown in Figure 1, divided into four subsets: Same, where both journals are from the same science; Hist and Exper with pairs of journals from different sciences, but the same type; and Diff indicates pairings of Historical journals with Experimental ones. The thick black line indicates the mean for each set, and the outlined box represents the standard deviation. As we see, journal pairs become more distinguishable as their methodological differences increase. Interestingly, Historical journals appear more stylistically homogenous than the Experimental journals, which is a subject for further study.

This shows that SFL is capable of discriminating between the different genres presented. We also examined the most important features across the 36 trials between different journals. The most consistently indicative-those features that are ranked highest for a class in at least 25 trials-are presented in Table 2. The table is arranged as a series of oppositions: the features on each row are in the same system, one side indicating Historical, the other Experimental.

In the system of Expansion, we see an opposition of Extension and Enhancement for Historical and Experimental sciences, respectively. This implies more independent information units in Historical science, and more focused storylines within Experimental science. Furthermore, there are oppositions inside both systems, indicating a preference for contrasting information (Adversative) and contextualization (Matter) in Historical science and for supplementary Information (Additive) and time-space (Spatiotemporal) relations in Experimental science.

The system of Comment also supports the posited differences in the sciences. The Experimental sciences' preference for Predictive comments follows directly from their focus on predictive accuracy. On the Historical side, Admissive comments indicate opinions (as opposed to factual claims), similarly Validative comments show a concern with qualifying the validity of assertions, comprising more of strong evidence than rigid proofs.

Finally in Modality we see interesting contrasted features. On the top level we have near-perfect opposition between Modalization and Modulation in general; Historical sciences speak of what is 'normal' or 'likely', while Experimental sciences assess what 'must' or 'is able' to happen.

5 Conclusion

This work is the first step in developing new automated tools for genre analysis, which promises the possibility of automatically analyzing large corpora efficiently or stylistic aspects while giving human interpretable results. The specific research presented has implications for the understanding of the relationship between scientific methodology and its linguistic realizations, and may also have some impact on science education. Future work (beyond the hand annotation and analysis already in progress) includes looking into stylistic variation within different article sections, as well as other analysis techniques (such as principle components analysis).

Journal	#Art	Avg. Words
J. Geology	93	4891
J. Metamorphic Geol.	108	5024
Biol. J. Linnean Society	191	4895
Human Evolution	169	4223
Palaeontologia Electronica	111	4132
Quaternary Research	113	2939
Physics Letters A	132	2339
Physical Review Letters	114	2545
J. Physical Chemistry A	121	4865
J. Physical Chemistry B	71	5269
Heterocycles	231	3580
Tetrahedron	151	5057

Table 1: Journals used in the study; the top represents historical fields with experimental sciences below.

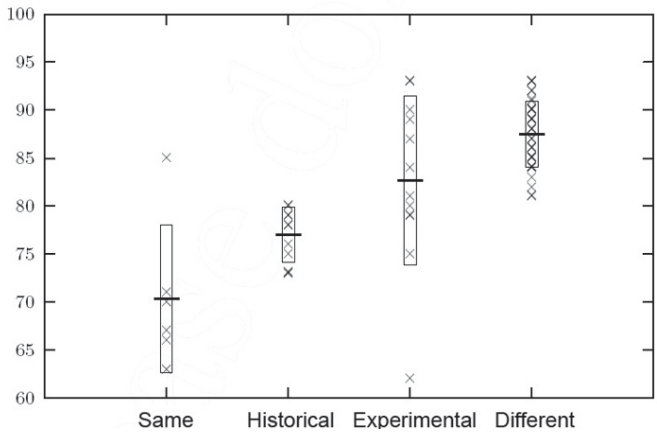

Figure 1: Learning accuracy for distinguishing articles in different pairs of journals. 'Same' are pairs where both journals are in the same field, 'Historical' and 'Experimental' represent pairs of journals in different Historical and Experimental fields, and 'Different' pairs of journals where one journal is experimental and the other historical. Means and standard deviation ranges are shown.

System	Historical	Experimental
Expansion	Extension(26)	Enhancement(31)
Elaboration		Apposition(28)
Extension	Adversative(30)	Additive(26)
Enhancement	Matter(29)	Spatiotemporal(26)
Comment	Admissive(30) Validative(32)	Predictive(36)
Modality Type	Modalization(36)	Modulation(35)
Modulation	Obligation(29)	Readiness(26)
Modality Value		High(27)
Modility Orientation	Objective(31)	Subjective(31)

Table 2. Consistent indicator features within each of the systems used in the study. Numbers in parentheses show in how many paired-classification tests the feature names was an indicator for the given class of documents.

References

Argamon, S., Chase, P., and Dodick, J.T. (2005). The *Languages of Science: A Corpus-Based Study of*

Experimental and Historical Science Articles. In Proc. 27th Annual Cognitive Science Society Meetings.

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. (2002). *Guns, Germs, & Steel*. (New York: W. W. Norton and Company).

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).

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

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

Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. In ECML-98,10th European Conference on Machine Learning, pp. 137-142.

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

Mitchell, T. (1997) *Machine Learning*. (McGraw Hill).

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

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

Combining Cognitive Stylistics and Computational Stylistics

Louisa CONNORS

School of Humanities and Social Science
The University of Newcastle, Australia

Studies in the computational analysis of texts have been successful in distinguishing between authors and in linking anonymously published texts with their authors but computational tools have yet to be accepted as mainstream techniques in literary analysis. Criticisms are generally centred around the belief that computational analyses make false claims about scientific objectivity and are in fact no less subjective than any other critical approach. This is perhaps because computational projects are in conflict, at a fundamental level, with contemporary post-structuralist notions of subjectivity, meaning and the arbitrary nature of language. This paper will argue that these objections rest on assumptions about language that need to be examined in light of developments in linguistics and cognitive psychology and that cognitive linguistics has the potential to bring a more interpretive framework to computational stylistics, a practice that has traditionally been applied in fairly narrow, empirical way.

Whilst computational analysis points to the possibility of subjectivity that is more coherent than some theoretical approaches imply, it does not necessarily diminish the role of culture and context in the formation of texts and subjectivity as highlighted by materialist readings. The application of cognitive linguistics in a computational study provides a model of syntax and semantics which is not independent of context but deeply bound up in context. Cognitive linguistics can explain the existence of computational results in a way that Saussurean based theories can not. It can offer a rich interpretive model that does not neglect the importance of author, reader, or context through its approach to language and literature as an expression of an innately constrained and embodied human mind.

Computational stylistics, particularly in studies of attribution, generally makes use of function words in