WILEY
InterScience®
DISCOVER SOMETHING GREAT

Stylistic Text Classification Using Functional Lexical Features

Shlomo Argamon
Linguistic Cognition Laboratory, Department of Computer Science, Illinois Institute of Technology, 10 W. 31st Street, Chicago, IL 60616. E-mail: argamon@iit.edu

Casey Whitelaw
School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia. E-mail: casey.whitelaw@gmail.com

Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan
Linguistic Cognition Laboratory, Department of Computer Science, Illinois Institute of Technology, 10 W. 31st Street, Chicago, IL 60616. E-mail: {chaspau, hotasob, gargnav, levishl}@iit.edu

Most text analysis and retrieval work to date has focused on the topic of a text; that is, *what* it is about. However, a text also contains much useful information in its style, or *how* it is written. This includes information about its author, its purpose, feelings it is meant to evoke, and more. This article develops a new type of lexical feature for use in stylistic text classification, based on taxonomies of various semantic *functions* of certain choice words or phrases. We demonstrate the usefulness of such features for the stylistic text classification tasks of determining author identity and nationality, the gender of literary characters, a text's sentiment (positive/negative evaluation), and the rhetorical character of scientific journal articles. We further show how the use of functional features aids in gaining insight about stylistic differences among different kinds of texts.

Introduction

A common goal in automated text analysis is to gain an understanding or summary of the topic(s) covered in the text. This may involve information extraction into frame-based semantic representations (Appelt, Hobbs, Bear, Israel, & Tyson, 1993; Cowie & Lehnert, 1996; McCallum, Freitag, & Pereira, 2000; Roth & Yih, 2001), text clustering and categorization (Kehagias, Petridis, Kaburlasos, & Fragkou, 2003; Lewis, Schapire, Callan, & Papka, 1996; Sebastiani, 2002), or retrieval of topic-relevant documents by one of a variety of heuristics (Grossman & Frieder, 1998; Ponte & Croft, 1998; Salton & McGill, 1983). However, it is now becoming clear that such "objective" representations of topical information in a text are not enough on their own to properly support users in interactive retrieval tasks (Belkin, 1993; Chen, Magoulas, & Dimakopoulos, 2005). Dealing with issues of "information quality" (Tang, Ng, Strzalkowski, & Kantor, 2003) and "authority" (Fritch & Cromwell, 2001) already have been identified as important for more effective user support; in this article, we examine another dimension: how to extract useful *stylistic* information from texts.

We view the full meaning of a text as much more than just the topic it describes or represents. Textual meaning, broadly construed, also can include aspects such as *affect* (What feeling is conveyed by the text?), *genre* (In what community of discourse does the text function?), *register* (What is the function of the text as a whole?), and *personality* (What sort of person, or who specifically, wrote the text?). These aspects of meaning are captured by the text's *style* of writing, which may be roughly defined as how the author chose to express a topic, from among a very large space of possible ways of doing so. We contrast, therefore, the *how* of a text (style) from the *what* (topic).

Immediate applications of stylistic text analysis include authorship attribution and profiling (Argamon, Koppel, Fine, & Shimony, 2003; Baayen, Halteren, & Tweedie, 1996; Burrows, 1987; de Vel, Corney, Anderson, & Mohay, 2002; Kjell & Frieder, 1992; McEnery & Oakes, 2000; Mosteller &Wallace, 1964; Stamatatos, Fakotakis, & Kokkinakis, 2000; Torvik, Weeber, Swanson, & Smalheiser, 2005), genre-based text classification and retrieval (Finn, Kushmerick, & Smyth, 2002; Karlgren, 2000; Kessler, Nunberg, & Schütze, 1997), sentiment analysis (Pang, Lee, & Vaithyanathan, 2002; Turney, 2002), and spam/scam filtering (Androutsopoulos, Koutsias, Chandrinos, Paliouras, &

Received August 2, 2005; revised June 18, 2006; accepted June 18, 2006

• Published online 26 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20553

Spyropoulos, 2000; Kushmerick, 1999; Patrick, 2004). Other applications include criminal and national security forensics (Chaski, 1999; McMenamin, 2002), mining of customer feedback (Berry & Linoff, 1997; McKinney, Yoon, & Zahedi, 2002), and aiding humanities scholarship (Argamon & Olsen, 2006; Holmes, 1998; Hoover, 2002; Matthews & Merriam, 1997). As the sheer quantity of texts available on every given topic grows exponentially, the need for effective automated extraction of more dimensions of meaning is becoming acute.

A key problem for stylistic text categorization, which we address here, is proper choice of textual features for modeling style. While topic-based text categorization can get quite far by using models based on "bags of content words," style is somewhat more elusive. We start from the intuitive notion that style is indicated by features representing the author's choice of one mode of expression from a set of equivalent modes for a given content. At the surface level, this may be expressed by a wide variety of possible features of a text: choice of particular words, syntactic structures, discourse strategy, or all of the above and more. The underlying causes of such variation are similarly heterogeneous, including the genre, register, or purpose of the text as well as the educational background, social status, and personality of the author and audience. What all these dimensions of variation have in common, though, is an independence from the "topic" or "content" of the text, which may be considered to be those objects and events to which it refers (as well as their properties and relations as described in the text). We may thus define the *stylistic meaning* of a text to be those aspects of its meaning that are *nondenotational*; that is, independent of the objects and events to which the text refers.

Most computational stylistics work to date has been based on hand-selected sets of content-independent features such as function words (Matthews & Merriam, 1997; Mosteller & Wallace, 1964; Tweedie, Singh, & Holmes, 1996), parts-of-speech and syntactic structures (Stamatatos et al., 2000), and clause/sentence complexity measures (de Vel, 2000; Yule, 1944; also see the survey in Karlgren, 2000). While new developments in machine learning and computational linguistics have enabled larger numbers of features to be generated for stylistic analysis, in almost no case is there strong linguistic motivation behind input feature sets that would relate features directly to stylistic concerns. Rather, the general methodology that has developed is to find as large a set of topic-independent textual features as possible and use them as input to a generic learning algorithm (preferably one resistant to overfitting, and possibly including some feature selection). Some interesting and effective feature sets have been found in this way (e.g., Karlgren, 2000; Koppel, Akiva, & Dagan, 2003); function words also have proven to be surprisingly effective on their own (Argamon et al., 2003; Argamon & Levitan, 2005; McEnery & Oakes, 2000). Nevertheless, we contend that without a firm basis in a linguistic theory of meaning, we are unlikely to gain any true insight into the nature of the stylistic dimension(s) under study. Proper choice of features also should, of course, aid classification accuracy.

Our goal, therefore, is to find a computationally tractable formulation of linguistically well-motivated features which permit text classification based on variation in stylistic meaning. We describe here a framework and methodology for constructing a lexicon using attribute-value taxonomies based on principles of Systemic Functional Grammar (SFG, Halliday, 1994), which we find to be useful for this purpose. In particular, SFG explicitly recognizes and represents *nondenotational* meaning as part of the general grammar, which makes it particularly applicable to stylistic problems. Our experimental results show that features based on systemic function features often improve stylistic classification effectiveness, and also enable us to gain insight into the nature of the stylistic varieties under study.

The system described here does no complex syntactic parsing, relying instead on mainly lexical features of the text. While more sophisticated processing will likely lead to improved results in the future, we believe that simpler methods can be quite effective as well and should be fully explored first. The text analysis methodology described in this article comprises four steps:

1. Tokenize texts and assign part-of-speech tags;
2. Extract instances of lexical units from each text, as specified in the lexicon (with some, but not complete, disambiguation);
3. Compute relative frequencies of semantic attribute values for each text, giving an overall "feature vector" describing the text;
4. Use machine learning to construct discrimination models for stylistic text classification tasks.

Using this methodology, this article shows how functional lexical features can improve stylistic classification results for several stylistic classification tasks:

> *Authorship attribution:* Determining who (in a given list of candidates) wrote a specific chapter of a literary work;
>
> *Gender attribution:* Determining whether a speech in one of Shakespeare's plays belongs to a male or a female character;
>
> *Sentiment analysis:* Determining if a movie review is positive or negative, based only on the text; and
>
> *Scientific rhetoric:* Determining if two similar scientific fields (Geology and Paleontology) differ in their reasoning and argumentation structure, by analyzing the text of peer-reviewed journal articles.

Functional Lexical Attributes

We first give an overview of the taxonomies underlying the functional lexical features that we have developed; more detail on the taxonomies can be found in the Appendix. This work is based on the theory of SFG, a functional approach to linguistic analysis (Halliday, 1994). SFG models the grammar of a language by a network of choices of meanings that can be expressed (Matthiessen, 1995), and so all lexical and structural choices are represented as the realizations of particular semantic and contextual meanings.

The theory, rooted in the earlier work of Firth (1968), takes a primarily *sociological* view of language, and has developed largely in the context of its use by applied linguists for literary/genre analysis and for studying language learning (An excellent overview of SFG and its relation to other functional accounts of grammar may be found in Butler, 2003.) Describing SFG's main concerns, Matthiessen (1983) eloquently observed:

> There are few grammatical mechanisms that have been developed within a framework with as impressive a tradition as Systemic Linguistics and with as wide a scope. The systemic framework is not just a non-transformational alternative to Chomsky's transformational grammar. It is different from Chomskyan work at the level of framework, not only at the level of mechanism and notation. Systemic linguists ask questions like "How does communication succeed?", "What are the relations between context and language use?", "What can a speaker of English do grammatically to achieve a particular purpose?", "What are the options for expressing grammatically a particular range of meanings?", "What functions does language serve?" and so on. . . . One consequence of questions of this type has been in Systemic Linguistics that text as a communicative unit is taken to be the basic linguistic unit rather than the sentences that are used to express texts . . . Obviously, this view has far-reaching effects on the conception of grammar. (p. 155, footnote 3)

We believe that the fact that SFG models language as a network of mutually exclusive options for expressing meaning (which can be structural or lexical) makes it particularly useful for modeling stylistic variation among texts.

Systemic functional grammars have been applied to automatic natural language processing in several contexts since the 1960s, though after the influential work of Winograd (1972) on natural language understanding, computational applications have been mostly limited to text generation (Fawcett & Tucker, 1990; Matthiessen & Bateman, 1991; Teich, 1995) rather than text analysis due to the complexity of parsing in the theory (but also see O'Donnell, 1993).

Briefly, SFG construes language as a set of interlocking choices for expressing meanings, with more general choices constraining the possible specific choices. A simple example in the pronominal system of English:

> If a pronoun is to be used, it may refer either to one of the discourse participants, or to a third party;
>
> - If to one of the participants, it may refer to the speaker (*I*, *me*), the speaker-plus-others (*we*, *us*), or the hearer (*you*);
> - If to a third party, it may refer either to one individual or to many (*they*, *them*);
>   - –If to a single individual, it may refer to a conscious individual or to a nonconscious individual (*it*);
>   - *If to a single conscious individual, it may refer to a male (*he*, *him*) or to a female (*she*, *her*);
>
> and so forth.

Note that a choice at one level may open up further choices at other levels, choices that are not open otherwise. For example, English does not allow a pronoun to distinguish between pluralities of conscious or nonconscious individuals. Furthermore, any specific choice of lexical item or syntactic structure is determined by choices from multiple systems at once, as the choice between "I" and "me" is determined by the independent choice governing the pronoun's syntactic role as either a subject or an object.

Thus, a *system* defines a set of *options* for meanings to be expressed. Each (nonroot) system has an *entry condition*, a propositional formula of options from other systems, denoting when that system is possible. Each option gives constraints (lexical, morphological, or syntactic) on utterances that express the option. Options (or logical combinations thereof) may serve as entry conditions for more specific systems. While some systems, as in the example described earlier, are *disjunctive* such that exactly one of their options must be chosen, others are *conjunctive* in that all of their options must be chosen—this enables combinatorial possibilities. For example, modal verbs (e.g., "may," "might," or "must") choose options from multiple systems, including "Modality Type" (likelihood, frequency, obligation, etc.) and "Modality Value" (median, high, low).

Relevant words and phrases are stored in a lexicon, which stores, for each lexical entry, a value for each of a set of *semantic lexical attributes* from the options in associated *system networks*. Each such network has a unique root, and we allow entry conditions to be only single option or conjunctions of options.[1] More formally, each system network in this conception is a directed acyclic *and/or* graph, whose nodes are systems and whose directed arcs are options. An Option O_2 is a *child* of Option O_1 if O_1's destination node is O_2's source node; descendants and ancestors in the graph are defined in the straightforward manner. If Option O_1 is chosen and it leads into a disjunctive node, then exactly one of its children also must be chosen; if it leads into a conjunctive node, then all of its children also must be chosen. Note that if an option is chosen, all of its ancestors also are chosen.

By viewing language as a complex of choices between mutually exclusive options, the systemic approach can enable effective characterization of variation in language use. As described more formally later, we use as features the *relative frequencies* of various options for each system node, which directly encode aspects of functionally relevant textual variation.

The remainder of this section describes the main system networks that we use here for computational analysis of textual style. They are divided into three categories: *Cohesion*, referring to how a text is constructed to "hang together;" *Assessment*, meaning how a text construes propositions as statements of belief, obligation, or necessity, contextualizing them in the larger discourse; and *Appraisal*, or how the text adjudges the quality of various objects or events. Note that the system networks we use are the result of decades of

[1] See Matthiessen (1995) for a discussion of the full SFG grammar representation (allowing disjunction in entry conditions), which we simplify to improve computational tractability.

DOI: 10.1002/asi

research on textual analysis within the SFG community, and are not ad hoc inventions for our particular purposes.

Cohesion

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 (e.g., pronominal reference, deictic expressions, ellipsis, etc.) as well as lexical repetition and variation, and different ways of linking clauses together. 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 current computational work considers just types of conjunctions, for feasibility of automated extraction. Automated coreference resolution, for example, is a very difficult unsolved problem.

Words and phrases that conjoin clauses (e.g., "and," "while," and "in other words") are organized in SFG in the CONJUNCTION system network. Types of CONJUNCTION serve to link a clause with its textual context by denoting how the given clause expands on some aspect of its preceding context (Matthiessen, 1995, pp. 519–528). The three top-level options of CONJUNCTION are Elaboration, Extension, and Enhancement, defined as:

- Elaboration: Deepening the content in its context by exemplification or refocusing.
- Extension: Adding new related information, perhaps contrasting with the current information.
- Enhancement: Qualifying the context by circumstance or logical connection.

A more detailed description of the CONJUNCTION taxonomy is given in the Appendix.

Assessment

Generally speaking, *assessment* may be defined as contextual qualification of the epistemic or rhetorical status of events or propositions represented in a text. Examples include assessment of the likelihood of a proposition, the typicality of an event, the desirability of some fact, or its scope of validity. Two important systems in SFG that address assessment are MODALITY, enabling expression of typicality and necessity of some fact or event, and COMMENT, enabling assessment of the writer's stance with respect to an assertion in the text.

The system of MODALITY enables one to qualify events or entities in the text according to their likelihood, typicality, or necessity. Syntactically, MODALITY may be realized in a text through a modal verb (e.g., "can," "might," "should," "must"), an adverbial adjunct (e.g., "probably," "preferably"), or use of a projective clause (e.g., "I think that. . . ." "It is necessary that. . . ."). Each expression of MODALITY has a value for each of four attributes (see the discussion in the Appendix for more detail):

- Type: What kind of modality is being expressed?
  –Modalization: How "typical" is it? (*probably*, *seldom*)
  –Modulation: How "necessary" is it? (*ought to*, *allowable*)
- Value: What degree of the relevant modality scale is being averred?
  –Median: The "normal" amount. (*likely*, *usually*)
  –Outer: An extreme (either high or low) amount. (*maybe*, *always*)
- Orientation: Relation of the modality expressed to the speaker/writer.
  –Objective: Modality expressed irrespective of the speaker/writer. (*maybe*, *always*)
  –Subjective: Modality expressed relative to the speaker/writer. (*We think . . .* , *I require . . .*)
- Manifestation: How is the modal assessment related to the event being assessed?
  –Implicit: Modality realized "in-line" by an adjunct or modal auxiliary. (*preferably . . .* , *maybe . . .*)
  –Explicit: Modality realized by a projective verb, with the nested clause being assessed. (*It is preferable . . .* , *It is possible . . .*)

The system of COMMENT provides a resource for the writer to "comment" on the status of a message with respect to textual and interactive context in a discourse. Comments are usually realized as adjuncts in a clause and may appear initially, medially, or finally. We use the eight categories of COMMENT listed by Matthiessen (1995): *Admissive*, message is an admission (e.g., '*we concur . . .*'); *Assertive*, emphasis of reliability (e.g., '*Certainly . . .*'); *Desiderative*, desirability of the content (e.g., '*Unfortunately . . .*'); *Evaluative*, judgment of the actors involved (e.g., '*Sensibly . . .*'); *Predictive*, coherence with predictions (e.g., '*As expected . . .*'); *Presumptive*, dependence on other assumptions (e.g., '*I suppose . . .*'); *Tentative*, assessing the message as tentative (e.g., '*Tentatively . . .*'); and *Validative*, assessing scope of validity (e.g., '*In general . . .*').

Appraisal

Finally, *appraisal* denotes how language is used to adopt or express an attitude of some kind toward some target (Martin & White, 2005). For example, in "I found the movie quite monotonous," the speaker adopts a negative *Attitude* ("monotonous") toward "the movie" (the *appraised object*). Note that attitudes come in different types; for example, "monotonous" describes an inherent quality of the appraised object while "loathed" would describe an emotional reaction of the writer. The overall type and orientation of appraisal expressed in the text about an object gives a picture of how the writer wishes the reader to view it (modulo sarcasm, of course). To date, we have developed a lexicon for appraisal adjectives as well as relevant modifiers (e.g., "very" or "sort of"). The two main attributes of appraisal, as used in this work, are (a) Attitude, giving the kind of appraisal being expressed, and (b) Orientation, giving whether the appraisal is *positive* (e.g., good, beautiful, nice) or *negative* (e.g., bad, ugly, evil) (There also are other attributes of appraisal, discussed in the Appendix.) The three main types of Attitude are: *affect*, relating to the speaker's/writer's emotional state (e.g., "happy," "sad"; *appreciation*, expressing evaluation of supposed intrinsic qualities of an object (e.g., "tall," "complex"); and *judgment*, expressing social evaluation (e.g., "brave," "cowardly"). More detail on the appraisal taxonomy as used in this work is given in the Appendix.

System Design

The core of our style analysis system is ATMan,[2] in which input texts are linguistically processed, creating "annotated texts" *(atexts)*, stored in a relational database. An atext comprises several tables in the database, as follows. The main component of an atext is a table of position-labeled *tokens*, each corresponding to a word, number, or punctuation mark, and each labeled with a set of *attributes* such as the token's position (in the text, in its sentence), part-of-speech (singular noun, past-tense verb, etc.), capitalization (lowercase, capitalized, all-uppercase, etc.), length in characters, and so forth. Raw text is converted into an atext in the database by an *import method,* which uses a library of tokenizers and token-analysis methods. Different import methods are defined for different input file formats (Usenet articles, SGML-tagged corpus files, etc.). Metadata about the text (its title, source, author, etc.) also are stored in the atext.

An atext also has an associated set of linguistic *units* that have been extracted based on a semantic lexicon of words and phrases, each corresponding to a short sequence of tokens in the text and described by a set of semantic attribute-value pairs (as described later). A separate data element in ATMan is the lexicon itself, used for semantic unit extraction, described later.

For use in machine learning, ATMan outputs ARFF files (in Weka format; Witten & Frank, 2000) comprising a list of labeled numeric vectors, each corresponding to an atext. Each vector element represents the relative frequency of some feature-value in the atext, conditional on some other possible feature-value (discussed in the last section), with symbolic labels indicating possible classes for text classification learning.

The lexicon comprises a set of *lexical items* (words or phrases), each described by one or more *lexical entries*. Multiple entries for a lexical item correspond to multiple possible meanings. Each lexical entry consists of both *constraints* on when the entry is applicable to the lexical item, and a set of *attribute values* describing syntactic and semantic properties. Currently, the only type of constraint is expressed as a set of allowed part-of-speech sequences, such that the entry is only instantiable for the lexical item if it is tagged with an allowed part-of-speech sequence (Recall that lexical items may be multiword phrases.) Each attribute is the name of a system network (described later), and is assigned a set of values, each corresponding to an option in that system network. All of the values are assigned conjunctively to that attribute in that entry.

This general structure is unexceptional; it is the choice of attributes and the semantic organization of their possible values that give the approach its power. We note again that our goal is *not* syntactic parsing, and hence, we rely as much as possible on locally computable properties of the text, specifically part-of-speech tags, to constrain interpretation. Furthermore, as the final goal of this processing is to compute an overall statistical representation of the document (as a vector of numeric feature frequencies), some local ambiguity in interpretation can be tolerated.

Lexicons in each system network described earlier were constructed using a semi-automated technique to find relevant terms and assign them appropriate attribute values. In each case, we started with *seed terms* taken from example words and phrases given for various combinations of system options in standard SFG references: Halliday's (1994) introduction to SFG, Matthiessen's (1995) grammar of modern English, and Martin and White's (2005) appraisal theory. Candidate expansions for each seed term were generated from multiple resources—WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) and from two online thesauruses (`http://m-w.com` and `http://thesaurus.com`). In WordNet, the members of each synset (i.e., set of synonyms) were taken as the related set; similarly, synonym and related word lists were taken from each thesaurus. Candidates were accepted only with the same part of speech as a seed term.

A list was generated, for each main category, of all such candidate terms, and they were then ranked by frequency of occurrence in the candidate list (total number of seed term/resource pairs generating that candidate). This provided a coarse ranking of relevance, enabling more efficient manual filtering. Uncommon words, unrelated words, or words arising from an incorrect sense of a seed term will tend to be ranked lower in the candidate list than those related to more of the seed terms and are present in more of the resources. As well as increasing coverage, using multiple thesauruses allows for more confidence votes and, in practice, increases the utility of the ranking. Each ranked list was manually inspected to produce the final set of terms used. This procedure enabled terms with low confidence to be automatically discarded early, reducing the amount of manual work required.

Stylistic Text Classification

We validate our claims for the usefulness of functional lexical features by applying them to a variety of stylistic text classification problems. In each experiment Weka's (Witten & Frank, 2000) implementation of the SMO version (Platt, 1998) of the Support Vector Machine learning algorithm (Cristianini & Shawe-Taylor, 2000) with a linear kernel was used for learning classification models (higher order kernels did not seem to make much difference); for the multiclass problems, a one-versus-all strategy was used to generalize the binary learner for multiple output classes. Except where otherwise noted, 10-fold cross-validation was used throughout to

[2] **A**nnotated **T**ext **Man**ager.

estimate out-of-training classification accuracy, with partitions held constant for the different feature sets on each task. SMO's stiffness (C) parameter was tuned in each case by a further 10-fold cross-validation over the training set (nested within evaluation cross-validation runs). Statistical significance of differences in accuracy between feature sets on the same task was estimated by the paired t test (over the cross-validation partitions).

Feature sets

As input to machine learning, the documents in each corpus were processed into numeric feature vectors using various combinations of the following feature sets:

FW: Features are the relative frequencies of a set *FW* of 675 function words, with the value of such a feature $w \in FW$ in a document d defined as:

$$\frac{\text{count}(w, d)}{\sum_{w' \in FW} \text{count}(w', d)}$$

BoW: Similarly, in some cases, we used "bag-of-words" features, defined as the relative frequencies of each distinct token in the given corpus, with each such feature (for a given token t) defined as:

$$\frac{\text{count}(t, d)}{\text{len}(d)}$$

where len(d) gives the total number of tokens in d.

Functional lexical features, based on the systemic functional taxonomies described earlier, are computed as follows. Recall that each lexical entry is a frame comprising a set of attribute values, where each attribute is the name of a system network, and each value is an option (or conjunction of noncontradictory options) in the system network. Numeric features may then be computed based on the lexical items occuring in a given text, where each feature is the relative frequency of some option O_1 with respect to some other option O_2. Given an atext d, define $N_d(O_1)$ to be the number of units in d with value O_1, similarly $N_d(O_1, O_2)$ to be the number with both O_1 and O_2. Then the *relative frequency of* O_1 *with respect to* O_2 is defined as

$$RF_d(O_1 | O_2) = \frac{N_d(O_1, O_2)}{N_d(O_2)}$$

For example, the frequency of sibling options relative to their shared parent allows direct comparison of how different texts prefer to express the parent via its different options. Alternatively, the frequency of options relative to a system network root enables a more global comparison of what types of meanings (with a given system) are expressed in a document. Specifically, the features that we considered in the experiments described in this article are as follows (refer to the section on functional lexical attributes and the Appendix):

Con: Each feature is the relative frequency (RF_d) of a node in the Conjunction system with respect to its parent.

Mod: This feature set consists of the union of two related feature sets:

- For each node in each Modality system (Type, Value, Orientation, and Manifestation), the relative frequency (RF_d) of the node with respect to its parent;
- For each pair of nodes in different Modality systems (e.g., Type and Value), the relative frequency (RF_d) of terms labeled by both nodes with respect to the conjunction of their parents.

Com: This set consists of the relative frequency (RF_d) of each node in the Comment system with respect to its parent.

Att: This feature set comprises, for each node in the Attitude system, the relative frequency (RF_d) of the node with respect to its parent;

App: This feature set comprises Att as well as, for each node n in Attitude, both RF_d(Positive | n) and RF_d(Negative | n).

Combinations of these feature sets (amounting to concatenating the relevant feature vectors) also were considered (termed, e.g., Con + Mod, denoting the union of Con and Mod).

Feature Analysis

In many cases, as we shall see, examining the most important features for stylistic classification can give useful insights. We measure the classification importance of each feature i by $\omega_i\sigma_i$, where ω_i is its weight in the linear model constructed by SMO, and σ_i is the sample *SD* of the feature's values over the entire corpus (as any given ω will affect classification more for a feature whose value varies more between texts).

To make explicit the relationship that the functional features indicating each of two document classes give us, we take the top features indicating each class and find all their *oppositions*, where an opposition is a pair of relative frequencies features, one of which indicates one class and the other indicates the other class, where the features' conditioning events are identical and their conditioned events are sibling nodes in some systemic taxonomy. For example, if CONJUNCTION/Extension (i.e., RF_d(Extension | CONJUNCTION)) is indicative of Class A and CONJUNCTION/Enhancement of Class B, we would have the opposition:

Condition	Class A	Class B
CONJUNCTION	Extension	Enhancement

A more complex example is where Class A is indicated by high values of

$$RF_d(\text{Median} \mid \text{VALUE,MODALITY TYPE/Modalization})$$

and Class B by high values of

$$RF_d(\text{Low} \mid \text{VALUE,MODALITY TYPE/Modalization})$$

In this case, the conditioning event is the conjunction of two nodes, one of which is the shared parent of the conditioned

DOI: 10.1002/asi

events. This gives the opposition:

Condition	Class A	Class B
MODALITY TYPE/Modalization: VALUE	Median	Low

In this case, when a text in Class A expresses Modalization (typicality of an event or proposition), it prefers to express Median (i.e., nonextreme) values whereas in similar situations, Class B prefers to express Low values. This may indicate that texts in Class A tend to be more cautious, not expressing even unexceptional statements as absolute fact (saying "he likely went home" rather than "he went home") while texts in Class B might only explicitly express Modalization when it is particularly low (saying "he went home" in the last case, but "she might have wanted him to stay," if the conclusion is uncertain). Interpretation will depend, of course, on the particular types of texts under consideration.

The oppositions given by such analysis give direct information about linguistic differences between two document classes, in that the two classes have differing preferences about how to express the conditioning event. In the first example presented earlier, Class A prefers to conjoin items by Expansion, indicating a higher density of more-or-less independent information units, whereas Class B prefers conjoining items by Enhancements, indicating a more closely focused structure dealing with a smaller number of independent information units.

Experimental Results

To validate the methodology of using functional lexical features for stylistic classification, we ran experiments on a number of different stylistic classification tasks, showing that (a) functional lexical features can aid classification, and (b) in many cases, analyzing indicative features can give insight into underlying phenomena.

Authorship Identification

Authorship attribution, the problem of determining who wrote an anonymous text, is perhaps the most classic stylistic text classification task. Ever since the influential work of Mosteller and Wallace (1964) on the authorship of the Federalist Papers introduced them, function word frequencies have proven remarkably resilient for this task (Argamon & Levitan, 2005), even though many other potentially useful features have been suggested. The intuition behind the utility of function words for authorship attribution is as follows. Due to their high frequency in the language and highly grammaticalized roles, function words are very unlikely to be subject to conscious control by the author. At the same time, the frequencies of different function words vary greatly across different authors and genres of text—hence the expectation that modeling the interdependence of different function-word frequencies with style will result in effective attribution. However, the highly reductionistic nature of such features seems unsatisfying, as they rarely give good insight into underlying stylistic issues. We suggest here that some of the systemic functional features developed in this work may both aid in accurate authorship attribution as well as give some insight.

The corpus. The corpus for this evaluation (see Table 1) was constructed from 20 19th-century novels by eight different authors (the same as those used in Hoover's, 2002, authorship

TABLE 1. The authorship attribution corpus, comprising the chapters in a set of 20 19-century novels. The number of chapters in each book and the average number of words per chapter in each book are as shown.

Author	Nationality	Book	#Chapters	Average words
W. Cather	American	My Antonia	45	1,826
		Song of the Lark	60	2,581
		The Professor's House	28	2,172
H. James	American	The Europeans	12	5,003
		The Ambassadors	36	4,584
S. Lewis	American	Babbit	34	3,693
		Main Street	34	4,994
		Our Mr. Wrenn	19	4,126
J. London	American	The Call of The Wild	7	4,589
		The Sea Wolf	39	2,739
		White Fang	25	2,917
J. Conrad	British	Lord Jim	45	2,913
		The Nigger of the Narcissus	5	10,592
T. Hardy	British	Jude the Obscure	53	2,765
		The Mayor of Casterbridge	45	2,615
		Tess of the d'Urbervilles	58	2,605
R. Kipling	British	The Jungle Book	13	3,980
		Kim	15	7,167
H. G. Wells	British	The Invisible Man	28	1,756
		The War Of The Worlds	27	2,241
Total			628	

DOI: 10.1002/asi

study). Each novel was divided into individual chapters, each of which was considered as a separate example for learning and classification.

Features. The feature sets used were FW, Con, Mod, and Com, as well as various combinations of these sets. Neither of the appraisal feature sets (App and Att) proved useful for classification nor did Con, Mod, or Com alone, so those results are not shown.

Results. We ran 10-fold cross-validation tests using SMO (as described earlier) for classification of chapters for book, author, and author nationality (American or British). For all feature sets, the best *C* parameter found (using nested cross-validation) was 1. Classification results are given in Figures 1, 2, and 3. Systemic features performed above baseline in all cases, though not as well as FW. In all cases, however, addition of systemic features improved classification accuracy (Author: FW + Con p = .00003, Book: FW + Con + Mod + Com p = .0018, Nationality: FW + Com + Mod p = .0016). However, the differences between the various combined feature sets were not statistically significant in any case.

We may tentatively examine differences between the tasks based on which functional features seemed to help the most. In authorship attribution, Con appears to be dominant, indicating variability in how different authors connect information units. Book attribution was aided by Con + Mod + Com, indicating a wide variety of differences in dialogue structure and modal assessment between books. Finally, nationality attribution was most aided by Com + Mod, suggesting differences between U.S. and British authors in how events and propositions are assessed.

To better understand which features contributed most to aid classification for each task, we considered the product of the *SD* and absolute weights assigned to them in the various linear models constructed by SMO for the highest accuracy feature set in each task. For the multiclass authorship and book attribution tasks, where opposition lists cannot be easily constructed, features were ranked by $\omega\sigma$ using the weights in each linear model (the multiclass problems use multiple linear models for classification) and for each feature, its ranks in the various model were summed, producing an overall measure of the influence of the feature on classification for that task. The most significant features for these two tasks by this "rank sum" measure are shown in Table 2.

Examination of the features shows some clear differences among the two tasks, in terms of which sorts of features were most significant for classification, from which we can draw some tentative conclusions. Book discrimination involves a

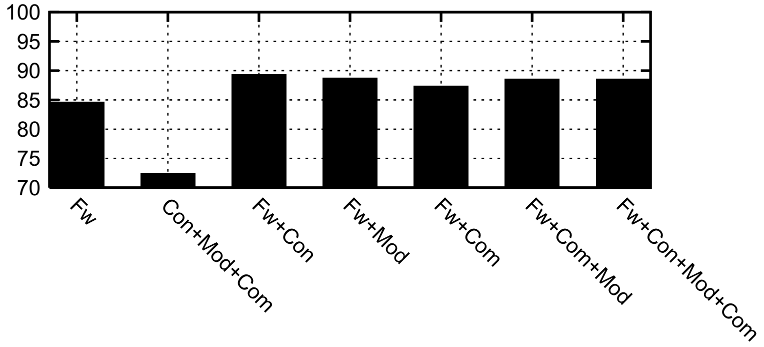

FIG. 1. Ten-fold cross-validation accuracy for authorship attribution in 19th-century literature. Baseline (majority class) classification would give 24% accuracy.

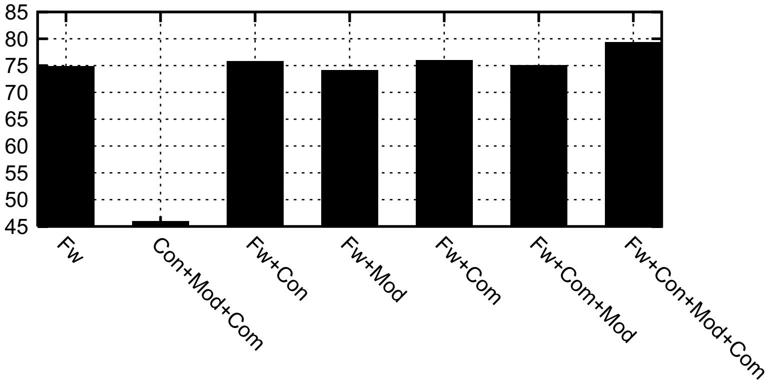

FIG. 2. Ten-fold cross-validation accuracy for book attribution in 19th-century literature. Baseline (majority class) classification would give 9.6% accuracy.

DOI: 10.1002/asi

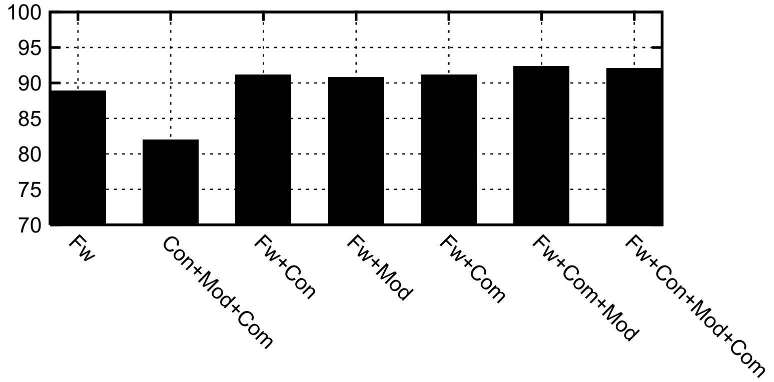

FIG. 3. Ten-fold cross-validation accuracy for nationality attribution in 19th-century literature. Baseline (majority class) classification would give 54.0% accuracy.

variety of features from all three functional systems used, with significant features drawn from CONJUNCTION (10 features), MODALITY (eight features), and COMMENT (two features). In author discrimination, we see functional features also focused on CONJUNCTION (14 features), but also see several kinds of function words of importance, including articles (“the,” “a”), personal pronouns (“I,” “she”), and connecting words (“and,” “of”). These results seem to indicate that while information structure (indicated by CONJUNCTION) varies among different authors and among different books, variation among authors involves different types of phoricity and phrase structure, and variation among books (even by the same author) involves variation in modal assessment.

TABLE 2. The top 20 features (by rank sum) for each of book, authorship, and nationality attribution.

Book attribution	Authorship attribution
COMMENT/Presumptive	CONJUNCTION/Extension
SPATIOTEMPORAL/Simple	CASUALCOND/Conditional
ELABORATION/Clarification	ELABORATION/Clarification
CONJUNCTION/Extension	SPATIOTEMP/Complex
COMMENT/Assertive	CASUALCOND/Causal
MODULATION/Obligation	CONJUNCTION/Enhancement
MODULATION/Obligation: MANIF/Implicit	SPATIOTEMP/Simple
CONJUNCTION/Enhancement	ELABORATION/Apposition
ELABORATION/Apposition	ENHANCEMENT/Matter
EXTENSION/Additive	ENHANCEMENT/CausalConditional
CASUALCOND/Conditional	EXTENSION/Additive
SPATIOTEMP/Complex	*the*
MODULATION/Obligation: VALUE/Low	EXTENSION/Adversative
CASUALCOND/Causal	ENHANCEMENT/Spatiotemporal
MODULATION/Obligation: ORIENT/Subjective	*and*
MANIF/Implicit	*I*
MANIF/Implicit: ORIENT/Subjective	*a*
ORIENTATION/Subjective	*she*
MODALIZATION/Usuality	*of*
ENHANCEMENT/ CausalConditional	EXTENSION/Verifying

Finally, Table 3 shows oppositions from the top 15 features (by $\omega\sigma$) for British and American authorship, respectively. We first note the striking difference in types of COMMENT found, where a British preference for Presumptive, Tentative, and Validative COMMENTs, which often function as hedges, constrasts sharply with a more muscular American preference for Admissive, Assertive, and Desiderative COMMENTs, which respectively project personal involvement, assertiveness, and personal desires. This interpretation, which sees a contrast in the features between a sort of British “personal genteel understatement” and an American “muscular assertive individualism,” is supported also by several oppositions found within MODALITY. For example, British MODALITY typically has Subjective ORIENTATION and Low VALUE while American has Objective ORIENTATION with High VALUE. The only seeming exception to this pattern is in Readiness, where British writers might prefer

TABLE 3. Oppositions from the 15 highest ranked systemic features for each class, with weights taken from the model learned using FW + Com + Mod for Nationality attribution.

Condition	British	American
COMMENT	Presumptive Tentative Validative	Admissive Assertive Desiderative
ORIENTATION	Subjective	Objective
VALUE	Low	High
TYPE/VALUE	Modalization/ Low	Modalization/ High
MODALIZATION/VALUE	Usuality/Low Probability/High	Usuality/High
MODULATION/VALUE	Obligation/Low Readiness/High	Obligation/High Readiness/Low
MANIFESTATION/VALUE	Implicit/Low	Implicit/High
MODALIZATION/ ORIENTATION	Usuality/ Subjective	Usuality/ Objective
MANIFESTATION/ ORIENTATION	Implicit/ Subjective	Explicit/ Subjective Implicit/Objective
TYPE	Modulation	Modalization

DOI: 10.1002/asi

High VALUE since High Readiness does not imply either certainty or necessity.

Characterizing Gender

We next examine the possibility of determining the gender (male or female) of literary characters based on their dialogue; for this, we considered characters from Shakespeare's plays. This extends previous results on classifying author gender (Argamon et al., 2003; de Vel et al., 2002) to examine the new question of whether and how a playwright might create linguistic distinctions between male and female characters.

We constructed a corpus of characters' speeches from 38 Shakespearean plays, based on text from the Moby Shakespeare (Shakespeare, n.d.). A text file for each character in each play was constructed by concatenating all the character's speeches in the play; characters' genders were catalogued. To improve robustness of the results, all characters

TABLE 4. Composition of the corpus of characters' speeches from Shakespeare. The table shows the number and average total speech length of selected characters of each gender from each play. The text has more detail on the construction of the corpus.

	Male		Female	
Play	Number	Avg. Length	Number	Avg. Length
All's Well That Ends Well	4	2,537	4	1,739
As You Like It	2	2,362	3	2,797
Cymbeline	5	2,583	2	2,734
Loves Labours Lost	1	2,384	4	1,000
Measure for Measure	3	3,522	2	1,698
Midsummer Nights Dream	0	n/a	3	1,393
Much Ado About Nothing	1	2,390	4	977
Pericles Prince of Tyre	1	4,688	4	757
The Comedy of Errors	2	2,346	4	904
The Merchant of Venice	2	2,737	3	1,914
The Merry Wives of Windsor	2	3,104	3	1,968
The Taming of the Shrew	1	3,892	2	1,080
The Tempest	1	4,880	1	850
Troilus and Cressida	4	3,262	1	2,133
Twelfth Night	3	2,384	3	2,019
Two Gentlemen of Verona	1	3,284	3	1,376
Winter's Tale	3	3,031	3	1,581
The first part of King Henry IV	4	4,155	2	342
The second part of King Henry IV	2	4,279	2	1,138
The Life of King Henry V	1	8,360	2	406
The first part of King Henry VI	1	1,910	2	302
The second part of King Henry VI	4	2,492	2	2,406
The third part of King Henry VI	2	2,802	3	708
The Life of King Henry VIII	2	3,032	2	1,672
The Life and Death of King John	3	3,136	0	n/a
The Life and Death of Richard II	3	3,790	2	829
The Life and Death of Richard III	2	4,372	4	1,524
Antony and Cleopatra	3	3,692	2	2,432
King Lear	3	3,528	2	1,086
Othello	3	5,481	3	1,551
Romeo and Juliet	4	2,856	3	2,397
The Life and Death of Julius Caesar	3	3,848	1	720
The Tragedy of Coriolanus	4	3,286	2	1,360
The Tragedy of Hamlet	5	4,618	1	1,325
Timon of Athens	2	4,113	0	n/a
Titus Andronicus	2	4,048	5	1,225

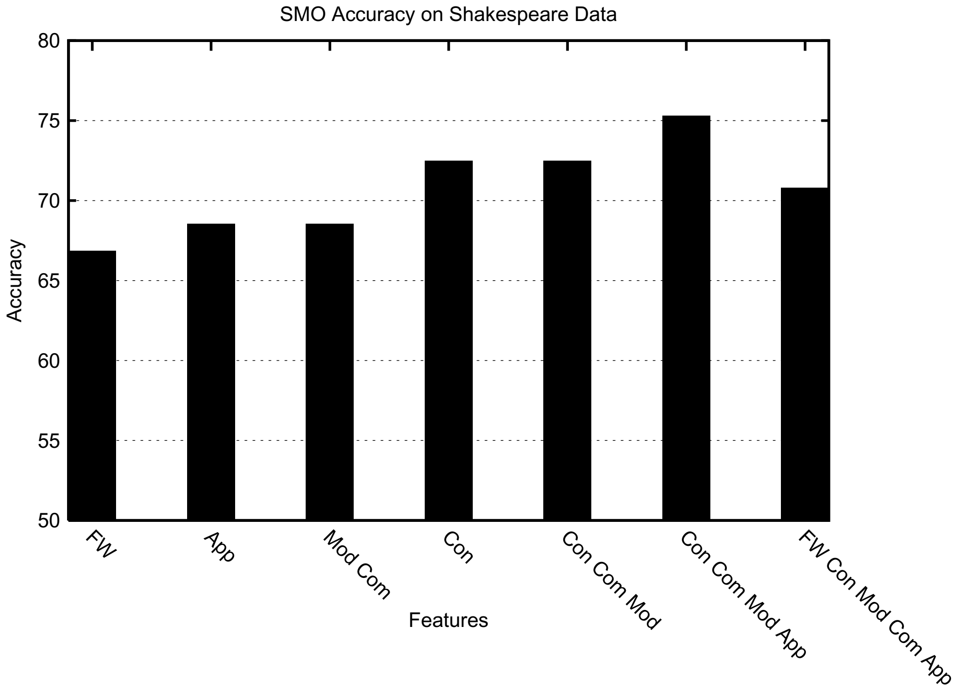

FIG. 4. Accuracies for various feature sets over the corpus.

with less than 200 total words in the corpus were discarded. We further balanced the corpus for gender by keeping all 89 female characters (with at least 200 words) together with the 89 male characters with the most words each, discarding the rest. The composition of the corpus is summarized in Table 4.

The feature sets used were FW, Con, Mod, Com, App, and their various combinations. Classification accuracies under 10-fold cross-validation are shown in Figure 4.

The highest classification accuracy obtained was 75.3% by the combination of all functional lexical features. When function words are added, accuracy noticeably decreases to 70.8% (though variance is high, giving $p = .1$). Note that this functional feature set contains just 141 features, compared to the 645 features in the FW feature set. This implies that the functional-lexical features are a better fit to this task than the function words; the addition of FW may be leading to some overfitting, reducing classification accuracy.

We also examined the top-ranked features (by $\omega\sigma$) from the Con + Mod + Com + App feature set for both genders; Table 5 shows all oppositions found among the 20 top male- and 20 top female-indicating features in the model learned for Con + Mod + Com + App. Several interesting differences between male and female characters are evident.

TABLE 5. Oppositions found in the top 20 features from both genders of Shakespearean characters. Note that they are organized categorically, not in rank order.

	Male		Female	
Condition	Feature	ωσ	Feature	ωσ
CONJUNCTION	Extension	4	Enhancement	3
EXTENSION	Additive	6	Adversative	6
ENHANCEMENT	Matter	5	CausalConditional	7
COMMENT	Validative	52	Admissive	8
	Presumptive	19		
	Desiderative	16		
	Tentative	14		
	Evaluative	9		
ATTITUDE/ORIENTATION	Affect/Positive	17	Affect/Negative	2
JUDGMENT/ORIENTATION	SocialEsteem/Negative	9	SocialSanction/Positive	3
	SocialSanction/Negative	6		
APPRECIATION ORIENTATION	ReactionQuality/Negative	6	ReactionImpact/Negative	3
GRADUATION	Force	14	Focus	11

DOI: 10.1002/asi

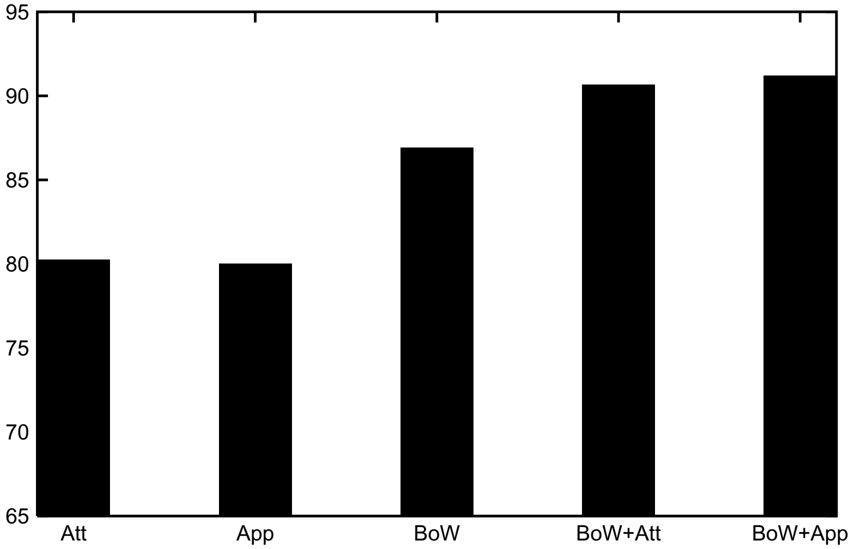

FIG. 5. Movie review classification results for using SMO with default parameters and a linear kernel with various feature sets; see text for further details.

First, we see that in CONJUNCTION, females tend to have Enhancement whereas males tend to have Extension. This means that male characters are more likely to link together many independent information units in their speeches, and so may imply that they have more monologues. This notion is supported by measuring the number of individual speeches that are greater than four lines, which shows that male characters in our corpus have a total of 2,388 such long speeches (26.8 on average) whereas the female characters have a total of 1,027 long speeches (11.5 on average).

Digging a little deeper, within EXTENSION, male characters prefer the additive variety whereas females prefer the adversative; females are more likely to contrast and compare. We further note that within ENHANCEMENT, female characters use CausalConditional constructs more than male characters do.

Male characters generally seem more likely to use COMMENTs, with the single exception of Admissive COMMENTs, which are preferred by female characters. This might indicate that males are generally more likely to actively contextualize their statements whereas females only do so as a form of hedging; this will require further work to elucidate.

Finally, examination of appraisal oppositions gives us an interesting set of patterns. Females express Negative Affect (feelings) while males tend more to express Positive feelings (perhaps suppressing negative ones). On the other hand, when expressing social evaluation through JUDGEMENT, males seem to more readily express Negative social attitudes than do females, consistent with a view of females being more socially accomodating and males being more aggressive.

Sentiment Analysis

Sentiment classification is the task of labeling a text as positive ("thumbs up") or negative ("thumbs down") based on the sentiment expressed by the author toward a target object (e.g., film, book, product, etc.). Important current applications include data and Web mining, market research, and customer relationship management.

Corpus. To test the usefulness of adjectival appraisal groups for sentiment analysis, we evaluated the effectiveness of the describes earlier feature sets for movie-review classification, using the publicly available collection of movie reviews constructed by Pang and Lee (2004). This standard testbed consists of 1,000 positive and 1,000 negative reviews, taken from the IMDb movie review archives.[3] Reviews with "neutral" scores (e.g., three of five stars) have been removed, giving a dataset with only clearly positive and negative reviews.

Features. Since the only lexical taxonomy in this article that is relevant to sentiment is Appraisal, the features used for sentiment analysis were Att and App (FW did not achieve appreciable accuracy.) In addition, since many content-bearing words bear sentiment of various types, we also included a "Bag-of-Words" feature set (*BoW*), defined as the relative frequencies (as for FW) of all words in the corpus. Combinations of FW and BoW with both appraisal feature sets also were considered.

Results. Figure 5 gives 10-fold cross-validation accuracy for SMO using different feature sets for sentiment analysis. Just using attitude-type of adjectival appraisal groups (Att) does surprisingly well, at 80.2% accuracy. This bears out our

[3] See http://www.cs.cornell.edu/people/pabo/movie-review-data/.

TABLE 6. Top 15 features for positive and negative reviews from BoW + App, showing ωσ (multiplied by 10,000 for scaling).

Negative Reviews		Positive Reviews	
Feature	ωσ	Feature	ωσ
APP/ReactionImpact:ORI/Negative	17.9	,	23.4
APP/ReactionQuality:ORI/Negative	12.8	ORI/Positive	18.8
ORI/Negative	12.7	ATT/Appreciation:ORI/Positive	16
ATT/Appreciation:ORI/Negative	12.2	*and*	13.6
bad	4.8	*is*	10.1
only	4.4	*he*	6.7
bad:APP/ReactionQuality:ORI/Negative	4.0	*as*	6.5
script	1.9	APP/ReactionImpact:ORI/Positive	5.5
nothing	1.7	*also*	3.2
worst	1.6	*well*	2.6
boring	1.4	*most*	2.6
unfortunately	1.0	*great*	2.3
mess	1.0	*quite*	1.8
awful:ATT/Affect:ORI/Negative	0.82	*fun*	1.5
awful	0.77	*hilarious*	1.0

hypothesis that attitude-bearing adjectives specifically are a key feature in the expression of sentiment. Using attitude type and orientation together (App) yields essentially the same accuracy (80.0%).

The limited-coverage appraisal feature sets are outperformed ($p < .001$), however, by standard bag-of-words classification using all words (BoW), which attains 87.05% accuracy using SMO, competitive with Pang and Lee's (2004) recent results for this dataset, based on classifying texts after using clustering to automatically extract subjective passages. More significantly, we clearly improve on that result when combining bag-of-words features (for coverage) with appraisal features, attaining 90.7% accuracy with BoW + Att ($p < .005$ vs. BoW) and 91.2% with BoW + App ($p < .0005$ vs. BoW). The difference between BoW + Att and BoW + App was not statistically significant. These results demonstrate that appraisal analysis can help sentiment classification.

The most significant 30 features (15 positive and 15 negative) are shown in Table 6. We see first of all that appraisal features are among the most important, with some lexical features adding fine-grained discrimination. Note that "bad" and "awful" occur both as bag-of-word features and as appraisal features, for negative reviews; this indicates that they are significant both in terms of their raw frequencies and in terms of the fraction of appraisal groups that contain them. Some non adjectival lexical features fill in gaps in our lexicon (e.g., "nothing," "mess"). Others seem to indicate some non evaluative stylistic features of positive reviews—the appearance of the comma, "and," "as," and "also" may indicate more complex syntax in positive reviews, "is" more direct assertions, and "he" more reference to people. A more detailed study of a larger corpus will be needed to further eluciate these features.

Little previous work on automated sentiment analysis has considered different attitude types as we do here. The only such work of which we are aware are recent works attempting to automatically determine the attitude type of a term. Kamps, Marx, Mokken, and Rijke (2002) used link-distances in WordNet to estimate three parameters–potency, activity, and evaluativity—based on Osgood, Succi, and Tannenbaum's (1957) theory of semantic distances. Taboada and Grieve (2004) also presented a method for automatically determining top-level attitude types in Appraisal Theory via application of Turney's (2002) PMI method. They observed that different types of reviews appear to contain different amounts of each attitude type, which our results partially confirm. Since the appraisal taxonomies used in this work are general purpose and were not developed specifically for sentiment analysis or movie-review classification, we expect appraisal group analysis to be highly portable to other related tasks.

Scientific Prose

Finally, we consider the question of whether different scientific fields have meaningfully distinctive language styles. To do so, we wanted to see if our functional lexical features could effectively classify peer-reviewed scientific articles from different fields, and if they can, whether the most indicative features give us any insight (also see our related study: Argamon, Dodick, & Chase, 2005, of experimental and historical science articles).

TABLE 7. Summary of the geology and paleontology journals used in the scientific literature corpus study, giving the number of articles from each journal in the corpus, and the average number of words per article.

Journal	# Articles	Avg. Words
Journal of Metamorphic Geology	108	5,025
Journal of Geology	93	4,891
Quaternary Research	113	2,939
Paleontologia Electronica	111	4,133

Corpus. For this experiment, we used a corpus comprising peer-reviewed articles from two geology and two paleontology journals (see Table 7):

> *Journal of Metamorphic Geology* focuses on metamorphic studies,[4] from the scale of individual crystals to that of lithospheric plates.
>
> *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.
>
> *Quaternary Research* publishes 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).
>
> *Paleontologica Electronica* publishes articles in all branches of paleontology as well as related biological or paleontologically related disciplines.

Features. The feature sets used were FW, Con, Mod, Com and App, and selected combinations thereof.

Results. Figure 6 shows the 10-fold cross-validation accuracies for the feature sets just mentioned. FW alone gave 84.9% accuracy; adding in App improved accuracy to 86.8% ($p = .059$), and adding in all systemic features yielded an accuracy of 87.5% ($p = .024$). The low accuracies obtained for each systemic feature set by itself indicate that the overall rhetorical structures used in these two fields tend to be similar, as we might have expected.

To examine the relevant stylistic differences between geology and paleontology, we next consider oppositions among the top 20 systemic features from the model constructed for the two article classes using the FW + Conj + Com + Mod + App feature set, as shown in Table 8. Appraisal is the most important, yielding the largest boost in classification power, as noted earlier, and accordingly generating many highly ranked oppositions. ORIENTATION is most important overall—geologists appear to prefer Positive appraisal while paleontologists prefer Negative. This opposition also is seen within JUDGEMENT and ATTITUDE. Such appraisal often appears when describing the results of the current, or of previous, research. Geology appears to prefer positive appraisal, stressing the cooperative nature of the research enterprise (e.g., ". . . collectively and *consistently* point to a single conclusion . . ."). On the other hand, paleontology tends to prefer negative orientation, seeming to stress inadequacies of the evidence or of previous work (e.g., in, ". . . records are *unfortunately* much more fragmented . . ."). As well, we see cases where a researcher will discredit previous work based on new evidence (e.g., ". . . the approach taken is fundamentally *flawed*"). It seems that

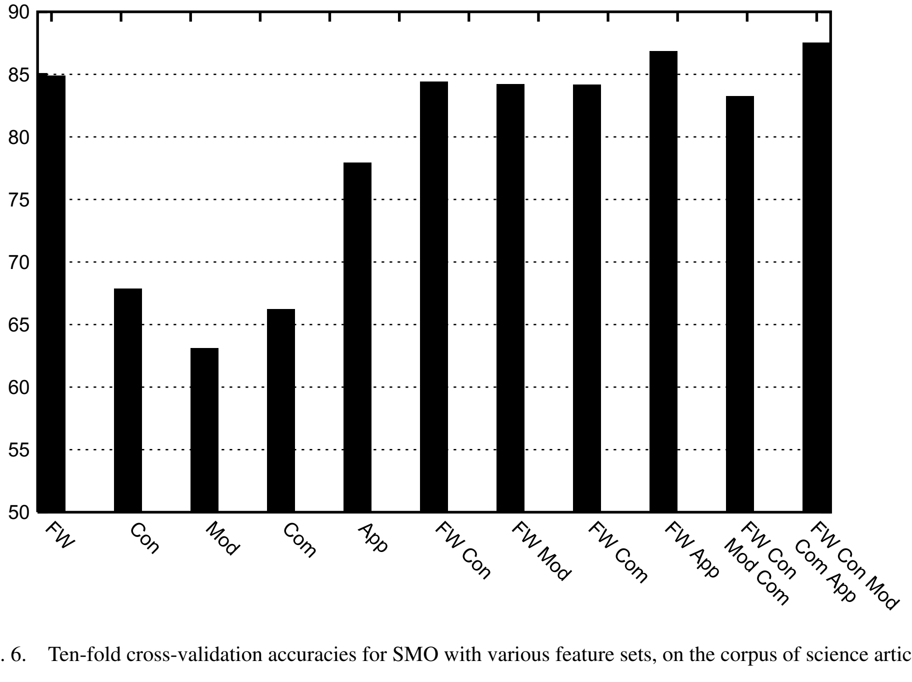

FIG. 6. Ten-fold cross-validation accuracies for SMO with various feature sets, on the corpus of science articles.

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

TABLE 8. Oppositions from the 20 highest ranked systemic features in geology and paleontology articles, from the model learned using FW + Con + Com + Mod + App.

Condition	Geology	ωσ	Paleontology	ωσ
ORIENTATION	Positive	7.02	Negative	3.94
JUDGMENT/SocialEsteem	ORIENT/Positive	3.24	ORIENT/Negative	3.03
JUDGMENT/SocialSanction	ORIENT/Positive	2.87	ORIENT/Negative	0.18
ATTITUDE/Judgement	ORIENT/Positive	2.67	ORIENT/Negative	1.60
ATTITUDE/Affect	ORIENT/Positive	1.84	ORIENT/Negative	2.96
APPRECIATION	ReactionQuality	4.26	CompositionComplexity	3.89
	ReactionImpact	3.49	CompositionBalance	3.04
SOCIALSANCTION	Propriety	5.02	Veracity	1.56
COMMENT	Assertive	3.05	Validative	16.4
	Desiderative	2.22	Presumptive	1.65
ENHANCEMENT	SpatioTemporal	2.27	CausalConditional	1.85

in a sense, geologists more often express positive views of previous work as they often add to it while paleontologists are more often negative, replacing old "truths" with new ones.

Next, oppositions in APPRECIATION indicate a distinction between a geological focus on Reaction (i.e., the effect of the object on an observer) and a paleontological focus on Composition (i.e., qualities of how the object is put together). This may indicate that paleontologists are more concerned with analyzing complex, multipart entities (e.g., fossils of various sorts) whereas geologists tend more toward uniform qualitative evaluations of specimens.

A similar distinction is seen in SOCIALSANCTION and in COMMENT. In SOCIALSANCTION, we see geologists more concerned with Propriety (i.e., how a methodology or a piece of evidence may fit with others) whereas paleontologists are more concerned with Veracity, in terms of how reliable particular methods or bits of evidence are on their own.

Similarly, we see two COMMENT types descriptive of geological prose: Assertive COMMENTs (e.g., "There is *surely* more to it . . ."), and Desiderative COMMENTs (e.g., "In doing so, *we hope* to deduce"), which is consistent with the apparent focus of geologists on "fitting in" noted earlier. Paleontologists, on the other hand, tend to use more Validative COMMENTs, expanding or contracting the scope of validity of a claim (e.g., "Holocene shells *generally* lack . . ."), and Presumptive COMMENTs, evaluating new claims in light of general background knowledge (e.g., ". . . which *apparently* are linked with . . .").

Finally, the single opposition we find within the CONJUNCTION system is in ENHANCEMENT, where geology prefers SpatioTemporal while paleontology prefers CausalConditional. Geological uses of SpatioTemporal conjunction tend to describe rock configurations and development over time as well as discourse organization (in a sort of descriptive "story line"). Paleontologists, however, are more often explicit about hypothesized causal links between specimens and historical events (e.g., ". . . perhaps *as a result* of migration . . .").

Conclusions

We have presented a novel set of linguistically motivated *functional lexical* features for stylistic text classification which have been shown to increase classification accuracy over the function word baseline for a variety of stylistic classification tasks, in some cases quite significantly. More significantly, our results show how different kinds of features are needed for different kinds of stylistic text classification, and that addition of irrelevant features often reduces performance (i.e., overfitting is difficult to avoid). Indeed, we have shown how the organization of these features in systemic taxonomies enables us to gain fairly detailed insights into the linguistic choices inherent in different text styles. This is because siblings in such a taxonomy represent functional alternatives in the language that can be meaningfully compared. Thus, one potentially important use of these methods is for *computational sociolinguistics*, extending the study of systematic linguistic variation from the very focused studies typical of traditional sociolinguistics (Labov, 1973; Trudgill, 2001) to larger scale studies examining lexical and grammatical variation related to geography, education, societal status, and other factors.

We are currently pursuing a number of directions to improve the methods presented in this article. The most immediate is extending the coverage of the functional lexical taxonomies, both increasing their completeness and adding new functional taxonomies. We also are currently developing methods for efficiently parsing phrases (e.g., adjectival groups) based on systemic functional principles; we recently presented early results (Whitelaw, Garg, & Argamon, 2005). We expect this to improve analysis by both enabling us to extract new functional attributes in a text and also correcting inaccuracies in our current purely lexical approach (since a group may have different functional attributes than its head—compare "good" with "not good"). Such shallow parsing also will help reduce ambiguity; for example, "even" on its own may be an adjective referring to APPRECIATION/QualityBalance whereas in a phrase "not even nice," it is clearly seen to be a modifier. More generally, recent work in such areas as Rhetorical Structure Theory (Marcu, 1997, 1999; Moore & Pollack, 1992) collocation analysis (Wiebe, McKeever, & Bruce, 1998; Wiebe, Wilson, & Bell, 2001), and coreference resolution (Cristea, Marcu, Ide, & Tablan, 1999; Ng & Cardie, 2002; Schauer & Hahn, 2001) also might be applied to improve and extend extraction of functionally relevant stylistic features.

Acknowledgments

Thanks to Ophir Frieder, Wai Gen Yee, Eric Jensen, and the anonymous reviewers for their careful readings of this article and many helpful comments, and to Jeff Dodick for enlightening discussions about the nature of the earth sciences. Casey Whitelaw was a visiting scholar at the IIT Linguistic Cognition Laboratory during November 2004.

References

Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., & Spyropoulos, C. (2000). An evaluation of naive bayesian anti-spam filtering. In Proceedings of the Workshop on Machine Learning in the New Information Age (pp. 9–17). New York: ACM Press.

Appelt, D., Hobbs, J., Bear, J., Israel, D., & Tyson, M. (1993). FASTUS: A finite-state processor for information extraction from real-world text. In Proceedings of the International Joint Conference on Artificial Intelligence (pp. 1172–1178).

Argamon, S., Dodick, J., & Chase, P. (2005). The languages of science: A corpus-based study of experimental and historical science articles. In Proceedings of the 26th Annual Meeting of the Cognitive Science Society (pp. 157–162).

Argamon, S., Koppel, M., Fine, J., & Shimony, A.R. (2003). Gender, genre, and writing style in formal written texts. Text, 23(3), 321–346.

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In Proceedings of the 2005 ACH/ALLC Conference. Retrieved from http://web.uvic.ca/hrd/achallc2005/abstracts.htm

Argamon, S., & Olsen, M. (2006, April). Toward meaningful computing. Communications of the ACM, 49(4), 33–35.

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.

Belkin, N.J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. Information Retrieval, 93, 55–66.

Berry, M.J., & Linoff, G. (1997). Data mining techniques: For marketing, sales, and customer support. New York: Wiley.

Burrows, J.F. (1987). Computation into criticism: A study of Jane Austen's novels and an experiment in method. Oxford, England: Clarendon Press.

Butler, C.S. (2003). Structure and function—A guide to three major structural-functional theories (No. 63–64). Amsterdam: John Benjamins.

Chaski, C.E. (1999). Linguistic authentication and reliability. In Proceedings of the National Conference on Science and the Law (pp. 97–148). San Diego, CA: National Institute of Justice.

Chen, S.Y., Magoulas, G.D., & Dimakopoulos, D. (2005). A flexible interface design for web directories to accommodate different cognitive styles. Journal of the American Society for Information Science and Technology, 56(1), 70–83.

Cowie, J., & Lehnert, W. (1996). Information extraction. Communications of the ACM, 39(1), 80–91.

Cristea, D., Marcu, D., Ide, N., & Tablan, V. (1999). Discourse structure and co-reference: An empirical study. In D. Cristea, N. Ide, & D. Marcu (Eds.), The relation of discourse/dialogue structure and reference (pp. 46–53). New Brunswick, NJ: Association for Computational Linguistics.

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

de Vel, O. (2000). Mining e-mail authorship. In ACM International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining. Boston. Retrieved from http://www.cs.cmu.edu/~dunja/WshKDD2000.html

de Vel, O., Corney, M., Anderson, A., & Mohay, G. (2002). Language and gender author cohort analysis of e-mail for computer forensics. In Proceedings of the Digital Forensic Research Workshop. Syracuse, NY. (pp. 7–9).

Fawcett, R.P., & Tucker, G.H. (1990). Demonstration of GENESYS: A very large, semantically based systemic functional grammar. In Proceedings of the 13th International Conference on Computational Linguistics (COLING-90) (pp. 47–49). Helsinki, Finland.

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 the 24th European Colloquium on Information Retrieval Research. Glasgow, United Kingdom: Springer Verlag, Heidelberg, DE.

Firth, J. (1968). A synopsis of linguistic theory 1930–1955. In F. Palmer (Ed.), Selected papers of J.R. Firth 1952–1959. London: Longman.

Fritch, J.W., & Cromwell, R.L. (2001). Evaluating internet resources: Identity, affiliation, and cognitive authority in a networked world. Journal of the American Society for Information Science and Technology, 52(6), 498–507.

Grossman, D., & Frieder, O. (1998). Information retrieval: Algorithms and heuristics. Dordrecht, The Netherlands: Kluwer.

Halliday, M.A.K. (1994). Introduction to functional grammar (2nd ed.). London: Arnold.

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

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

Hoover, D. (2002). Frequent word sequences and statistical stylistics. Literary and Linguistic Computing, 17, 157–180.

Kamps, J., Marx, M., Mokken, R.J., & Rijke, M. de. (2002). Words with attitude. In Proceedings of the 1st International Conference on Global WordNet. Mysore, India (pp. 332–341).

Karlgren, J. (2000). Stylistic experiments for information retrieval. Unpublished doctoral dissertation, SICS.

Kehagias, A., Petridis, V., Kaburlasos, V., & Fragkou, P. (2003). A comparison of word-and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems, 21(3), 227–247.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In P.R. Cohen & W. Wahlster (Eds.), Proceedings of the 35th annual meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (pp. 32–38). Somerset, NJ: Association for Computational Linguistics.

Kjell, B., & Frieder, O. (1992). Visualization of literary style. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (pp. 656–661). Chicago: IEEE Press.

Koppel, M., Akiva, N., & Dagan, I. (2003). A corpus-independent feature set for style-based text categorization. In Workshop on Computational Approaches to Style Analysis and Synthesis, 18th International Joint Conference on Artificial Intelligence.

Kushmerick, N. (1999). Learning to remove internet advertisement. In O. Etzioni, J.P. Müller, & J.M. Bradshaw (Eds.), Proceedings of the 3rd International Conference on Autonomous Agents (agents'99) (pp. 175–181). Seattle, WA: ACM Press.

DOI: 10.1002/asi

Labov, W. (1973). Sociolinguistic patterns. Philadelphia: University of Pennsylvania Press.

Lewis, D., Schapire, R.E., Callan, J.P., & Papka, R. (1996). Training algorithms for linear text classifiers. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval (pp. 298–306). New York: ACM Press.

Marcu, D. (1997). The rhetorical parsing of natural language texts. In Meeting of the Association for Computational Linguistics (pp. 96–103). Morristown, NJ: ACL.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. In Proceedings of the ACL'99 (pp. 365–372). Morristown, NJ: ACL.

Martin, J.R., & White, P.R.R. (2005). The language of evaluation: Appraisal in English. London: Palgrave. http://www.grammatics.com/appraisal/

Matthews, R.A.J., & Merriam, T.V.N. (1997). Distinguishing literary styles using neural networks. In E. Fiesler & R. Beale (Eds.), Handbook of neural computation (pp. 8). New York: IOP Publishing and Oxford University Press.

Matthiessen, C. (1983). Systemic grammar in computation: The nigel case. In Proceedings of the Meeting of the European Association for Computational Linguistics (pp. 155–164). Morristown, NJ: ACL.

Matthiessen, C. (1995). Lexico-grammatical cartography: English systems. Tokyo: International Language Sciences.

Matthiessen, C., & Bateman, J.A. (1991). Text generation and systemic-functional linguistics: Experiences from English and Japanese. London, New York: Pinter, St. Martin's Press.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. In Proceedings of the 17th International Conference on Machine Learning. Stanford, CA (pp. 591–598).

McEnery, A., & Oakes, M. (2000). Authorship studies/textual statistics. In R. Dale, H. Moisl, & H. Somers (Eds.), Handbook of natural language processing (pp. 234–248). Philadelphia: Dekker.

McKinney, V., Yoon, K., & Zahedi, F.M. (2002). The measurement of web-customer satisfaction: An expectation and disconfirmation approach. Information Systems Research, 13(3), 296–315.

McMenamin, G. (2002). Forensic linguistics: Advances in forensic stylistics. Boca Raton, FL: CRC Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Wordnet: An on-line lexical database. International Journal of Lexicography, 3(4), 235–312.

Moore, J.D., & Pollack, M.E. (1992). A problem for RST: The need for multi-level discourse analysis. Computational Linguistics, 18(4), 537–544.

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

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 104–111). Morristown, NJ: ACL.

O'Donnell, M. (1993). Reducing complexity in a systemic parser. In Proceedings of the 3rd International Workshop on Parsing Technologies. Tilburg, the Netherlands (pp. 10–13).

Osgood, C.E., Succi, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd ACL (pp. 271–278). Morristown, NJ: ACL.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (pp. 79–86). Morristown, NJ: ACL.

Patrick, J. (2004). The ScamSeek project: Text mining for finanical scams on the internet. In S. Simoff & G. Williams (Eds.), Proceedings of the 3rd Australasian Data Mining Conference (pp. 33–38).

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C.J.C. Burges, & A.J. Smola (Eds.), Advances in Kernel Methods—Support vector learning. Cambridge, MA, MIT Press.

Ponte, J.M., & Croft, W.B. (1998). A language modeling approach to information retrieval. In Proceedings of ACM SIGIR. New York: ACM Press.

Roth, D., & Yih, W. (2001). Relational learning via propositional algorithms: An information extraction case study. In Proceedings of the International Joint Conference on Artificial Intelligence (pp. 1257–1263).

Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Schauer, H., & Hahn, U. (2001). Anaphoric cues for coherence relations. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, & N. Nikolov (Eds.), Proceedings of the Euroconference Recent Advances in Natural Language Processing (RANLP-2001) (pp. 228–234). Tzigov, Bulgaria.

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

Shakespeare, W. (n.d.). The complete Moby Shakespeare. http://www-tech.mit.edu/Shakespeare/

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

Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In AAAI Spring Symposium on Exploring Attitude and Affect in Text. Menlo Park, CA: AAAI Press.

Tang, R., Ng, K.B., Strzalkowski, T., & Kantor, P.B. (2003). Toward machine understanding of information quality. In Proceedings of Annual Meeting of American Society for Information Science and Technology (Vol. 40, pp. 213–220).

Teich, E. (1995). A proposal for dependency in systemic functional grammar—Metasemiosis in computational systemic functional linguistics. Unpublished doctoral dissertation, University of the Saarland and GMD/IPSI, Darmstadt, Germany.

Torvik, V.I., Weeber, M., Swanson, D.R., & Smalheiser, N.R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. Journal of the American Society for Information Science and Technology, 56(2), 140–158.

Trudgill, P. (2001). Sociolinguistics: An introduction to language and society (4th ed.). New York: Penguin.

Turney, P.D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting of the ACL (pp. 417–424). Morristown, NJ: ACL.

Tweedie, F., Singh, S., & Holmes, D. (1996). Neural network applications in stylometry: The Federalist Papers. Computers and the Humanities, 30(1), 1–10.

Whitelaw, C., Garg, N., & Argamon, S. (2005, May). Using appraisal taxonomies for sentiment analysis. In Proceedings of the 2nd Midwest Computational Linguistic Colloquium (MCLC 2005).

Wiebe, J., McKeever, K., & Bruce, R. (1998). Mapping collocational properties into machine learning features. In Proceedings of the 6th Workshop on Very Large Corpora (pp. 225–233). Morristown, NJ: ACL.

Wiebe, J., Wilson, T., & Bell, M. (2001). Identifying collocations for recognizing opinions. In Proceedings of ACL/EACL 2001 Workshop on Collocation (pp. 24–31).

Winograd, T. (1972). Understanding natural language. Orlando, FL: Academic Press.

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

Yule, G.U. (1944). Statistical study of literary vocabulary. Cambridge, UK: Cambridge University Press.

Appendix: Functional Lexical Taxonomies

This Appendix describes in more detail the four types of functional lexical features used in the article, with their associated taxonomies of attribute values.

Conjunction

Words and phrases that conjoin clauses (e.g., "and," "while," and "in other words") are organized in SFG in the CONJUNCTION system network. Types of CONJUNCTION serve to link a clause with its textual context by denoting how the given clause expands on some aspect of its preceding context (Matthiessen, 1995, pp. 519–528). Similar systems also operate at the lower levels of noun and verbal groups, "overloading" the same lexical resources while denoting similar relationships, (e.g., *and* usually means "additive extension"). The three top-level options of CONJUNCTION are Elaboration, Extension, and Enhancement:

- Elaboration: Deepening the content in its context by exemplification or refocusing.
- Extension: Adding new related information, perhaps contrasting with the current information.
- Enhancement: Qualifying the context by circumstance or logical connection.

A more detailed picture of the system network, with examples of lexical items corresponding to each option, is given in Figure A1.

Different patterns of CONJUNCTION usage lead to markedly different textual styles. Frequent use of Extension can give a text with high information density which can give a "panoramic" effect of touring through a conceptual landscape, but if done poorly may overwhelm and lose a reader in too many facts. On the other hand, Elaboration can be used to good effect to create textual coherence around a single focused storyline. We note, too, that many of the standard function words traditionally used in computational stylistic studies are types of CONJUNCTION, which further argues for this system's importance for stylistic text analysis.

Modality

The system of MODALITY enables writers to qualify events or entities in the text according to their likelihood, typicality, or necessity. Syntactically, MODALITY may be realized in a text through a modal verb (e.g., "can," "might," "should," "must"), an adverbial adjunct (e.g., "probably," "preferably"), or use of a projective clause (e.g., "I think that . . .", "It is necessary that . . ."). Each expression of MODALITY has four attributes, corresponding to simultaneous choice of options within four system networks. These networks, with their top-level options, are as follows (The complete picture as we have implemented is shown in Figure A2.):

- Type: What kind of modality is being expressed?
  –Modalization: How "typical" is it?
  –Modulation: How "necessary" is it?
- Value: What degree of the relevant modality scale is being averred?
  –Median: The "normal" amount.
  –Outer: An extreme (either high or low) amount.
- Orientation: Relation of the modality expressed to the speaker/writer.
  –Objective: Modality expressed irrespective of the speaker/writer.
  –Subjective: Modality expressed relative to the speaker/writer.
- Manifestation: How is the modal assessment related to the event being assessed?
  –Implicit: Modality realized "in-line" by an adjunct or modal auxiliary.
  –Explicit: Modality realized by a projective verb, with the nested clause being assessed.

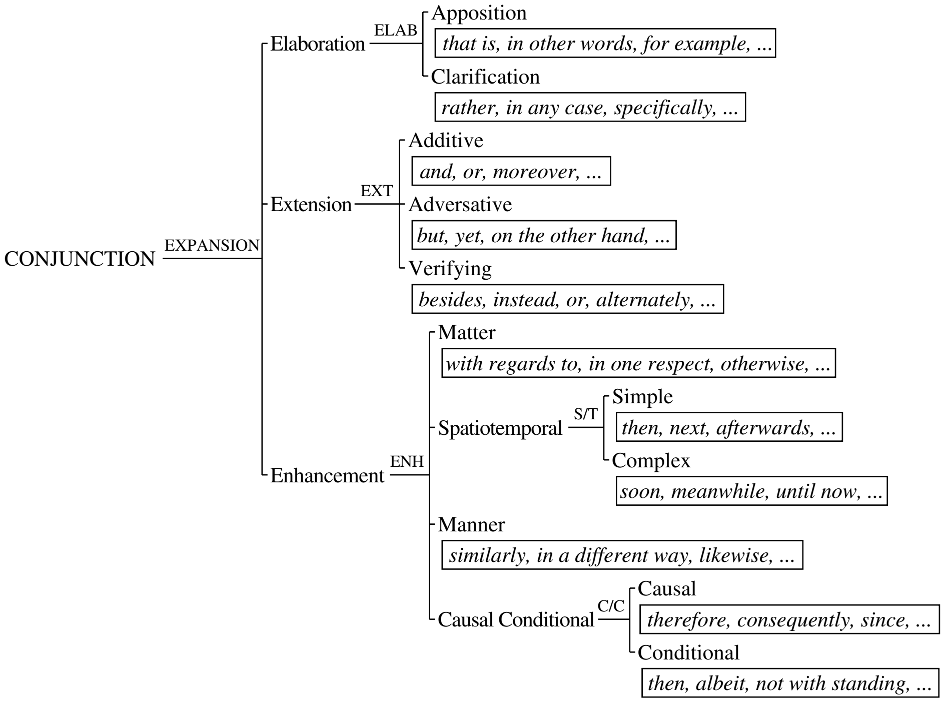
CONJUNCTION
EXPANSION
Elaboration
ELAB
Apposition
that is, in other words, for example, ...
Clarification
rather, in any case, specifically, ...
Extension
EXT
Additive
and, or, moreover, ...
Adversative
but, yet, on the other hand, ...
Verifying
besides, instead, or, alternately, ...
Enhancement
ENH
Matter
with regards to, in one respect, otherwise, ...
Spatiotemporal
S/T
Simple
then, next, afterwards, ...
Complex
soon, meanwhile, until now, ...
Manner
similarly, in a different way, likewise, ...
Causal Conditional
C/C
Causal
therefore, consequently, since, ...
Conditional
then, albeit, not with standing, ...

FIG. A1. The CONJUNCTION system (Matthiessen, 1995). Options here are disjunctive; examples of lexical realizations for the leaves are given in italics.

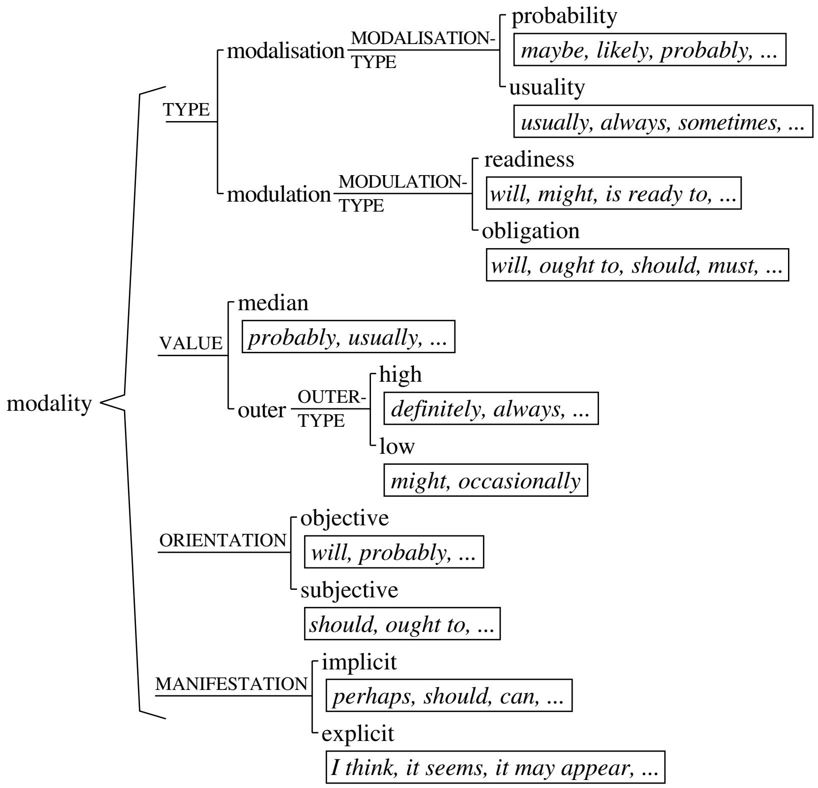

FIG. A2. The MODALITY system networks (Matthiessen, 1995).

			Type:Modalization		Type:Modulation	
			Probability	Usuality	Readiness	Obligation
Objective	Explicit	Median	*is likely*	*is frequent*	—	*is preferable*
Objective	Explicit	High	*is undeniable*	—	—	*is required*
Objective	Explicit	Low	*is possible*	*is infrequent*	—	*is permitted*
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*

FIG. A3. 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.

Any given expression of MODALITY will choose options in parallel from these four networks, though some combinations are rare or nonexistent. Figure A3 gives examples of lexical items for each possible combination of attributes.

Comment

The system of COMMENT provides a resource for the writer to "comment" on the status of a message with respect to textual and interactive context in a discourse. Comments are usually realized as adjuncts in a clause and may appear initially, medially, or finally. Matthiessen (1995), following Halliday (1994), lists eight COMMENT options, as follows:

- Admissive: Message is an admission (e.g., "Frankly . . .")
- Assertive: Emphasis of reliability (e.g., "Certainly . . .")
- Desiderative: Desirability of the content (e.g., "Unfortunately . . .")
- Evaluative: Judgment of the actors involved (e.g., "Sensibly . . .")

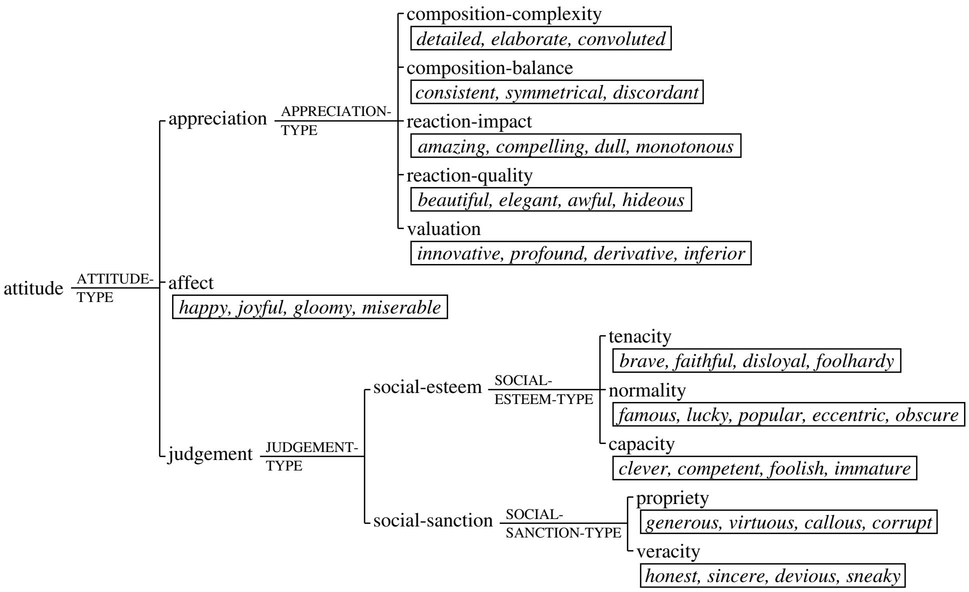
attitude
ATTITUDE-TYPE
appreciation
APPRECIATION-TYPE
composition-complexity
detailed, elaborate, convoluted
composition-balance
consistent, symmetrical, discordant
reaction-impact
amazing, compelling, dull, monotonous
reaction-quality
beautiful, elegant, awful, hideous
valuation
innovative, profound, derivative, inferior
affect
happy, joyful, gloomy, miserable
judgement
JUDGEMENT-TYPE
social-esteem
SOCIAL-ESTEEM-TYPE
tenacity
brave, faithful, disloyal, foolhardy
normality
famous, lucky, popular, eccentric, obscure
capacity
clever, competent, foolish, immature
social-sanction
SOCIAL-SANCTION-TYPE
propriety
generous, virtuous, callous, corrupt
veracity
honest, sincere, devious, sneaky

FIG. A4. Options in the Attitude network, with examples of appraisal adjectives from our lexicon.

- Predictive: Coherence with predictions (e.g., "As expected . . .")
- Presumptive: Dependence on other assumptions (e.g., "I suppose that . . .")
- Tentative: Assessing the message as tentative (e.g., "Tentatively . . .")
- Validative: Assessing scope of validity (e.g., "In general . . .")

Appraisal

Appraisal denotes how language is used to adopt or express an attitude of some kind toward some target. For example, in "I found the movie quite monotonous," the speaker adopts a negative *Attitude* ("monotonous") toward "the movie" (the *appraised object*). Note that attitudes come in different types; for example, "monotonous" describes an inherent quality of the appraised object while "loathed" would describe an emotional reaction of the writer. The overall type and orientation of appraisal expressed in the text about an object gives a picture of how the writer wishes the reader to view it (modulo sarcasm, of course). To date, we have developed a lexicon for appraisal adjectives as well as relevant modifiers (e.g., "very" or "sort of"); we are currently developing a shallow parser that will be able to extract adjectival appraisal groups as well as identify the appraiser and the appraised object.

Following Martin and White (2005), we define five appraisal attributes: Attitude, Orientation, Force, Focus, and Polarity:[5]

Attitude gives the type of appraisal being expressed as either *affect, appreciation*, or *judgment*. Affect refers to a personal emotional state (e.g., "happy," "angry"), and is the most explicitly subjective type of appraisal. The other two options express evaluation of external entities, differentiating between evaluation of intrinsic *appreciation* of object properties (e.g., "slender," "ugly") and social *judgment* (e.g., "heroic," "idiotic"). Figure A4 gives a more detailed view of the various options in Attitude, together with illustrative adjectives of each type. In general, attitude may be expressed through nouns (e.g., "triumph," "catastrophe") and verbs (e.g., "love," "hate") as well as adjectives; we are currently working on expanding our lexicon to include nouns and verbs as well.

Orientation is whether the appraisal is *positive* or *negative* (often termed "sentiment").

Force denotes the intensity of the appraisal being expressed; for example, "good" will have neutral Force while "great" will have high Force, and "the best" will have maximal Force.

Focus is another aspect of the graduation of appraisal, referring to the "prototypicality" of the appraisal being expressed; for example, the modifier "truly . . ." is a Focus sharpener while "sort of . . ." is a Focus softener.

Polarity of an appraisal is *marked* if it is scoped in a polarity marker (e.g., "not"), or *unmarked* otherwise. Other attributes of appraisal are, of course, affected by negation; for example, "not good" expresses a different sentiment from "good."

Appraisal adjectives take on attribute values from all five appraisal attributes as described earlier. Appraisal modifiers, on the other hand, have values for just the latter four attributes, as Attitude type cannot be modified.

A value for each appraisal attribute is stored for each appraisal adjective; for example, the lexical entry for

[5] Note we use the term "Polarity," in its SFG sense to denote the grammatical notion of "explicit negation of a quality or assertion within the scope of the particle "not" or the equivalent," although this term has sometimes been used to mean what we refer to as "Orientation."

DOI: 10.1002/asi

"beautiful" reads:

'beautiful'	
Attitude:	appreciation/reaction-quality
Orientation:	positive
Force:	neutral
Focus:	neutral
Polarity:	unmarked

Modifiers, mostly adverbs, give transformations for one or more appraisal attributes, for example:

'very'	
Force:	increase

or polarity modification:

'not'	
Orientation:	negate
Force:	reverse
Polarity:	marked

Modifiers can specify effects on multiple appraisal attributes at once (e.g., "really" functions both as an intensifier of force and a sharpener of focus).

The experiments reported in this article only consider the Attitude and Orientation attributes of appraisal adjectives; elsewhere, we have presented some early results using shallow parsing of adjectival groups (Whitelaw et al., 2005).