Systemic Functional Features in Stylistic Text Classification

Casey Whitelaw
School of Information Technologies
University of Sydney, NSW, Australia
casey@it.usyd.edu.au

Shlomo Argamon
Department of Computer Science
Illinois Institute of Technology
10 W. 31st Street, Chicago, IL 60616, USA
argamon@iit.edu

Abstract

We propose that textual 'style' should be best defined as 'non-denotational meaning', i.e., those aspects of a text's meaning that are mostly independent of what the text refers to in the world. To make this more concrete, we describe a linguistically well-motivated framework for computational stylistic analysis based on Systemic Functional Linguistics. This theory views a text as a realisation of multiple overlapping *choices* within a network of related meanings, many of which relate to non-denotational (traditionally 'pragmatic') aspects such as cohesion or interpersonal distance. Variations between relative frequencies of options within these systems corresponds to stylistic variation. Though full parsing in SFL remains a difficult unsolved problem, we present a software architecture which allows for efficient modelling and extraction of SFL entities for use in stylistic analyses. In support, we show results for two applications: classifying texts as financial scams and classifying scientific articles by field.

1 Introduction

Modern computational stylistics seeks to apply statistical and machine learning techniques to features extracted from text, in order to answer style-related questions such as authorship attribution, genre characterisation and so forth. Its origins are in the field of *stylometrics* in which small numbers of features of a text were sought that statistically significantly predict a given stylistic difference. Traditionally, stylometric models for categorization have typically been based on hand-selected sets of content-independent, lexical [21], syntactic [25], or complexity-based [30] features. In nearly all cases, the input feature sets used are not theoretically motivated, from a linguistic perspective, as being directly related to style—rather the methodology that has developed is to find as large a set of topic-independent features as possible and use them as input to a generic learning algorithm (preferably one resistant to overfitting). While some interesting and effective feature sets have been found in this way (such as [15; 16]; function words have also proven to be surprisingly effective on their own [1; 20]), we believe 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.

How then should we approach finding more theoretically-motivated stylistic feature sets? First consider the well-accepted notion that style is indicated by features that indicate the author's choice of one mode of expression from among a set of equivalent modes. At the surface level, this may be expressed by specific choice of words, syntactic structures, discourse strategy, or combinations. The causes of such surface 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* of the text, which may be considered to be those objects and events that it refers to (as well as their properties as described in the text). We thus provisionally define the *stylistic meaning* of a text to be those aspects of its meaning that are *non-denotational*, i.e., independent of the objects and events to which the text refers. Indeed, this definition of style as characteristic choice unrelated to basic functionality is applicable outside the textual domain.

Our goal, therefore, is to find a computationally tractable formulation of linguistically well-motivated features which permit expression of non-denotational meanings. We believe that Systemic Functional Linguistics (SFL) is a particularly apposite theory for this purpose. As we will detail further below, SFL [12] analyses a text as a realisation of particular choices of meanings. For work that seeks to categorise a text according to the meanings that it makes rather than just the words that it uses, SFL presents an extremely useful model. In particular, the fact that SFL explicitly recognises and represents *non-denotational* meanings makes it particularly applicable to stylistical problems.

We describe in this paper a software architecture for computing aspects of language use in an SFL paradigm. Full parsing of SFL is a difficult and unsolved problem [22] that has been seen as holding back the applicability of systemic functional theory to NLP applications. We avoid this complexity by developing a partial parsing approach, based on the notion of identifying *semantic units* in the text, which are specific instances of systems defined in the underlying SFL grammar. Application of our methods to stylistic text categorization shows that SFL is well-suited to identifying document-

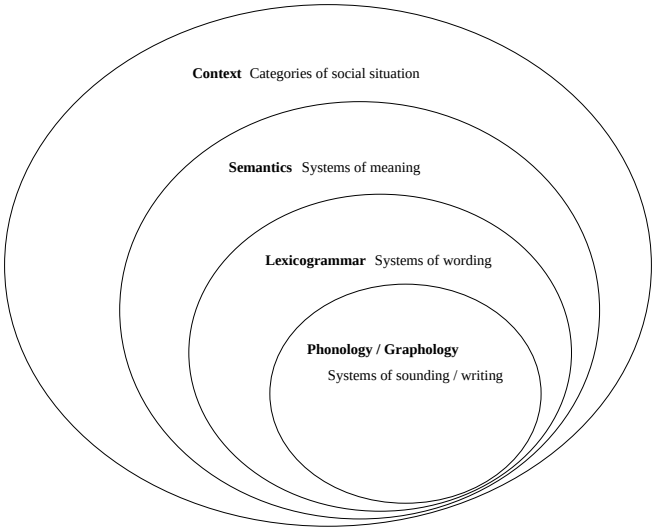

Figure 1: Modelling language stratally (after [14]). Units in each stratum are realised by configurations of units in the enclosed stratum.

level characteristics of language use.

2 Systemic Functional Linguistics

Systemic Functional Linguistics (SFL) is a framework for describing and modeling language in functional rather than formal terms. The theory is *functional* in that language is interpreted as a resource for making meaning, and descriptions are based on extensive analyses of written and spoken text [12]. The theory is also *systemic* in that it models language as a system of choices [19]. SFL has been applied in natural language processing in various contexts since the 1960s, but has been used most widely in text generation [18; 26], due to the difficulty of full parsing in the theory.

As noted, SFL construes language as a set of interlocking choices for expressing meanings, with more general choices constraining the possible specific choices. For example: "A message is either about doing, thinking, or being; if about doing, it is either standalone action or action on something; if action on something it is either creating something or affecting something preexistent," and so on. Thus a system is a set of options for meanings to be expressed, with entry conditions denoting when that choice is possible—for example, if a message is not about doing, then there is no choice possible between expressing standalone action or action on something. Each system has also a realization specification, giving constraints (lexical, featural, or structural) on statements expressing the option. Options often serve as entry conditions for more specific (or *delicate*) systems.

By viewing language as a complex of choices between mutually exclusive options, the systemic approach enables effective characterisation of variation in language use. As will be seen, a systemic specification allows us to create features related to high-level linguistic variation. A general preference for one or another non-denotational option, will usually indicate individual or social/contextual factors. Such *stylistic* preferences can be measured by evaluating the relative probabilities of different options by tagging their realisations in a corpus of texts [11].

2.1 Stratification and Realisation

One key global dimension is the hierarchy of stratification. Language itself is modelled as an ordered series of levels or strata, as shown in Figure 1. In general, we are interested in systems that capture an aspect of the meaning of the text (located within the semantic stratum) and is expressed as a pattern of word usage (realised within the lexicogrammatic stratum).

Register

A *register* is a group of texts whose language selections vary from the general language system in similar ways. Register often denotes functional distinctions in language use related to the context of such use [8], or it may derive from the particular history of a community of discourse. A register is realised by a particular skewing 'of probabilities relative to the general systemic probabilities' [17]. We thus view writing styles as registers realised by skewing the probabilities of non-denotational meanings in the language. These meanings in a document in turn are realised in the semantics and lexicogrammar, and so may be analysed on these terms. In particular, register differences may be exposed through the patterns of language choice within a system.

For example, Plum and Cowling [24] show a relation between speaker social class and choice of verb tense (past/present) in face-to-face interviews. Similarly, Hasan [13] has shown, in mother-child interactions, that the sex of the child and the family's social class together have a strong influence on several kinds of semantic choice in speech. The methodology for such studies has involved hand-coding a corpus for systemic-functional and contextual variables and then evaluating how systemic choice probabilities vary with contextual factors using multivariate analysis (e.g., principal components analysis). Here, by contrast, we develop automated methods for recognising entities which realise such variables in the text, allowing us to use machine learning techniques to automatically build accurate classification models for large numbers of SFL-based features.

2.2 Metafunction

Key to our use of SFL for stylistic analysis is the notion of *metafunction*, referring to three separate strands of meaning that in parallel contribute to the overall meaning in the text [12]. Briefly, the three metafunctions, deployed simultaneously in a text, are as follows:

The textual metafunction provides 'resources for presenting information as text in context' [19]. These resources enable individual utterances to be evaluated as messages, as well as related to the context of utterance (either other utterances or extra-textual features). A simple example is the use of 'textual theme', which sets up a clause as connected to a previous clause in a text, as in the conjunctive 'However...' or the continuative 'Yes, but...'.

The interpersonal metafunction provides the resources for enacting social roles and relations as meaning in a

text. This includes a variety of resources for the author/speaker to construct a text as a dialogue with the reader. Realisations of such meanings are given, for example, in use of personal pronouns, clausal mood (e.g., declarative vs. interrogative sentences), level of formality, and so forth.

The ideational metafunction provides the resources for construing our experience of the world in terms of objects, events, and relations between them. This may be divided into the **experiential** metafunction, allowing objects and events to be symbolised in language, and the **logical** metafunction, allowing conjunctive, logical, and causal meanings to be expressed.

Putting together the above notions of register and of metafunction, our initial rough 'definition' of language style as 'non-denotational meaning' may be made (provisionally) more precise. We posit that stylistic features of a text can usefully be considered those features which (a) constitute realisations of the textual and interpersonal metafunctions (and perhaps the logical), and (b) vary systematically with context (register). In this vein, the remainder of this section will flesh out some important systems in the textual and interpersonal metafunctions which we think particularly important for computational stylistics. Note that the systems we describe below are the result of years of research on textual analysis within the SFL community, and are not *ad hoc* inventions for our particular purposes.

2.3 The Textual Metafunction

The textual metafunction is that which enables expression of how parts of a text are related to other parts of the text or to the greater context. Resources in the textual metafunction enable a clause to be assessed as a 'message', related to other clauses and the context of discourse. One main purpose of the textual metafunction is to achieve coherence of the text, i.e., to structure it so that it forms a single whole that 'hangs together'. Texts may cohere in different manners; the way in which a specific text coheres expresses how the author organises ideas and relates them to each other. We discuss here the central textual system of CONJUNCTION, which we have used for stylistic analysis [19, p. 519–528].

Conjunction

On the discourse level, the system of CONJUNCTION serves to link a clause with its textual context, by denoting how the given clause expands on some aspect of its preceding context. Similar systems also operate at the lower levels of noun and verbal groups, 'overloading' the same lexical resources while denoting similar logico-semantical relationships, e.g., *and* usually means "additive extension". The three options within CONJUNCTION are Elaboration, Extension, and Enhancement. Each of these options gives entry to a more specific system with its own options. The main options, and options they give entry to, are:

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

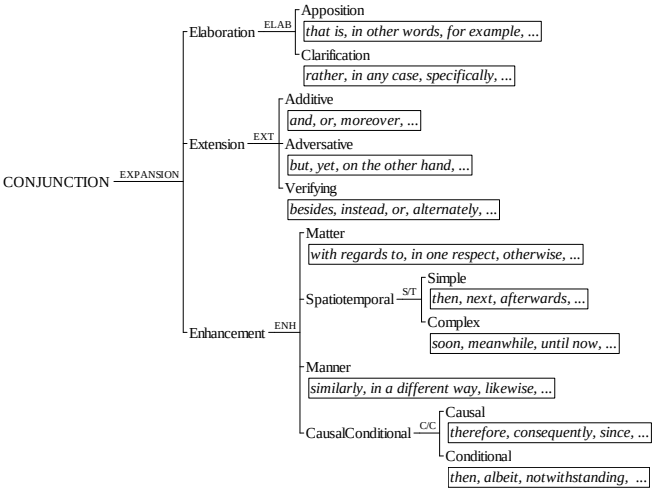

Figure 2: The CONJUNCTION system [19]. Options here are disjunctive; examples of lexical realisations for the leaves are given in italics.

- Enhancement: Qualifying the context by circumstance or logical connection.

A more detailed picture, with examples of lexical realisations, is given in Figure 2.

Different patterns of CONJUNCTION usage lead to markedly different textual styles. Frequent use of Extension gives 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 realisations of CONJUNCTION, which further argues for this system's importance in this context.

2.4 The Interpersonal Metafunction

The interpersonal metafunction deals with the way in which an writer/speaker establishes a dialogue with a reader/listener. We consider here three phenomena within the interpersonal metafunction which we find useful for stylistic text analysis: *interpersonal distance*, which relates to the tenor of the relationship between the writer and reader [9], *comments* which assess the status of a message relative to its (textual and interpersonal) context, and *modality* which relates to how events or assertions are framed for their typicality or necessity. All these phenomena construe a clause as a 'move' taken by one of the participants in a discourse, either relating the move to the participants' relationship, or clarifying the move's place in the larger discourse.

Nominals and Determination

The system of PRONOMINALDETERMINATION(Figure 3) describes how a referent can be identified. It positions the referent in relation to the speaker and listener, contributing to the establishment of the *interpersonal distance* of a text.

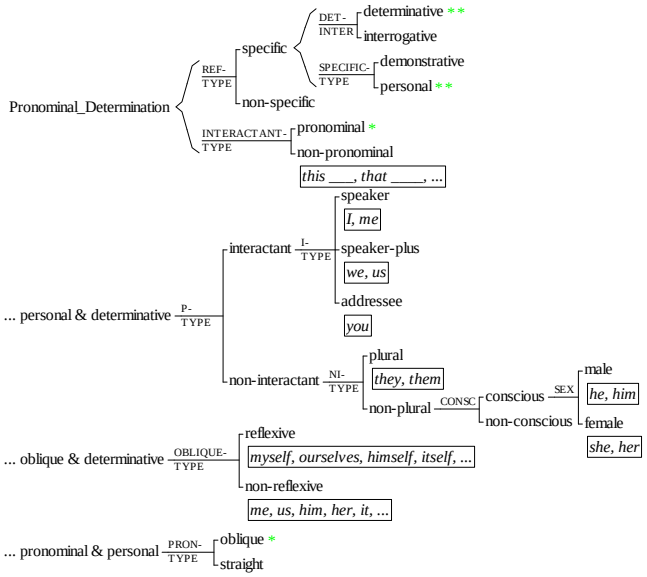

Figure 3: The PRONOMINALDETERMINATION system [19]. Options in curly braces are conjunctive (i.e., one option in every child must be chosen), others are disjunctive; examples of lexical realisations for the leaves are given in italics. Note the conjunctive entry conditions to the three lower systems.

Interpersonal distance relates to the tenor of the relationship between the writer and reader expressed in a text. Typically, spoken discourse with both oral and visual contact is representative of minimal interpersonal distance whereas written discourse with no visual, oral or aural contact represents maximal interpersonal distance.

Interpersonal distance can be determined by analysing various systemic language choices made within a text. Examples of such an analysis might include measuring the degree and frequency of participant nominalisations deployed within a text as well as the frequency and type of interactant reference [9; 7].

An example of a text with very close interpersonal distance would be one that includes direct speech, such as the following [4]:

> Kupe went to Muturangi's village and spoke of the bad behaviour of the animal with regard to his people's bait, saying, '**I** have come to tell **you** to kill **your** octopus.', Muturangi replied, '**I** won't agree to **my** pet being killed. Its home is in the sea.' 'Well', said Kupe, 'if **you** won't take care of **your** pet, **I** will kill it.'

In the above text, degree and frequency of nominalisation is low and selections from the Interactant system (bold face) are high.

A written history text, on the other hand, gives a good example of a text that constructs maximum interpersonal distance partly by making no selections from within the Interactant system [5]:

> The discovery of Hawaii from the Marquesas was a remarkable achievement, but at twenty degrees north latitude Hawaii is still within the zone of the trade winds that blow steadily and predictably for half of each year. New Zealand lies far to the South of the trade winds, in the stormy waters and unpredictable weather of the Tasman Sea.

It is expected that very close interpersonal distance in a text would be characterised by frequent selections from the INTERACTANT systems.

Comment

The system of COMMENT is one of message assessment. It comprises a variety of types of 'comment' on a message, assessing the status of the message with respect to its greater context, such as the writer's attitude towards it, or its validity or evidentiality. Comments are usually realized as adjuncts in a clause and may appear initially, medially, or finally. Matthiessen [19], following Halliday [12], lists eight COMMENT options, as follows:

- Admissive: Message is an admission (*Frankly...*)
- Assertive: Emphasis of reliability (*Certainly...*)
- Desiderative: Desirability of the content (*Unfortunately...*)
- Evaluative: Judgement of the actors involved (*Sensibly...*)
- Predictive: Coherence with predictions (*As expected...*)
- Presumptive: Dependence on other assumptions (*I suppose that...*)
- Tentative: Assessing the message as tentative (*Tentatively...*)
- Validative: Assessing scope of validity (*In general...*)

Modality

The system of MODALITY qualifies events or identities in the text according to their likelihood, typicality, or necessity. There are two main types: Modalization, which quantifies levels of likelihood or frequency (e.g., "probably", "might", "usually", "seldom"), and Modulation, which quantifies ability or necessity of performance (e.g., "ought to...", "should...", "allows..."). Based on Halliday's [12] analysis of the Modality system, as formulated by Matthiessen [19], MODALITY comprises simultaneous choices of options within four systems (outlined here, see Fig. 4 for a more complete picture):

- Type: What kind of modality?
  - Modalization: How 'typical' is it?
  - Modulation: How 'necessary' is it?
- Value: What degree of the relevant modality scale is being averred?
- Orientation: Relation of the modality expressed to the speaker/writer.
  - Objective: Modality expressed irrespective of the speaker/writer.

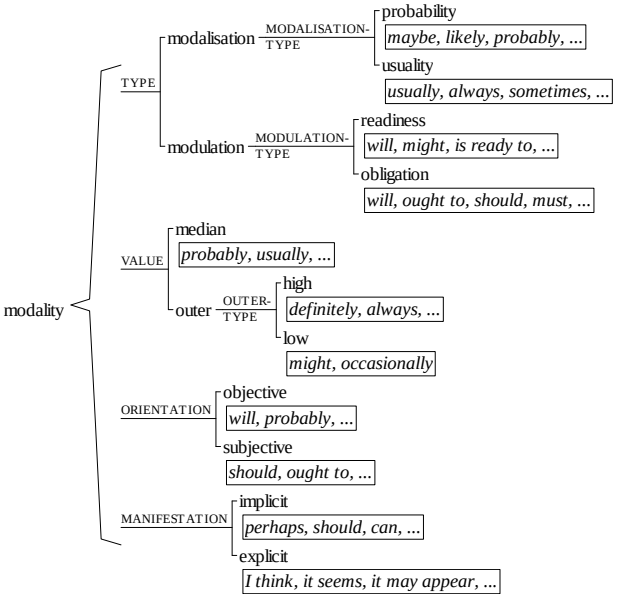

Figure 4: The MODALITY system [19], notation as above.

  - 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 finite verb.
  - Explicit: Modality realized by a projective verb, with the nested clause being assessed.

Any given expression of MODALITY will choose in parallel from these four systems, though some combinations are rare or non-existent.

3 Systemic Features

3.1 Modelling systemic features

The SFL notion of system networks is a powerful formalism for describing systemic phenomena in text. This descriptive framework was built to be a comprehensive and flexible mechanism to be used by linguists, but was not designed with computation in mind. In order to start computing systemic features, we use a new formalism that has a clearer relationship to a processing architecture. The intention is to capture as much of the descriptive power of the original as possible, while respecting the need for efficient processing. The model is based on the notion of *semantic units* and relationships between them. There are three types of units:

Categories: A category contains links to its member units. If **any** of the member units are present, the category is present also. This is the most prevalent relationship in the SFL systems given in Section 2. For example, CONJUNCTION is a Category whose members are the units for Elaboration, Extension, and Enhancement (Figure 2).

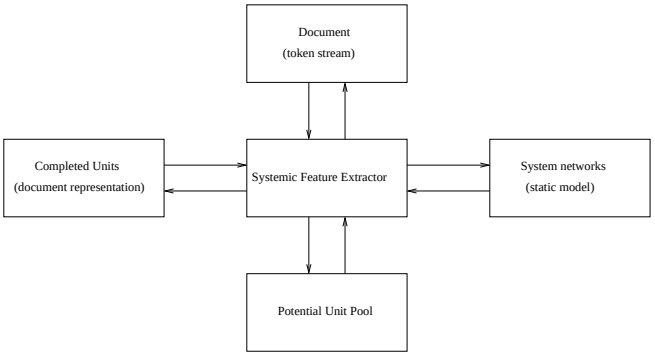

Figure 5: System architecture for systemic feature extraction

Sets A Set consists of two or more member units. If **all** the member units are found in the text, the Set is found also. Unit order is disregarded. Sets are most often used when dealing with the interaction between systems; for instance, the occurence of interactant-speaker from PRONOMINALDETERMINATION together with high-valued modalisation from MODALITY.

Sequences A Sequence is a Set with an imposed ordering. Like Sets, Sequences can be used to model conjunction in entry conditions, and are also used to describe phrases as collocations of individual lexical units.

Both Sets and Sequences may be constrained to a span of text, defined in terms of logical boundaries at word, clause, sentence, or paragraph. At the lowest level, a unit is realised as a property of a token in the text. Each token in the document may be assigned any arbitrary properties, such as its lexis (surface form), lemma, morphological information, parse location, sense tag, as well as features useful in stylistics such as word length etc. Each of these properties functions individually and independently within the system network.

This descriptive model is not specific to describing systemic functional linguistic phenomena. Any systemic features of a token stream, textual or otherwise, can be modelled in the same way. While we focus here on style, system networks can be used to generalise topic-based features as well.

3.2 Extracting systemic features

System networks, described using the model given above, can form arbitrarily large and complex features that may span an entire document. Relationships are not constrained to occur between certain classes of unit; a Sequence may be defined to look first for an exact word ("Dear", as a letter opener) at the beginning of a document, followed by a high-level construct ('well-wishing') at some later point. This flexibility makes the system networks useful for describing a broad range of useful systemic features, but also brings challenges for efficient feature extraction.

In a system network, a single unit may play multiple roles; it may belong to multiple categories, or as part of multiple sets or sequences. This heterogeneity of meaning is not resolved to a single unambigious decision, but is embraced as an essential part of the system network. Textual features indeed

contribute to multiple meanings, and this multidimensionality is preserved in our semantic extractor. For example, the word "*should*" realises several options within MODALITY (Fig. 4) simultaneously: Modulation/obligation, Value/median, Orientation/objective, and Manifestation/implicit, any or all of which may be relevant to a given stylistic question[1].

Systemic features do not define a full grammar that may be applied to the text; the are both ambiguous and incomplete. Only some sections of the text are modelled. This is in contrast to a traditional parsing problem, where an unambiguous structure must be found for all tokens in a sentence.

Figure 5 shows the basic architecture of our systemic feature extractor. The process is performed linearly, taking a single token from the document at a time. Each property of the token is checked against the pre-defined system model, and may result in the creation of one or more instances of basic units. These in turn contribute, via a sort of 'marker-passing', to the creation of other instances; if a completed unit is a member of a category, an instance of that category is also created. If it is a member of a set, or the initial member of a sequence, a *potential instance* of that unit is created. This instance is potential in that, for it to be fully instantiated, it requires other instances to be found. A pool of potential instances are managed separately and notified of the completion of their required instances, or removed if their conditions are not met within their lifespan. The completed instances are collected and form the full representation of the document as described by the system networks used. These can then be used as the basis of systemic feature representations, as described below.

This extraction model is very general, and can extract instances of any system defined according to our formalism. The architecture is designed specifically to efficiently handle the types of features that are common in SFL analysis; these features may be long-range, rely on both lexical and syntactic properties, and be defined in terms of other existing systems.

This systemic extractor relies on a system network 'bottoming out' in instantiation as lexis. Most grammatical systems do not describe a full system to the level of lexis; these systems cannot be extracted directly, but must be described in terms of other, lower-level features. These may be derived from SFL theory, or existing semantic resources such as WordNet or VerbNet. Since we are interested in the non-denotational meaning of a text, the systems we use are generally small and highly lexicalised, making them particularly amenable to a shallow processing approach.

3.3 Representing systemic features

Features in a system are not independent. They are correlated specifically by their relationships in the system network. This can be used to create new meaningful feature representations not available to naive representations such as bag-of-words. This is one of the strengths of using systemic features with machine learning.

In a standard 'bag-of-words' approach, the contribution of a word to a document is given by its relative frequency: how rarely or often that word is used. This implicitly assumes that all words in a text occur independently of one other. Crucially, this does not and cannot take into account the *choice* between words, since there is no representation of this choice. Placing words within a system network provides a basis for richer and more informative feature representation.

There are two main advantages gained by adding systemic information for feature representation. Firstly, it allows for categorical features that are based on semantically-related groups of words, at all levels in the network. By collecting aggregate counts, individual variations within a category are ignored. For a given register, it may be the case that important and characteristic language choice occurs at a very fine level, distinguishing between usage of individual words. This word-level information is kept intact, as in a bag-of-words approach. In another register, it may be the usage of a category, such as 'interactant' within PRONOMINALDETERMINATION, that is characteristic. The usage of any words within the category may appear to be entirely random, while the category itself is used very consistently. These higher-level features are not available in a traditional bag-of-words approach, in which these patterns may be lost as noise.

The second and more important difference to traditional feature representation is the representation of language choice. Not only can a system instance calculate the frequency of usage for categories within a system, it can calculate the relative usage within a category. *System contribution* is simply the ratio of sub-category occurence count to super-category occurence count, or a normalisation across elements within a category. This gives rise to features such as 'interactant usage versus non-interactant usage'. This directly models the fact that in using language, a choice is made. It is a choice not between one word and any other (choosing between unrelated words such as 'dog' and 'elegant'), but between semantic categories within a system. Comparative features such as these can only be used together with a sensible basis for comparison, which is provided here through the use of SFL.

4 Applications

We have applied these techniques of systemic feature extraction and representation to two separate stylistic document classification problems. Systemic functional theory has shown to be useful in classifying general documents based on their interpersonal distance [27], and the style of scientific prose [2].

4.1 Classifying financial scams

An initial application of extracting systemic features was for the detection of financial scam documents. This was as part of a larger project that included manual SFL analysis of texts. This manual analysis suggested that interpersonal distance was a key indicator for particular classes of financial scams. To test this hypothesis, we modelled PRONOMINALDETERMINATION as has been described, and used features from this

[1] Orthogonal to this is the question of ambiguity, in that "*should*" can also indicate Modalisation/probability with Orientation/subjective, depending on context. We preserve such ambiguity, and for stylistic text categorisation we rely on the fact that consistent differences in systemic preferences should still show up in the aggregate.

system to classify documents from three corpora. These corpora were chosen from three clearly separated registers: scam texts, newswire, and spoken texts.

Previous work has examined the use of the determination system in so-called 'Nigerian emails'. These are fraudulent emails in which the author attempts to establish an illegal business relationship (money transfer) with the recipient. One of the most salient characteristics of this register is the way in which the author, despite having no prior relationship with the reader, works to set up a sense of familiarity and trust. These semantic strategies suggest closer interpersonal distance than would usually be expected in the setting up of a legitimate business relationship, particularly since the texts are written rather than spoken. This corpus contained 67 manually collected Nigerian emails.

The Nigerian emails were contrasted with a collection of newspaper articles taken from the standard Reuters text classification corpus. Only texts with more than one thousand words were kept, resulting in 683 documents. It was expected that this register constructs greater interpersonal distance between author and reader.

The third register was taken from the British National Corpus and consists of 195 documents marked as belonging to the 'spoken / leisure' category. These are mostly transcriptions of interviews and radio shows covering a wide range of topics. As stated above, the interpersonal distance constructed in spoken text is almost always much closer than that constructed in written texts. Including this corpus allowed us to explore whether the perceived close interpersonal distance in the Nigerian email corpus would be confused with the close interpersonal distance that is typical of spoken texts.

These corpora differ greatly in both field and tenor, and can be separated easily using standard bag-of-words techniques. In using these corpora, we aim not to show improved performance, but to show that the determination system provides sufficient evidence to separate documents on the basis of interpersonal distance. For this to be possible, the words and categories in this system must be used in a regular and learnable fashion, which reflects the semantic positioning of the text.

Features Used

In its entirety, the system consists of 109 nodes including 48 lexical realisations. From these, various subsets were used to test the performance and robustness of the system.

all All 109 system and lexis nodes.

lexis The 48 lexical realisations in the system.

system All 61 non-lexical features.

top10 Top 10 features on the basis of information gain

top5 Top 5 features on the basis of information gain

Each set of features was computed once using term frequency (percentage of document) and again using system contribution (percentage of supersystem). Classification was performed using three different machine learners, all commonly used in text classification tasks: a Naive Bayes probabilistic classifier (NB), a decision tree (J48), and a support vector machine (SVM). All implementations are part of the publicly available WEKA machine learning package [29]. As a baseline, we used a standard bag-of-words approach using the top 500 features (ranked by information gain) represented using term frequency. Since the system contribution relies on a structured feature set, no baseline was applicable.

	#atts	NB	J48	SVM
all	109	92.8%	98.2%	98.3%
lexis	48	93.8%	98.1%	**98.4**%
system	61	93,9%	98.4%	98.3%
top10	10	96.1%	**98.6%**	97.9%
top5	5	**97.3%**	98.1%	97.8%
baseline	500	98.4%	97.5%	100%

Table 1: classification accuracy using term frequency

	#atts	NB	J48	SVM
all	109	**99.4**%	97.9%	**99.6**%
lexis	48	98.6%	**98.6**%	**99.6**%
system	61	98.6%	98.1%	99.5%
top10	10	98.9%	97.7%	98.6%
top5	5	96.2%	98.1%	98.2%

Table 2: classification accuracy using system contribution

Results

Results from using term frequency and system contribution are shown in Tables 1 and 2 respectively. All of the feature sets and classifiers produced clear separation of the classes, using only features from the determination system. The best result of 99.6% came from an SVM using the system contribution data of either all features or lexical features. It is clear from these results that these corpora are separable using features related to interpersonal distance.

Better results were achieved using system contribution than term frequency. By measuring the system choice, rather than system usage, this feature representation highlights the salient aspects of language use. This contrastive description is made possible by placing words in a system network.

In all tests, the Nigerian and Reuters corpora were clearly separated. These registers have markedly different and strongly characteristic interpersonal distance. The spoken corpus exhibited a small amount of confusion with the Nigerian texts, showing evidence that their language is more like spoken than written text.

4.2 Stylistic character of scientific prose

As a second application of SFL for stylistic text analysis, we describe some first steps towards a linguistic comparative analysis of scientific writing in experimental and historical sciences (also see [2]). Our main goal is to see if linguistic features can be found indicative of different classes of scientific articles, which may be usefully correlated with the differing methodologies of different sciences. We have studied genre variation between articles in a historical science (paleontology) and an experimental science (physical chemistry). We hypothesize that rhetorical differences between articles in the respective fields will be found that correlate with posited

differences in methodology. Use of systemically-motivated features is key in allowing this connection to be made, as we see below.

The motivation of this work is to seek empirical support for philosophers' increasing recognition that the classical model of a single "Scientific Method" (usually based on that of experimental sciences such as physics) can be a disservice to sciences such as geology and paleontology, which are historically oriented. Instead, differences in method may stem directly from the types of phenomena under study[6]. *Experimental* science works to formulate general predictive laws, and so relies heavily on repeatable series of controlled experiments which test hypotheses. *Historical* science, on the other hand, deals with contingent phenomena, involving the study of specific phenomena in the past, in an attempt to find unifying explanations for effects caused by those phenomena. Because of this, reasoning in historical sciences consists largely of reconstructive reasoning, and differs thus from the predictive reasoning from causes to possible effects characteristic of experimental science [10].

We would thus expect consistent stylistic differences should exist between scientific communication in different fields. Specifically, we hypothesise that:

- Writing in historical science should have more features expressing the weight, validity, likelihood, or typicality of different assertions or pieces of evidence
- Writing in experimental science should have more features typical of explicit reasoning about predictions and expectations.

The Corpus

We tested these hypotheses using a corpus of over 400 recent (2003) research articles drawn arbitrarily from four peer-reviewed journals in two fields: *Palaios* and *Quaternary Research* in paleontology, and *Journal of Physical Chemistry A* and *Journal of Physical Chemistry B* in physical chemistry (chosen in part for ease of electronic access). *Palaios* is a general paleontological journal, covering all areas of the field, whereas *Quaternary Research* focuses on work dealing with the quaternary period (from roughly 1.6 million years ago to the present). The two physical chemistry journals are published in tandem but have separate editorial boards and cover different subfields of physical chemistry, specifically: studies of molecules (*J. Phys Chem A*) or of materials, surfaces, and interfaces (*J. Phys Chem B*).

Features used

The systemic features we used were relative frequencies of options within three systems: CONJUNCTION, MODALITY, and COMMENT. MODALITY and COMMENT relate directly to how propositions are assessed in evidential reasoning (e.g., for likelihood, typicality, consistency with predictions, etc.), while CONJUNCTION is a primary system by which texts are constructed out of smaller pieces, and so may be expected to reflect possible differences in overall rhetorical structure. Keyword sets for leaf options were constructed by starting with the lists of typical words and phrases given by Matthiessen, and expanding them to related words and phrases taken from Roget's Interactive Thesaurus (manually filtered for relevance). These lists were constructed entirely independently of the target corpus.

	Historical		Experimental	
	P	*QR*	*PCA*	*PCB*
Palaios	–	26%	**9%**	**9%**
Quat Res	26%	–	**17%**	**14%**
Ph Ch A	**9%**	**17%**	–	32%
Ph Ch B	**9%**	**14%**	32%	–

Table 3: Average error rates for linear SMO for pairs of journals using 20-fold cross-validation.

System	Historical	Experimental
CONJUNCTION	Extension	Enhancement
COMMENT	Validative	Predictive
MODALITY/Type	Modalization	Modulation
Modalization & Manifestation	Implicit	Explicit
Modulation & Manifestation	Explicit	Implicit

Table 4: Consistently indicative features (see text) for Paleontology or Physical Chemistry.

Results

For analysis, we used the SMO learning algorithm [23] as implemented in the Weka system[28], using a linear kernel, no feature normalization, and the default parameters. (Other options did not appear to improve classification accuracy, so we used the simplest option.) Generalization accuracy for different tests was measured using 20-fold cross-validation.

We first compared inter-class and intra-class discriminability (Tab. 3). In all four cross-disciplinary cases, error rates are 17% or less, while in the two intra-disciplinary cases, accuracy is noticeably lower; *Palaios* and *Quat. Res.* are significantly less distinguishable at 26% error, while *J. Phys. Chem. A* and *J. Phys. Chem. B* are entirely undistinguishable.

We now consider if a consistent picture emerges of rhetorical differences between the two classes (paleontology and physical chemistry), based on feature weights in the learned models. We ran SMO on the entire corpus (without reserving test data) for each of the four pairs of a paleontology with a physical chemistry journal, and ranked features according to their weight for one or the other journal in each weight vector. We then consider those features strong (i.e., among the 30 with the highest absolute weights, out of all 101 features) for a single class across all journal pairs (Tab. 4).

First, in COMMENT, one opposition that emerges is between preferences for Validative comments by paleontologists and for Predictive comments by physical chemists. This linguistic opposition directly supports both our hypotheses. The historically-oriented paleontologist has a rhetorical need to explicitly delineate the scope of validity of different assertions, as part of synthetic thinking [3] about complex and ambiguous webs of past causation [6]. This is not a concern of the experimentally-oriented physical chemist, however; her main focus is prediction: the predictive strength of a theory

and its predictive consistency with the evidence.

Next, consider MODALITY. At a coarse level, we see a primary opposition in Type. The preference of the (experimental) physical chemist for Modulation (assessing what 'ought' or 'is able' to happen) is consistent with a focus on prediction and manipulation of nature. On the other hand, the (historical) paleontologist's preference for Modalization (assessing 'likelihood' or 'usuality') is consistent with the outlook of a "neutral observer" who cannot directly manipulate or replicate outcomes.

A supportive pattern is seen within features representing option pairs in MODALITY Type and Manifestation. Implicit variants are more likely to be used for options that are well-integrated into the expected rhetoric, while Explicit realizations are more likely to be used for less characteristic types of modal assessment, as more attention is drawn to them in the text. Keeping this in mind, note that Modalization is preferably Implicit in paleontology but Explicit in physical chemistry; just the reverse holds for Modulation. This shows that Modalization is integrated smoothly into the overall environment of paleontological rhetoric, and similarly Modulation is a part of the rhetorical environment of physical chemistry.

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

5 Conclusions

Systemic functional linguistics brings a wealth of theoretical resources to the study of stylistics. By considering a text as a realisation of possible options in a system network, SFL enables us to describe stylistic variation at many levels of delicacy. This provides us with a well-reasoned basis for the selection of stylistic features.

We have shown that, by explicitly modelling these systems as completely and accurately as possible, we can create features that are applicable to a range of stylistic classification problems. For this we use a partial parsing approach that enables efficiently identifying instances of system networks in unparsed text.

Individual systems describe different aspects of the meaning of a text. While in some cases a single system is sufficient to capture stylistic variation, in general it is only by modelling a set of systems and their interactions that we can build a coherent picture of language use. Note that the structural composition of each system is important; variation can occur at any level. Thus by capturing variation in systemic features, we may find useful stylistic characterisations more effectively than by examining individual lexical features or groupings.

Since systemic features are domain independent, in that text in any domain must choose to use one or another section of a system, we expect systemic features to be highly portable to other problems. Here we have described a number of useful systems taken from the SFL literature; there are many more systems that may be applicable to identifying non-ideational meaning, which future work will address.

References

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

[2] Shlomo Argamon and Jeff Dodick. Linking rhetoric and methodology in formal scientifi c writing. In *Proc. 26th Annual Meeting of the Cognitive Science Society*, August 2004. (to appear).

[3] V.R. Baker. The pragmatic routes of american qua-ternary geology and geomorphology. *Geomorphology*, 16:197–215, 1996.

[4] Bruce Biggs. *In the Beginning, The Oxford Illustrated History of New Zealand*. Oxford University Press, Oxford, 1990.

[5] Bruce Biggs. *He Whirwhiringa Selected Redaings in Maori*. Auckland University Press, Auckland, 1997.

[6] C.E. Cleland. Methodological and epistemic differ-ences between historical science and experimental science. *Philosophy of Science*, 2002.

[7] Maria Couchman. Transposing culture: A tri-stratal exploration of the meaning making of two cultures. Master's thesis, Macquarie University, 2001.

[8] S. Eggins and J. R. Martin. Genres and registers of discourse. In T. A. van Dijk, editor, *Discourse as structure and process. A multidisciplinary introduction*, Discourse studies 1, pages 230–256. London: Sage, 1997.

[9] S. Eggins, P. Wignell, and J. R. Martin. *Register analysis: theory and practice*, chapter The discourse of history: distancing the recoverable past, pages 75–109. Pinter, London, 1993.

[10] S. J. Gould. Evolution and the triumph of homology, or, why history matters. *American Scientist*, pages 60–69, 1986.

[11] M.A.K. Halliday. Corpus linguistics and probabilistic grammar. pages 30–44, 1991.

[12] Michael A. K. Halliday. *Introduction to Functional Grammar*. Edward Arnold, second edition, 1994.

[13] R. Hasan. Language in the process of socialisation: Home and school. In J. Oldenburg, Th. v Leeuwen, and L. Gerot, editors, *Language and socialisation: Home and school*. North Ryde, N.S.W.: Macquarie University, 1988.

[14] R. Hasan. *Ways of saying, ways of meaning: selected papers of Ruqaiya Hasan*. Cassell, London, 1996.

[15] J. Karlgren. *Stylistic Experiments for Information Retrieval*. PhD thesis, SICS, 2000.

[16] M. Koppel, N. Akiva, and I. Dagan. A corpus-independent feature set for style based text categorization. In *Working Notes of the IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003.

[17] C. M. I. M. Matthiessen. *Register analysis: theory and practice*, chapter Register in the round: diversity in a unifi ed theory of register, pages 221–292. Pinter, London, 1993.

[18] C. M. I. M. Matthiessen and J. A. Bateman. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press, London and New York, 1991.

[19] Christian Matthiessen. *Lexico-grammatical cartography: English systems*. International Language Sciences Publishers, 1995.

[20] A. McEnery and M. Oakes. *Authorship studies/textual statistics*, pages 234–248. Marcel Dekker, 2000.

[21] F. Mosteller and D. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Massachusetts, 1964.

[22] Michael O'Donnell. Reducing complexity in a systemic parser. In *Proceedings of the Third International Workshop on Parsing Technologies*, Tilburg, the Netherlands, August 1993.

[23] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.

[24] G. A. Plum and A. Cowling. Social constraints on grammatical variables: Tense choice in english. In Ross Steele and Terry Threadgold, editors, *Language topics. Essays in honour of Michael Halliday*. Amsterdam: Benjamins, 1987.

[25] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic text categorisation in terms of genre and author. *Computational Linguistics*, 26(4):471–495, 2001.

[26] Elke Teich. *A Proposal for Dependency in Systemic Functional Grammar – Metasemiosis in Computational Systemic Functional Linguistics*. PhD thesis, University of the Saarland and GMD/IPSI, Darmstadt, 1995.

[27] C. Whitelaw and M. Herke-Couchman. Identifying interpersonal distance using systemic features. In *Proceedings of AAAI Workshop on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

[28] I. H. Witten and Frank Eibe. Weka 3: Machine learning software in Java, 1999. `http://www.cs.waikato.ac.nz/ ml/weka`.

[29] Ian H. Witten and Frank Eibe. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[30] G. U. Yule. *Statistical Study of Literary Vocabulary*. Cambridge Univ. Press, 1944.