Chapter 16

# Identifying Interpersonal Distance using Systemic Features

**Casey Whitelaw and Jon Patrick**
*University of Sydney*
*Language Technology Research Group*
*School of Information Technologies*
*University of Sydney*
*NSW 2006 Australia*
Email: {casey, jonpat}@it.usyd.edu.au

**Maria Herke-Couchman**
*Macquarie University*
*Centre for Language in Social Life*
*Division of Linguistics and Psychology*
*Macquarie University*
*NSW 2109 Australia*
Email: maria.couchman@ling.mq.edu.au

**Abstract**

This chapter uses Systemic Functional Linguistic (SFL) theory as a basis for extracting semantic features of documents. We focus on the pronominal and determination system and the role it plays in constructing interpersonal distance. By using a hierarchical system model that represents the author's language choices, it is possible to construct a richer and more informative feature representation with superior computational efficiency than the usual bag-of-words approach. Experiments within the context of financial scam classification show that these systemic features can create clear separation between registers with different interpersonal distance. This approach is generalizable to other aspects of attitude and affect that have been modelled within the systemic functional linguistic theory.

**Keywords**: interpersonal distance, document classification, machine learning, feature representation, systemic functional linguistics, register.

## 1. Introduction

This paper explores the categorization of text based on meaning. Rather than classify on the content matter of a document, we aim to capture elements of the manner in which the document is written. In particular, we use a computational model of a part of Systemic Functional Linguistic theory to identify the interpersonal distance of a text.

Previous work has looked at extracting other semantic properties of documents. This has included the subjectivity or objectivity of whole texts (Kessler et al., 1997) or individual sentences (Wiebe, 1990; Riloff et al., 2003), and classifying reviews as positive or negative (Turney, 2002). Here, we investigate the interpersonal distance of an entire document, which partially describes the type of relationship established between author and reader.

Much of the prior research has focused on semantic categories of adjectives (Turney, 2002) and nouns (Riloff et al., 2003). This paper focuses on the closed class of pronominals and determiners. These are terms that have often been placed in stop lists, due to their frequent usage and apparent lack of relevance to classification tasks. While the use of these individual words may provide some semantic information, it is through placing them in a system of language choice that patterns of usage may be correlated with interpersonal distance.

Interpersonal distance is, briefly, the relationship established between the author / speaker and reader / listener in a text. It sets the environment in which the information is presented, and can affect the way this information is processed. This research was undertaken as part of a study into the language used in financial scams on the internet. One common characteristic of some types of scams was their ‘friendly’ and ‘casual’ manner, and it was hoped to exploit these traits within a traditional document classification task.

Systemic Functional Linguistics (SFL) provides the necessary framework to approach notions such as interpersonal distance. SFL uses multistratal analysis that encompasses both the ideational and non-ideational phenomena of a text. Attitude, affect, judgement, and subjectivity are all addressed within SF theory.

It is important to understand the fundamental bases of SFL, and its approach to describing language in general, and specifically its characterization of interpersonal distance. Section 2 includes a brief introduction to SFL theory, and explains why it is appropriate for use in this field. For SFL to be applied in practise, a suitable computational model is required. Section 3 describes such a computational representation that covers some elements of SFL, and how this can be used within a standard document classification and machine learning environment. This approach is evaluated in Section 4 through an initial series of experiments in classifying financial scams. The results confirm the usefulness of this approach and show that SFL is well-suited to identifying document-level characteristics of language use, especially the aspects of non-denotational meaning that have traditionally confounded keyword-based classification systems.

## 2. Systemic Functional Linguistics

Systemic Functional Linguistics (SFL) is a framework for describing and modelling 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 naturally occurring written and spoken text (Halliday, 1994). The theory is also *systemic* in that it models language as systems of choices (Matthiessen, 1995).

System networks have been used in SFL for more than 40 years as a way of representing the paradigmatic organization of choices within the language system (Matthiessen, 2000: 65). Initially formalized and applied by Halliday in his work on intonation (Halliday, 1963), network diagrams have been used extensively in all areas of theoretical, descriptive and applied SFL research. Systems are organized in terms of increasing *delicacy*, enabling language choice to be viewed from the most general to the most specific. The system network has served as a useful resource in computational linguistics for more than thirty years (Matthiessen 2000:66); Section 3 proposes one approach for its use in document classification.

Systemic Functional theory is a linguistic theory that describes a text in terms of the multiple meanings that it makes. While these meanings are realized by words or orthographic strings, both grammatical and lexical, a text in the first instance is viewed as a semantic unit (Halliday and Hasan, 1985). In a study such as this that seeks to categorize texts according to the meanings that they make in a systematic way, rather than just the set of words that it uses most frequently, SF theory presents itself as an extremely useful model. The proposed methods for computing aspects of SFL operate on raw text. The present research is not dependent on external semantic resources or parsers, and the text representation used is constructed only according to the specific SF meaning system under focus.

SF theory describes the use of language in context. It conceptualizes language as a multi-dimensional semiotic space showing the organization of language both globally as a meaning making system and locally as sub-systems of language use. Here we will focus on the three global dimensions that are implicated in the semantic phenomenon of interpersonal distance - the hierarchy of stratification, the spectrum of metafunction and the cline of instantiation.

### 2.1 The Hierarchy of Stratification

One key global dimension of SFL is the hierarchy of stratification. Language itself is modelled as an ordered series of levels or strata encompassing semantics, lexicogrammar and graphology / phonology, as shown in Figure 1. This in turn is modelled as being embedded stratally within context. Interpersonal distance, the phenomenon being investigated here, is located as a pattern of meaning within the semantic stratum. This pattern of meaning is realized as patterns of wording in the lexicogrammar, and it is at this level that it is exposed to current NLP techniques. The aim, then, is to recreate the semantic characterization of interpersonal distance through modelling its visible effects in the lexicogrammar of a text.

The outermost stratum shown in Figure 1 is that of context, and is frequently overlooked in NLP tasks. The social situation in which a text takes place influences and is influenced by all aspects of language choice; this is partially captured by the notion of register, as discussed below.

### 2.2 The Spectrum of Metafunction

The metafunctions refer to the three separate strands of meaning that contribute to the overall meaning in the text (Halliday, 1994). These three metafunctions are deployed simultaneously and are the textual, the interpersonal and the ideational:

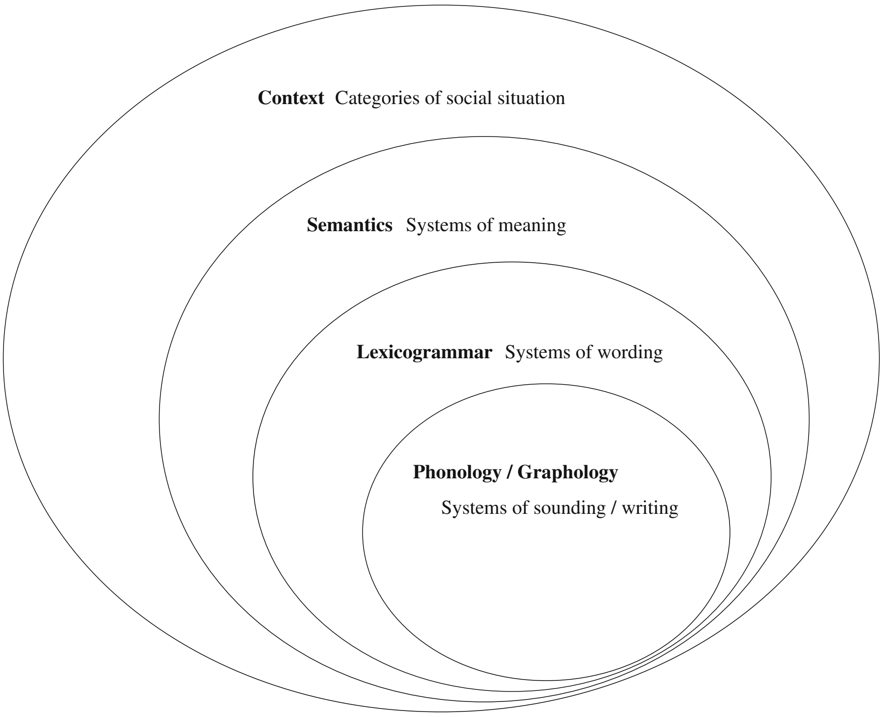

**Context** Categories of social situation
**Semantics** Systems of meaning
**Lexicogrammar** Systems of wording
**Phonology / Graphology**
Systems of sounding / writing

*Figure 1. Modelling language stratally (Hasan, 1996)*

- The textual metafunction provides ‘the resources for presenting information as text in context’ (Matthiessen, 1995).
- The interpersonal metafunction provides the resources for enacting social roles and relations as meaning.
- The ideational metafunction provides the resources for construing our experience of the world.

Interpersonal distance is located within the interpersonal metafunction and relates to the tenor of the relationship between the writer and reader within the context. Metafunction is orthogonal to the strata shown in Figure 1; the interpersonal metafunction is evidenced in context, semantics, lexicogrammar, and orthography.

### 2.3 Characterizing Registers

A register is a group of texts whose language selections vary from the general language system in similar ways. A register can be characterized by properties of its field, tenor, and mode. Registers are skewings “of probabilities relative to the general systemic probabilities” (Matthiessen, 1993). Register is the instantiation of particular situation types within the system. This characterization of registers as probabilistic is key to SFL and to this study: a register is relatively, not rigidly. This provides a formal basis for the feature representation choices made in Section 3.

Register is a realization of diversification in the context of situation (Matthiessen, 1993:234) and, in turn, register is realized by variation in meanings and wordings. While a register groups documents on the basis of the meanings they make, these meanings are realized in the semantics and lexicogrammar of the texts, and so may be analysed on these terms. In particular, registerial differences should be exposed through the patterns of language choice within a system.

A register can be described in terms of the language selections it makes within all the various language sub-systems. While the meaning of a particular text is constructed by all the selections across all the sub-systems simultaneously, examination of just a single system will still give insight into a specific type of meaning being made within the text. If this system contains characteristic variation between registers, it may be a strong enough basis for classification without further unpacking.

### 2.4 Attitude, Affect, and SFL

The systemic functional approach to language is well suited to the broader study of attitude and affect in text. Traditional areas of interest such as semantic orientation, opinion, polarity, and modality are developed, from one perspective, within the framework of Appraisal theory (Martin, 2004). Appraisal, encompassing systems such as Judgement and Appreciation, is itself only one element of the interpersonal metafunction. The study of attitude and affect must be a study of all aspects of the interpersonal; and it is through the application of these systems, both individually and as interacting elements, that a deeper understanding will be reached.

As computational techniques for identifying systemic features improve, these models of attitude and affect provide a reasoned and functional basis for classification at all levels of a text. The current level of sophistication is showing results in analysing modality (Argamon and Dodick, 2004) and appraisal (Taboada and Grieve, 2004), as well as this work on interpersonal distance.

### 2.5 Interpersonal Distance

*Interpersonal distance* is a measure of the distance being constructed by the text in the relationship between the speaker or writer and the addressee (Eggins et al., 1993). Typically, spoken discourse that unfolds in a context of maximum 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 nominalization deployed within a text as well as the frequency and type of interactant reference (Couchman, 2001).

An example of a text with very close interpersonal distance would be one that includes direct speech, such as the following (Biggs, 1990):

> *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.' Kupe went back home and said to his people, 'Prepare **my** canoe as well.' Maataa-hoorua was made ready and Kupe set off to go.*

In the above text, degree and frequency of nominalization is low and selections from the Interactant system, shown in bold face, are high.

A written history text is a good example of a text that constructs maximum interpersonal distance, partly by making no selections from within the Interactant system (Biggs, 1997):

> *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. The Southern hemisphere, moreover, has no Pole Star to provide a constant compass point.*

Work on Nigerian emails has indicated that close interpersonal distance might be characteristic of that particular register (Herke-Couchman, 2003). As mentioned above, interpersonal distance can be analysed through various systemic language choices. One key system, and that focused upon here, is the closed set of pronominals and determiners.

### 2.6 The Pronominal & Determination System

The Pronominal and Determination system (DETERMINATION) is a language system that includes within it the interpersonal resource for modelling the relationship between the interactants in the dialogue. The system is a closed grammatical system that includes realizations of both interactant (speaker, speaker-plus and addressee) and non-interactant reference items. A portion of the full DETERMINATION system is shown in Figure 2.

It is expected that very close interpersonal distance in a text would be characterized by frequent selections from the interactant system. For example, a text seeking to establish patterns of familiarity between author and reader would show foregrounded patterns of speaker (*I, me, my, mine*) and addressee (*you, your, yours*) usage. Contrastively, a text that is constructing a more formal and distant tenor will typically make little use of the interactant system but may instead show strong patterns of usage of more generalized alternative meaning systems.

The full list of terms included in this system are as follows: *my, mine, i, me, our, ours, we, us, your, yours, you, her, hers, she, his, he, him, its, it, their, theirs, they, them, one's, one, whose, who, whom, this, these, that, those, the, which, what, no, not any, no one, noone, nobody, nothing, each, each one, every, everyone, everybody, both, all.*

## 3. Representing System Networks

For systemic information to be extracted from a document, there must be a suitable computationally-feasible language model. While SFL is a comprehensive and multidimensional linguistic theory, and is not obviously computationally tractable, we can develop a more restricted model that allows us to work with specific systems such as DETERMINATION.

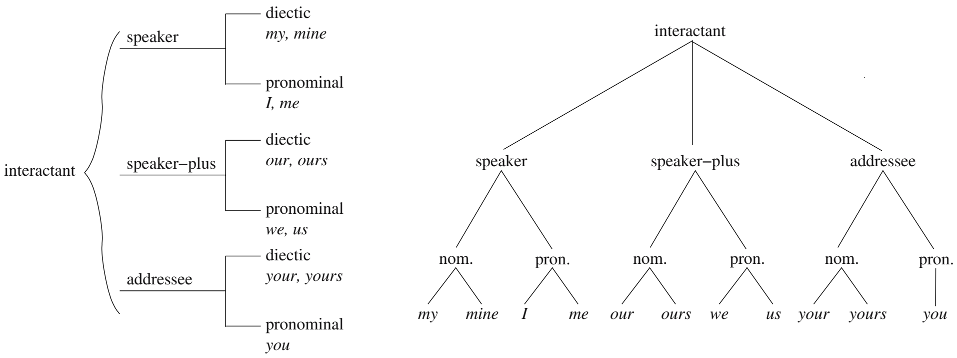


*Figure 2. The interactant subsystem modelled systemically and as a tree*

### 3.1 SFL in Computational Linguistics

Most of the computational work using systemic functional grammar has focused upon generation. The multi-stratal approach of SFG has been shown to be very effective at generating individual sentences (Mann and Matthiessen, 1985) and rhetorically linked texts using Rhetorical Structure Theory (Mann and Thompson, 1988). Functional parsing has proved more problematic in part due to the need for manual creation of broad-coverage grammars (O'Donnell, 1994).

Much less work has been done in performing automated functional analysis. Many view parsing as the fundamental basis of any text analysis, but this is not necessarily the case; machine learning techniques have been used to classify sentences by function (O'Donnell, 2002), and to automatically induce the functional properties of nominal groups (Munro, 2003). By putting to one side the complexities of full parsing, relevant aspects of SFG have been used successfully in practise.

By modelling the SFL system of DETERMINATION, the aim is to produce a model of the relevant elements of a text's meaning, and in doing so be able to efficiently classify documents based on the interpersonal distance they create. The representation used must be sufficient to capture the range of expression displayed in the system, but be amenable to use with current machine learning techniques.

The richness and reach of SF theory has meant that the linguistic analysis has typically been associated with manual qualitative text analysis. However, it is important to remember that the development of the theory has been firmly based on quantitative observations about language (Matthiessen, 2003).

### 3.2 System as Hierarchy

As is shown in Figure 2, this system can intuitively be modelled as a tree. Each internal node represents a subsystem or category: a pattern of possible language choice. Each leaf gives a realization of its parent system as a word or phrase. A system may contain both lexical realizations (leaves) and subsystems (nodes).

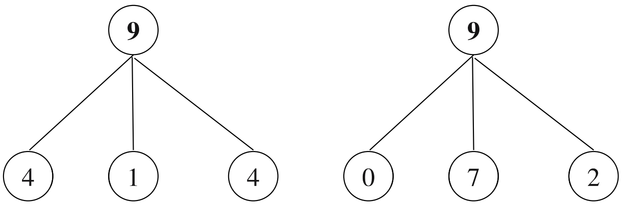


*Figure 3. Aggregating counts smoothes differences at greater delicacy*

This is an impoverished but still useful view of a system network. Language choice does not always result in a specific word or phrase; an in-depth manual analysis of a text would show that grammatical and lexical units of various sizes contribute to the overall meaning. Further, interaction between systems can result in networks that are not strictly hierarchical, and richer representations will be required to model these processes effectively. A more general computational model for extracting systemic features is proposed by Whitelaw and Argamon (2004) as an extension of this purely hierarchical approach. The current representation is sufficient to capture language choice for a system such as DETERMINATION, which is a closed class and fully lexically realized.

The usage of a system in a document can be represented by a system instance. Each occurrence of each lexical realization in the document is counted, and these counts are accumulated upwards through the network. The count at an internal node is the sum of the counts of its sub-categories. This process is no more costly than constructing a feature vector in traditional text classification methods.

### 3.3 Leveraging Systemic Structure

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 uses a language model in which all words are independent of each 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 to be gained from systemic information:

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. Figure 3 shows the raw counts of the same system in two documents; at the lower level, closer to lexis, the distributions of counts are highly dissimilar. At the higher level, these differences have been smoothed, and the documents look the same.

For a given register, it may be the case that important and characteristic language choice occurs at a very fine level, distinguishing between usages 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, that is characteristic. The usage of any words within the category may appear random while maintaining consistent category usage. These higher-level features are not available in a traditional bag-of-words approach, hence these patterns may be lost as noise.

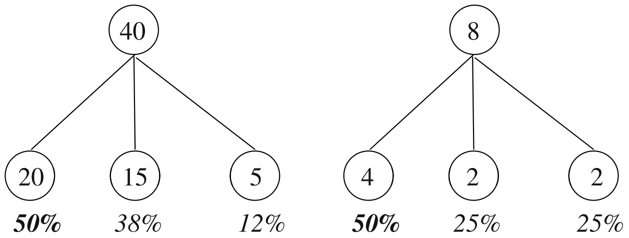


*Figure 4. Proportional features are a local and size-independent measure*

The second and more important difference to traditional feature representation is the representation of language *choice*. SF theory treats language use as a series of selections within systems; at any point in the system network, or tree as it has been modelled here, the selection is restrained to the immediate sub-systems. The choice is not between one word and any other, or even one system and any other, but a series of semantically-driven choices within the system. A bag-of-words model can model only choice between one word and any other; a choice between arbitrary words such as ‘dog’ and ‘elegant’. Comparative features such as these can only be used within an appropriate theory-driven structure, which is provided here through the use of SFL and system networks. Figure 4 shows the potential for comparative features to reveal similarities not immediately apparent in a text.

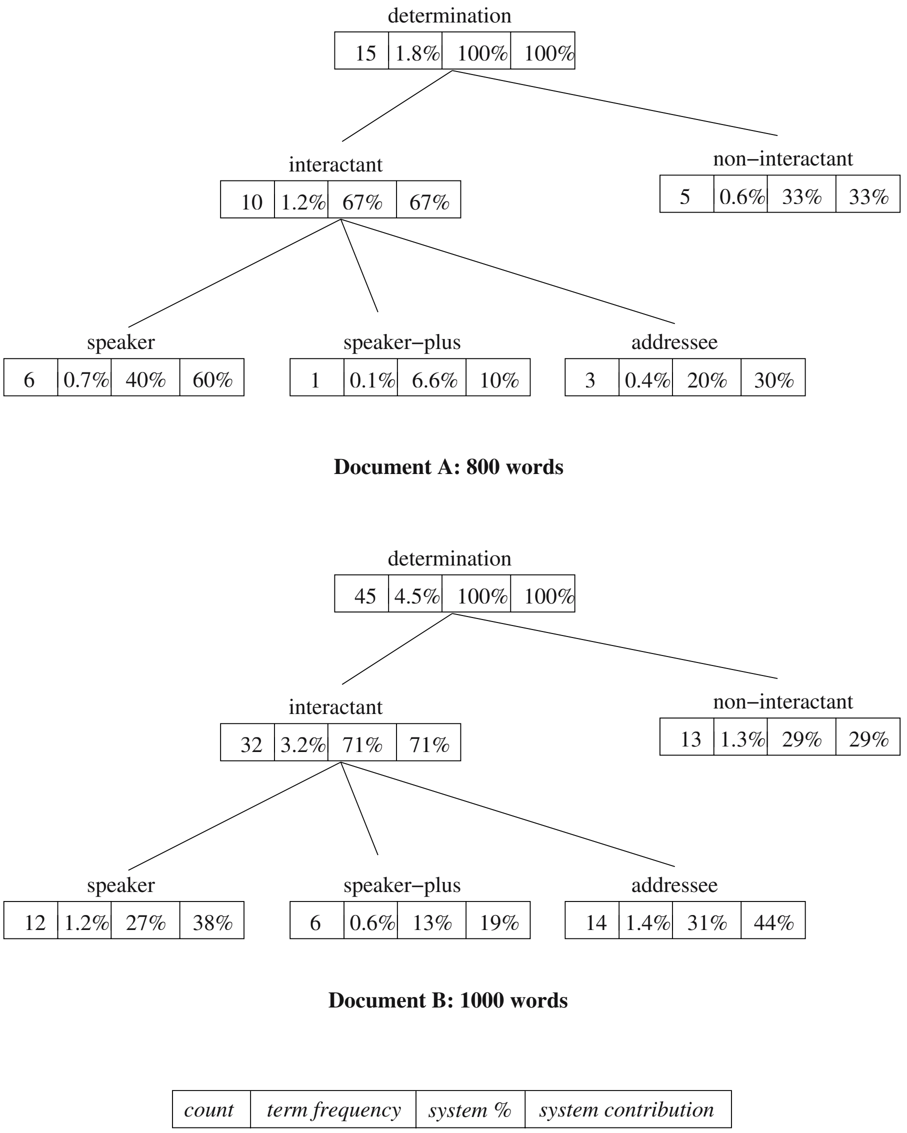
### 3.4 Representing Systemic Features

Figure 5 shows a portion of the DETERMINATION system for two documents of different sizes, belonging to the same register. Four possible feature representations are given: from left to right, each node shows the total count, term frequency, system percentage, and system contribution. Each feature representation captures a different aspect of system usage in a document and register.

**Raw counts** (first column). The summed feature count, shown in the leftmost column, presents these two documents as highly dissimilar. Note also that this is only the top portion of the system, and that multiple levels exist below those shown. Raw term counts are usually not used directly as features, as they are heavily influenced by document length.

**Term frequency** (second column) is the standard basis for bag-of-words representations; it gives the proportion of the document accounted for by this term. Term frequency is commonly used since it normalizes for document length; most topic-based document classification assumes that the document length is not important (Sebastiani, 2002). In creating features for each sub-system, this representation can still take advantage of the aggregation and smoothing provided by the system, but does not take further advantage of the known structure.

**System percentage** (third column) gives the proportion of total system usage made up by this sub-system. In Document A, *addressee* occurs three times from a total of fifteen occurrences of *determination* in the document, giving it a system percentage of 20%. Within a document, system percentage is directly proportional to term frequency, but is independent to system *density*. If another 800 words were added to Document A, but no more uses of DETERMINATION, the term frequency for a feature would halve while the system percentage remained constant. This makes it a suitable representation where distinctions are made not on how often a feature occurs, but the manner of its use. The system percentage of *speaker* is higher in Document A than Document B, despite higher term frequency in the latter.



**Document A: 800 words**

**Document B: 1000 words**

*Figure 5. Different feature representations portray a text differently*

**System contribution** (fourth column) shows the ratio of sub-system to super-system occurrence. Again in Document A, *speaker* occurs six times and its super-system, *interactant*, occurs ten times, giving a system contribution of 60%. This is a strictly local measure of usage, and captures most directly the systemic notion of choice: once the decision to use a given super-system has been made, how often was this sub-system chosen as the realization? This is a relative feature, and as such is independent of document length, total system usage, and usage of other portions of the system (see Figure 4). Despite the differences in lower-level choices, and in the raw counts of system usage, the system contribution of *interactant* in Documents A and B are very similar.

System contribution is not proportional or strongly correlated to term frequency, and the two measures provide useful and complementary information. Term frequency reports the percentage of a document that is made up of a given term. Within a system instance, term frequency can be used to report the term frequency not just of terms but of systems as well. Unlike term frequency, system contribution does not capture how often a system is used, but rather its usage in relation to the other possible choices. In the same way as a register may be characterized by choice, it may also be characterized by frequent usage of a particular system. The three complementary representations given here may each be useful in discerning characteristic system usage in general, and interpersonal distance in particular.

In implementing these representations, it is worth noting that not all system contribution features are necessary. Systems with a single child will always give 100%, and do not add information since there is no choice. In a system with a binary choice, either one of the features may be discarded since they have unit sum. Both system percentage and system contribution are meaningless at the root level, and system percentage and system contribution are identical at first level below the root. This feature reduction can be performed deterministically before any further feature selection.

By mapping only the relevant portions of a document's meaning, systemic features also have the potential to increase computational efficiency by reducing the number of attributes used in machine learning systems, in comparison to broader bag-of-words methods.

## 4. Identifying Registers

As discussed in Section 2, a register is constrained in the types of meanings it is likely to construct. A register may be characterized as establishing a certain interpersonal distance. If the choice within the determination system reflects this semantic position, it should be possible to classify documents on this basis.

Not all registers are distinguishable by interpersonal distance. This is but one of many of the semantic properties that characterize documents, such as formality, modality, and evaluation. Note also that the identification of a register is not the same as identifying the *topic* of a document; instances of the 'newspaper article' register may have very different content that is all presented in the same fashion.

### 4.1 Corpora

We chose corpora that were clearly separated into different registers. From prior manual analysis, it was expected that these registers would have different characteristic interpersonal distance.

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. Since many of the newswire texts are very short, only texts with more than one thousand words were kept, resulting in 683 documents. As a result of the context in which they unfold, it was expected that the Reuters newswire texts would make different language choices in order to realize the different meanings they construct. More specifically, it is expected that this register constructs a greater and more formal 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. The systemic organization proposed by SFL is only one possible structure; if this is a sensible semantic description of language use, as SFL asserts it to be, the resulting systemic features should be useful in classification.

### 4.2 Features Used

The behaviour of a system within a document can be represented as a system instance. As discussed in Section 3, a system instance stores hierarchical information at every level from the full system to individual lexical realizations. System usage may differ at any or all of these levels: some registers may make very specific lexical choices, while others may be differentiable by more general trends. In its entirety, the determination system consists of 109 nodes including 48 lexical realizations. 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 realizations 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

| | #attributes | 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 1. Classification accuracy using system contribution*

| | #attributes | 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 | 109 | 98.4% | 97.5% | 100% |

*Table 2. Classification accuracy using term frequency*

Each set of features was computed once using term frequency and again using system contribution. 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 (Witten and Eibe, 1999).

### 4.3 Results

Results from using system contribution and term frequency 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 the use of 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.

Feature selection exhibits different effects on the two types of features used. Best performance for system contribution features came from using all features, or only lexical features. Best performance for term frequency features, however, came from using fewer features. Since there is a high degree of correlation between term frequencies within a system network, this can skew results when using classifiers that assume independent features, as Naive Bayes does.

| system | term | bag-of-words |
|---|---|---|
| addressee / pronominal | addressee / nominal | *your* |
| addressee / nominal | *your* | *my* |
| speaker / pronominal | *my* | *you* |
| speaker / nominal | interactant | *me* |
| *I* | addressee | *said* |
| interactant | speaker / nominal | *I* |
| Non-interactant | addressee / pronominal | *er* |
| *your* | *you* | *that's* |
| *me* | *I* | *erm* |
| *my* | speaker / pronominal | *us* |

*Table 3. Top ten features show pronominals and speech markers*

Table 3 shows the top 10 features as ranked using the information gain metric (Quinlan, 1993). For systemic features, almost all are located within the *interactant* subsystem. This is further confirmation that the discerning features are not random discrepancies between classes, but are evidence of the underlying semantic intent. Also shown are the most significant features in the bag-of-words approach. Despite being informed by all the words in the documents, the most significant were still those located in the determination system, together with transcribed discourse markers such as 'er' and 'erm', which were of use in separating the spoken texts of the BNC documents.

## 5. Conclusion

SFL is fundamentally a theory of meaning. As such, language choices can be identified as both formal lexical or grammatical selections as well as in terms of systemic meaning selections. The relationship between these two complementary perspectives is one of abstraction or generalization; a meaning system is more abstract than the grammar or lexis that realizes it (Martin and Rose, 2003). This realization ensures that a meaning phenomenon such as interpersonal distance is characterizable in terms of both systemic choice and lexicogrammatical structure.

In this paper, we have shown that one aspect of the interpersonal distance of a document can be characterized by the use of the determination system. We have further shown that registers that construct variable interpersonal meaning can be separated solely using the features from the Pronominal and Determination system. This can be achieved by modelling SFL at the lexical level without specific external resources.

Interpersonal distance is but one property of the tenor of a document. Similarly, the determination system is but one small part of SFL theory. As our ability to computationally model and extract system networks increases, these systems and their interactions will provide more features by which the semantic properties of a document may be discerned.

## 6. Bibliography

Argamon, S. and Dodick, J. *Conjunction and Modal Assessment in Genre Classification: A Corpus-Based Study of Historical and Experimental Science Writing*. In this volume, Shanahan J. G., Qu Y., Wiebe J. (Eds.) *Computing Attitude and Affect in Text*. Springer, Berlin.

Biggs, B. (1990) *In the Beginning, The Oxford Illustrated History of New Zealand*. Oxford University Press, Oxford.

Biggs, B. (1997) *He Whirwhiringa Selected Readings in Maori*. Auckland University Press, Auckland.

Couchman, M. (2001) *Transposing culture: A tri-stratal exploration of the meaning making of two cultures*. Honours thesis, Macquarie University.

Eggins, S., Wignell, P. and Martin, J. R. (1993) *Register analysis: theory and practice*. In *The discourse of history: distancing the recoverable past*, 75-109, Pinter, London.

Halliday, M. A. K. and Hasan, R. (1985) *Language, Context and Text: a social semiotic perspective*. Geelong, Victoria: Deakin University Press.

Halliday, M. A. K. (1994) *Introduction to Functional Grammar*. Edward Arnold, second edition.

Halliday, M. A. K. (1995) Computing Meaning: some reflections on past experience and present prospects. Paper presented to PACLING95, Brisbane, April, 1995.

Hasan, R. (1996) *Ways of saying, ways of meaning: selected papers of Ruqaiya Hasan.* Cassell, London.

Herke-Couchman, M. A. (2003) *Arresting the scams: Using systemic functional theory to solve a hi-tech social problem*. In Australian SFL Association Conference 2003.

Iedema, R., Feez, S. and White, P. (1995) *Media literacy*. Sydney, Metropolitan East Disadvantaged Schools Program.

Kessler, B., Nunberg, G., and Shutze, H. (1997) *Automatic detection of text genre*. In Philip R. Cohen and Wolfgang Wahlster (Eds), *Proceedings of the Thirty-Fifth Annual Meeting of the ACL and Eigth Conference of the EACL*, 32-38, Somerset, New Jersey.

Mann, W. C. and Matthiessen, C. M. I. M. (1985). *Nigel: A systemic grammar for text generation.* In Benson, R. and Greaves, J. (Eds), *Systemic Perspectives on Discourse: Selected Papers from the 9th International Systemic Workshop*. Ablex.

Mann, W. C. and Thompson, S. A. (1988) *Rhetorical structure theory: Toward a functional theory of text organisation.* Text, 8 (3), 243-281.

Martin, J. R. and Rose, D. (2003). *Working with Discourse: Meaning Beyond the Clause*. Continuum, London and New York.

Martin, J. R. (2004) *Mourning: how we get aligned.* In Discourse & Society 15.2/3 (Special Issue on 'Discourse around 9/11'). 321-344.

Matthiessen, C. M. I. M (1993) *Register analysis: theory and practice*. In *Register in the round: diversity in a unified theory of register*. 221-292. Pinter, London.

Matthiessen, C. M. I. M (1995) *Lexico-grammatical cartography: English systems.* International Language Sciences Publishers.

Matthiessen, C. M. I. M. (2003). *Frequency Profiles of some basic grammatical systems: an interim report*. Macquarie University.

Munro, R. (2003) *Towards a computational inference and application of a functional grammar,* Honours thesis, University of Sydney, 2003.

O'Donnell, M. (1994) *Sentence analysis and generation: a systemic perspective*. PhD thesis, University of Sydney.

O'Donnell, M. (2002) *Automating the coding of semantic patterns: applying machine learning to corpus linguistics*. In *Proceedings of the 29th International Systemic Functional Workshop*. University of Liverpool.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Riloff, E., Wiebe, J. and Wilson, T. (2003), *Learning subjective nouns using extraction pattern bootstrapping*. In *Proceedings of CoNLL-2003*, 25-32. Edmonton, Canada.

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

Sebastiani, F. (2004). Text Categorization. In Alessandro Zanasi (Ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK.

Taboada, M. and Grieve, J. *Analysing Appraisal Automatically.* In Shanahan J. G., Qu Y., Wiebe J. (Eds.) *Computing Attitude and Affect in Text*. Springer, Berlin.

Turney, P. (2002) *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.* In *Proceedings 40th Annual Meeting of the ACL (ACL'02)*, 417-424

Wiebe, J. (1990) *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text.* PhD thesis, State University of New York at Buffalo.

Witten, I.H. and Eibe, F. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.

Whitelaw, C. and Argamon, S. (2004) *Systemic Functional Features for Stylistic Text Categorization*. In *Proceedings of the AAAI 2004 Fall Symposium on Style and Meaning in Language, Art, Music, and Design.* AAAI Press.

Wu, C. (2000) *Modelling Linguistic Resources: A Systemic Functional Approach.* Unpublished PhD thesis, Macquarie University, Sydney.