Stylistic Features as Meaning Representation:
Text as Phase Portrait

Maria Herke-Couchman and Canzhong Wu

Capital Markets Co-operative Research Centre
Centre for Language in Social Life
Macquarie University
Sydney, Australia
mcouchman@ling.mq.edu.au   canzhong.wu@ling.mq.edu.au

Abstract

Typically, high volume text processing has been based on representations of the texts such as Parts of Speech (POS), parsed formalizations or bag of words. This paper proposes an alternative representation of the text based on linguistic analysis. Drawing on systemic functional theory and its application to stylistics, we will develop a text 'phase portrait' comprised of the patterns of systemic features that are foregrounded in each text. We will only focus on computationally tractable features to ensure that textual phase portraits are easily identifiable by machine learner technology, a method that has proved successful in previous similar research (Herke-Couchman & Whitelaw 2003).

Introduction

From the evolving research field of text categorization, Sebastiani (2004) draws attention to the pursuit of alternative and additional methods of text representation:

> ... while the bag of words model is still the unsurpassed text representation model, researchers have not renounced the belief that a text must be something more than a mere collection of tokens, and that the quest for models more sophisticated than the bag of words model is still worth pursuing.

The present paper will propose such an alternative representation. This research is guided generally by systemic functional theory, and influenced by the work of Halliday (1995), Matthiessen (1995) and Butt & O'Toole (2003), in which text is conceived of as a semantic event and a mobilization of the resources of meaning.

Using stylistic features, we will identify texts in terms of 'phase portraits' (Butt & O'Toole 2003). These phase portraits are representations of texts constructed from the systemic choices revealed by linguistic analysis. Previous research (Koppel, Argamon and Shimoni 2003; Herke-Couchman and Whitelaw 2003) has demonstrated the success of document identification based on various systemic features. The current research will expand on Herke-Couchman and Whitelaw (2003) by exploring not only the occurrence and frequency of systemic features, but also the patterns of collocation of these systems with features from a limited number of other systems. In other words, we will explore the meaning that collocating systems contribute to a sample of text types. These meanings will then be mapped as phase portraits that will be both characteristic of the texts as well as being computationally tractable.

This research uses computational tools for both high-level manual analysis and low-level automatic analysis to locate a limited set of systemic language patterns. In the first instance, we use features that are easily extracted computationally. When computational identification of the systemic features presents a challenge beyond orthographic matching, we either use or point to readily available computational tools that, in a high volume application, could replace the manual component of this research. It is a goal of the research that the phase portraits be easily transportable to machine learners where the methodology can be usefully employed in high-volume document classification. The machine-learning component, which falls outside of the scope of the present paper, is a topic for further research.

Systemic functional theory

Systemic functional theory originated in the tradition of European linguistics that developed following the work of Saussure, grew out of J.R. Firth's work, and was further developed by M.A.K. Halliday, and has now been developed into a theory of language in context centred around the notion of language function. This theory is semantics-oriented rather than syntax-oriented, functional rather than formal, and its object of study is thus the text rather than the sentence as in formal linguistics.

The global dimensions

In the systemic interpretation, language is conceptualised as a multi-dimensional space for making and expressing meanings in context (Halliday 1978 & 1991; Matthiessen 1995; Halliday and Matthiessen 1999; Wu 2000). It provides the potential for mapping out languages in general terms. This space is organized along a number of interrelated dimensions: stratification, metafunction and instantiation (see Figure 1).
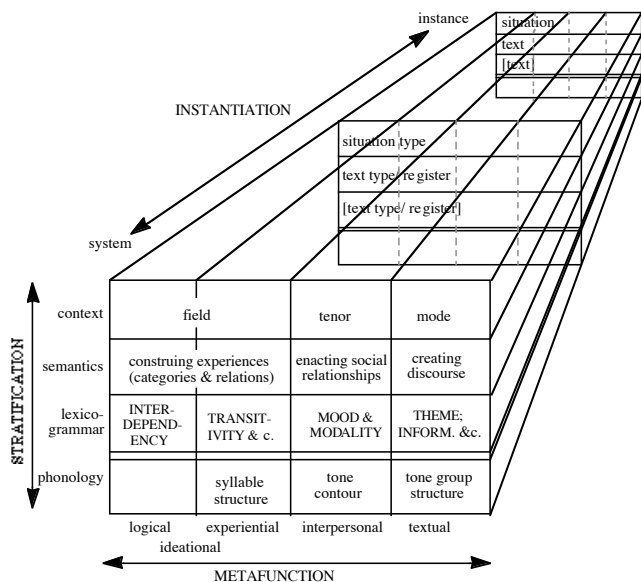
Figure 1 Multidimensional space for making meanings

The dimension of **stratification** orders the space into four strata or levels: context, semantics, lexicogrammar and phonology/ graphology, with the last three concerned with language. These strata are related by means of realization: each higher stratum constitutes the semiotic environment for the immediately lower one, and a lower stratum realizes the immediately higher one. Context is realized by semantics, which is realized by lexicogrammar, which is, in turn, realized by phonology or graphology. Semantics, the highest stratum of language, interfaces with our experience of the world whereas phonology, the lowest stratum, interfaces with the human body.

The dimension of **metafunction** divides the space into three simultaneous but functionally diversified modes of meanings: ideational, interpersonal and textual. The three functions (often called metafunctions) are equally important, each contributing to one strand of meaning. The ideational function is concerned with construing our experience of the world around and inside us. There are two ideational subtypes: experiential and logical. The former construes experience in terms of configurations, such as the configurations of processes, participants and circumstances while the latter construes experience as chains through highly generalized relations such as modification and exemplification. The interpersonal function is concerned with enacting social relations between the speaker and the listener, and the textual function with creating 'text' and constructing ideational and interpersonal meanings into information that is relevant to context.

The cline of **instantiation** relates the system or potential to the text or instance. The system is language seen as semiotic potential from a distance, and the text is language as instances from close up (Halliday 1995). As observers of language, we can locate ourselves at any point along the cline of instantiation. We can position ourselves at the instance end, and work on the text, e.g., translating one particular text into another, and extracting information from text. We can position ourselves at the potential end, and focus on the system, e.g., compiling thesauruses and dictionaries, and developing reference grammars. We can also position ourselves at an intermediate point along this cline, moving in from one end or from the other.

However, this does not mean that we are constrained to a particular region between text and system along the cline of instantiation. Rather we have a degree of freedom to move around and draw on the resources along the way. We can investigate a text, and how a particular linguistic system is instantiated in this particular text; we can also go beyond the single text, and look for recurrent patterns in sets of texts in order to categorize text types or the overall linguistic system. This is a typical methodology in corpus linguistics, and a common approach to domain modelling in computational linguistics.

It is often the case that we simply concentrate on the system or the text along the cline of instantiation, but it is sometimes necessary to keep both system and text in focus. In multilingual work, the process of translation takes place at the instance end, but relies heavily on systemic potentials to do the multilingual mappings. In corpus linguistics, the process of tagging is also carried out at the instance end, but the tagging rules are part of the linguistic resources located at the potential end. Similarly, text categorization draws heavily on systemic potentials when a document is classified.

The local, intra-stratal dimensions

While the global organization of language is determined by stratification, metafunction and instantiation, the internal organization of each stratal region is determined by axis, rank and delicacy (see Figure 2).
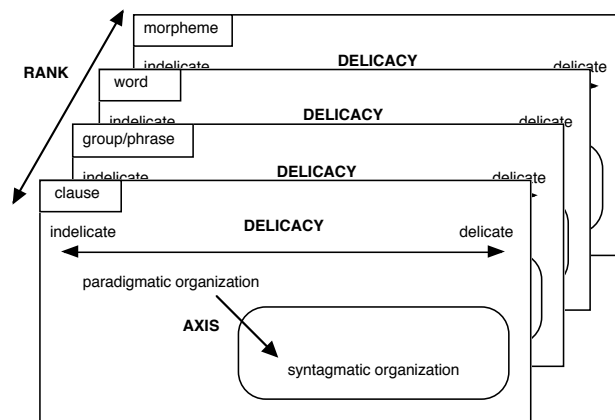

Figure 2 Local, stratal organization

The dimension of stratification divides language into a number of strata, each of which is organized locally into a hierarchy of **ranks**. The highest rank consists of units of the next rank below, each of which in turn consists of units of the next rank below, and so on until the lowest rank is reached where there is no internal constituency structure. In English as well as many other languages, the rank scale of the lexicogrammatical stratum is clause—group/phrase—word—morpheme, and that of phonology is tone group—foot—syllable—phoneme.

Each rank is organized internally along two **axes** of organization: systemic or paradigmatic axis and structural or syntagmatic axis. The paradigmatic organization is represented by system networks (networks of interrelated systems of options), and the syntagmatic organization is represented by function structures (configurations of constituent functions). The paradigmatic organization is primary and the syntagmatic organization is secondary; syntagmatic relations are derivable from paradigmatic selections by means of realization statements.

The paradigmatic options are ordered in **delicacy** from more general systems to more specific ones, and the systems are organized into a system of systems or a system network. A system is a set of options or features together with an entry condition, such that if the entry condition is satisfied, one option or feature must be chosen from the system. As shown in Figure 3, given the entry condition *clause*, a feature must be chosen from the system of PROCESS TYPE (existential / relational / verbal / mental / behavioural / material) and a feature from the system of AGENCY (effective / middle) since there are two simultaneous systems here.
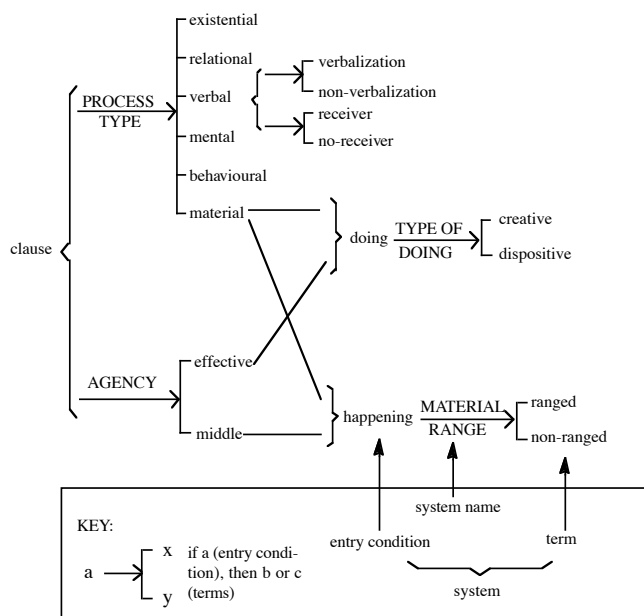

Figure 3 Fragment of a system network

Stylistics

Text categorization has typically relied on classification according to the content words of the text in question (Koppel, Argamon and Shimoni 2003). However, as systemic functional theory makes clear, the meaning of a text is constructed by a large number of lexical, grammatical, semantic and contextual systems, of which the informational or experiential meanings (content words) are just one component. These meanings are layered in a text and, as such, occur simultaneously. As a consequence, the mappings and the tools that systemic functional linguistics provides have proved extremely useful to stylistic studies for many years (Butt 1984; Hasan 1984; Couchman 2001).

As Figure 1 above indicates, the areas in which these meaning systems are located can be conceptualized as a three dimensional grid, the axes of which are the dimensions of instantiation, stratification and metafunction (Halliday, 1995). Each of the areas on the grid is representative of a number of systems, many with explicit formal realizations, which means that once specified, they should be easily identifiable computationally. So while the layerings and multi-dimensionality of systemic functional theory might suggest that the task of operationalizing the theory computationally is a daunting one, the explicit realizations that point to the meaning systems implicated in a given text ensure that, provided the mappings are followed, the task is manageable.

Methodology and Data

This research makes use of a suite of computational tools developed by Wu and Matthiessen at Macquarie University. Among these tools, SysAm is a database system for supporting manual linguistic analysis from the perspective of systemic functional linguistics, and for producing statistical results of the analysis, while SysConc enables the user to organize the search items into a feature-based network, and quickly extract lexical patterns, including systemic sets of lexis. SysConc allows the analyst at-a-glance insights into individual texts of variable length on the one hand or entire corpora on the other (Wu, 2000).

The use of both SysAm and SysConc in this research has enabled us to not only process much larger volumes of texts than would otherwise have been possible by manual analysis, it has also ensured that the features we have used have been computationally identifiable.

Focus systems

In this research, we will focus on just a small selection of all the possible systems mapped onto the systemic meaning space. As each of the systems that we will be working with

are introduced, we will pin-point the system's semiotic address in terms of Figure 1 above (see also Halliday 1994; Matthiessen 1995).

Within each of these systems, we will look at the 20 most frequently occurring features and, while we make no statistical claims, we will construct textual phase portraits using relative average frequencies.

INDICATIVE MOOD PERSON

This system is located on the stratum of lexicogrammar and within the interpersonal metafunction. It is selected for indicative clauses, and may determine the mood features of person and number. It has an entry condition of 'indicative', and two features of 'interactant' (realized by Subject: *you*, *we* and *I*) and 'non-interactant' (realized by Subject: third person of traditional grammar such as *she*, *he*, *they*, etc.)

Herke-Couchman & Whitelaw (2003) used this system as a measure of interpersonal distance and successfully classified documents based on these features. As a starting point for the present research, we have taken just two features from this system used in the previous work, the pronominals *he* and *she*.

We will then use the Key Word in Context (KWIC) component of SysConc and identify all occurrences of 'he' and 'she' as well as the collocating item immediately to the left (L+). The 20 most frequently occurring collocations for each pronominal within each text will then be analysed generally in terms of the following systems.

PROCESS TYPE

This system is one of the most general systems in the system of TRANSITIVITY, and is located on the stratum of lexicogrammar and within the experiential metafunction. It has the entry condition of 'clause', and six systemic features of 'material', 'behavioural', 'mental', 'verbal', 'relational' and 'existential' according to which we will classify the process types (or verbs) of the clauses that contain the search items. This classification will be performed by SysVerb, a component of SysAm. SysVerb cross-matches verbs with both English verb classes (Levin 1993) and Systemic verb categories. It is the Systemic categories of 'material', 'behavioural', 'verbal', 'mental' and 'relational' that will be used in this investigation.

MODALITY

The system of MODALITY is about a speaker's judgement of the probabilities or obligations in his or her proposal or proposition. It is located on the stratum of lexicogrammar and within the interpersonal metafunction. In this research, we will look at the degree to which each Non-interactant (*she* or *he*) collocates with a modal element, whether it be a modal finite such as 'could' or a modal adjunct such as 'never'.

TENSE

The system of TENSE represents the relationship between the present and the time of the occurrence of an event. It is located on the stratum of lexicogrammar and within the interpersonal metafunction.

In this research the tense of each process type will be explored. Future tense is not easily determined by these computational methods since it is easily confused with modal finites. However, as ambiguity contributes to the meaning of a text by suggesting that both meanings may be possible, all modal finites that could be interpreted as future tense will be recounted as such in the study of the tense system.

EXPANSION

Occasionally, the collocating L+ item will be a conjunctive of expansion and these too will be taken into account. While expansion typically refers to the relationship between clauses in a clause complex, relationships of expansion may also be found further down the rank scale linking groups. These relationships may be realised implicitly or explicitly by conjunctions. It is these explicit conjunctions that the present research will explore. It is important to note that, in the present research, no differentiation is made as to the rank at which the identified expansion is realised.

Data

To explore text as phase portraits, we have selected as our corpus a few texts that are freely available in electronic format. The corpus includes a poem ('The Man from Snowy River' by Banjo Patterson), a novel ('Wuthering Heights' by Emily Bronte) and a small collection of obituaries. In addition, we have used a corpus of horoscopes as well as a small subset of the Reuters news-wire corpus. However the use of both of these was restricted to the initial exploration of the system of INDICATIVE MOOD PERSON.

Results and Analysis

The results are reported according to the systems that were explored.

Indicative Mood Person: While the degree to which this system differentiates the texts is not the focus of this research, the system is nonetheless implicated since it is the collocations of one part of this system that are being explored.

As Figure 4 indicates, in four of the sample texts or text types, *he* occurs more frequently than *she*. In the horoscope corpus there are zero occurrences of either *he* or *she*.
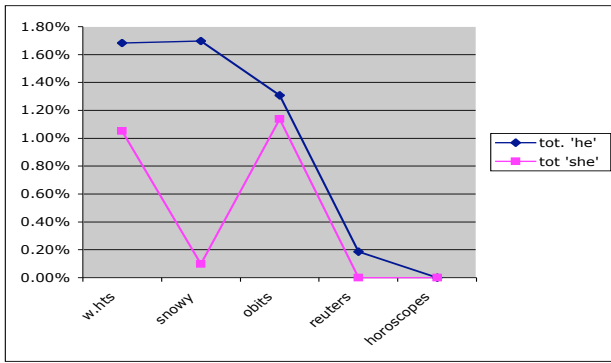
Figure 4 Indicative Mood Person

Figure 5 summarizes the selection frequencies of the L+ collocations in terms of the three systems of process type, expansion and modality.
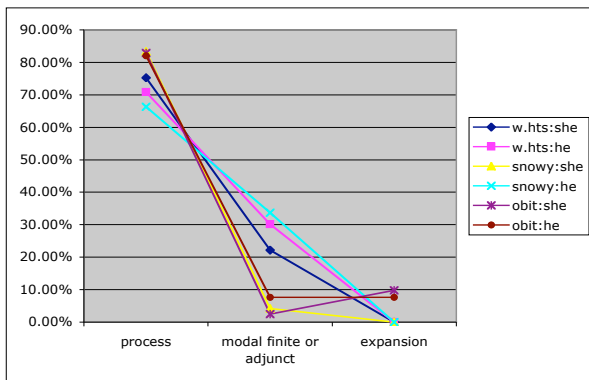

Figure 5 Collocating systems (1)

Process type, Modality, Expansion and Tense: There are a number of points to note about the results reported here. Firstly, because this research examines only the first left collocational element, we are unable to analyze complex verbal groups, should they occur. Categorization of the process type is restricted to the first element of the verbal group. It may be possible to overcome this problem by POS tagging the texts so that the entire collocating verbal group could be identified.

Secondly, while the present research has categorized verbs in terms of their most general systemic features, using a tool such as SysVerb (a database system that organizes English verbs according to systemic functional categories), it is actually possible to categorize in terms of much more delicate distinctions. Nonetheless, as indicates in Figure 6, selection of process type as well as frequency of both expansion and modality usage as well as tense selections do vary across the texts.

Coupled with the above findings, the remainder of the findings comprise the phase portraits for each text. For the sake of comparison, these have been overlayed (see Figure 6 below).
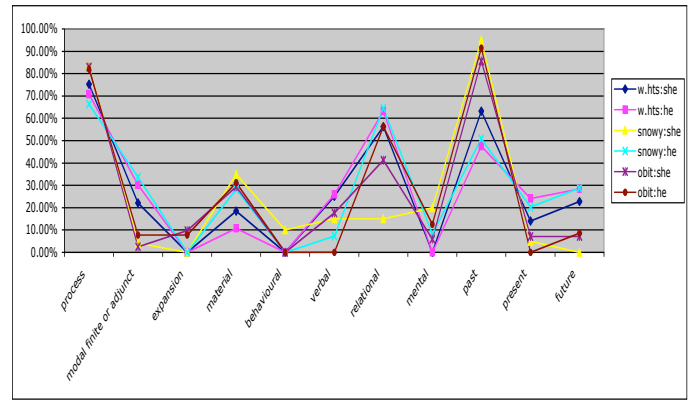

Figure 6 Collocating systems (2)

Interpretation

According to Manning and Schutze (1999: 575-579) the traditional approach to text categorization from computational linguistics seeks to classify texts based on a representation of the text that is exclusive of many grammatical items. In other words, this 'bag of words' approach classifies texts according to a text representation comprised almost entirely of content words. While this method has been shown to yield reasonably successful text categorisation within a limited domain, the success rate falls as the range of domains from which the texts are drawn expands.

In interpreting the present research, it is useful to first locate the bag of words method of text categorization in terms of both the global and local dimensions of SF theory.

Stratally, a bag of words representation is located in the lexicogrammar. Metafunctionally, since this representation foregrounds the construing of experience in the text, it is located primarily within the experiential metafunction (see Figure 1).

In terms of the local dimensions, we can sharpen our focus on the lexicogrammatical stratum. In doing so, two points become clear.

The first point is that, while the presence of adjectives represents a very limited coverage of the interpersonal metafunction, a bag of words representation is comprised largely of nouns and verbs and so is essentially metafunctionally restricted to just the experiential metafunction.

The second point is that even this coverage of the experiential metafunction is impoverished since its reliance on single lexical items means that it often gives only clues as to the participants and actions in the texts. The text coverage that a bag of words approach offers in terms of the local dimensions of function and rank is indicated by solid blue ovals in Figure 7 below.

rank	[class]		ideational		interpersonal	textual	
			logical	experiential			(cohesive)
clause	clause		INTER-DEPENDENCY	TRANSITIVITY	MOOD / MODALITY / POLARITY	THEME / CULMINATION / VOICE	SUBSTITUTION / ELLIPSIS / REFERENCE (CONJUNCTION)
group/phrase	phrase	prep.		MINOR TRANSITIVITY	MINOR MOOD	CONJUNCTION	
	group	nominal	MODIFICATION	THING TYPE / CLASSIFICATION / EPITHESIS / QUALIFICATION	PERSON / ATTITUDE	DETERMINATION	
		verbal	TENSE	ASPECT / EVENT TYPE	FINITENESS / DEICTICITY	VOICE	
		adverbial	MODIFICATION	QUALITY	COMMENT	CONJUNCTION	
word			DERIVATION	DENOTATION	CONNOTATION		
			INFORMATION TAXIS	ACCENTUATION	KEY	INFORMATION FOCUS	
			complexes	simplexes			

LOGICO-SEMANTIC RELATION

Figure 7 Location of 'bag of words' and 'phase portraits' within SF dimensions

While the systems of THING TYPE and EVENT TYPE have been highlighted, as mentioned above, the bag of words focus on single lexical items or words means that, in the case of multi-word items, we are unable extract either complete thing type or complete event type information.

We will now switch our focus from the bag of words model to the phase portrait model explored in the present paper. By locating this exploration in terms of the global and local dimensions of systemic functional (SF) theory, we will better understand the contribution of our proposal.

As Figure 7 above indicates (see dotted red ovals), the present research offers a more extensive coverage of the local lexicogrammatical systems of language than a bag of words approach offers.

 In addition to offering a more comprehensive coverage of systems already represented in the experiential metafunction by the bag of words approach, an SFL phase portrait representation extends on this by representing texts in terms of systemic features from both the interpersonal (modality, person) and textual (conjunction) metafunctions.

Further, in this research we have characterized texts in terms of both systemic stylistic features as well as in terms of combinations of these systems.

In terms of the global dimensions, while this research is still largely situated on the stratum of lexicogrammar, representing the texts in terms of meaningful groupings of

lexical and grammatical realisations allows us to push up onto the semantic stratum and gain insights into not just grammatical and lexical differences, but also differences in the meanings constructed across different texts.

The system of expansion is particularly important in this regard; explicit conjunction at the strata of lexicogrammar gives us insights into the rhetorical structure of a text. The research here has only been a cursory overview of a very interesting and promising area of exploration into computing semantics using tools such as Rhetorical Structure Theory (Matthiessen, 2000).

In addition to gaining insights into both the semantic stratum and how future research might more comprehensively compute meaning in texts at this stratum, we have also achieved insights into the context. By constructing textual phase portraits for texts that have unfolded in different situations, we can begin to compare the meanings that specific situations bring to bear on a text. This too, opens up a very large area for future computational research.

Conclusion

By using SF theory as a navigation tool, the present research has attempted to offer a representation of text that not only locates the traditional approach to text categorization in terms of linguistic theory, but has also

sought to expand on this representation offering richer, more meaningful and computationally operationalizable text representations.

Returning to Sebastiani's quotation above, we have not so much offered an alternative text representation as built upon the traditional method to which he refers. We have shown that a bag of words is in fact a more meaningful representation than a 'mere collection of tokens'. By locating it within SF theory, we are able to see that, while limited, the approach does offer us insights into the experiential aspects of a text's meaning.

More importantly though, we have demonstrated that SF theory can be used to navigate through the meaning space of language by shunting between the global and local dimensions. In this research, we have identified some of the explicit and, therefore, computationally accessible language systems. We have shown that different texts make different meanings in terms of these systems and we have suggested that Systemic Functional phase portraits as text representations have much to offer high volume text categorization across varied domains.

Finally, it is important to note that in locating both our own work as well as previous work in terms of the Systemic Functional dimensions, the direction for future work is clearly indicated. From the perspective of both local and global dimensions, the areas that have received little or no attention from computational linguists become very clear and it is to these areas of language that future research can be directed.

References

Argamon, S. and J. Dodick. 2004. Conjunction and Modal Assessment in Genre Classification: A Corpus-Based Study of Historical and Experimental Science Writing. Exploring Attitude and Affect in Text: Theories and Applications. Working Notes AAAI-2004 Spring Symposium Series.

Butt, D. G. 1984. Perceiving as making in the poetry of Wallace Stevens. *Nottingham Linguistic Circular* 13: 124-146.

Butt, D. and M. O'Toole. 2003. Transactions between Matter and Meaning: A Functional Theory for the Science of Text. In M. Amano (ed.) *Creation and Practical Use of Language Texts*. Proceedings of the Second International Conference Studies for the Integrated Text Science, Graduate School of Letters, Nagoya University.

Couchman, M.A. 2001. *Transposing Culture: A tri-stratal Exploration of the Meaning Making of Two Cultures.* MA Honours thesis, Dept of Linguistics, Macquarie University, Sydney.

Halliday, M.A.K. 1978. Language as a social semiotic: the social interpretation of language and meaning. London: Edward Arnold.

Halliday, M.A.K. 1991. *Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M.A.K. 1991. Towards Probabilistic Interpretations. In E. Ventola (ed.), *Functional and systemic linguistics: approaches and uses*. The Hague: Mouton.

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

Halliday, M.A.K. 2002 Lexis as a Linguistic Level 1966. In J. J. Webster (Ed.), *On Grammar* M.A.K.Halliday (Vol. 1). London & New York: Continuum.

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

Halliday, M.A.K. & C. Matthiessen. 1999. Construing experience through meaning: a language-based approach to cognition. London: Cassell.

Hasan, R. 1985. *Linguistics, language and verbal art*. Geelong, Vic., Deakin University Press.

Herke-Couchman, M. and C. Whitelaw. 2003. Identifying Interpersonal Distance Using Systemic Features. Australasian Language Technology Workshop, Melbourne, Australasian Language Technology Association.

Koppel, M., J. Argamon and A. Shimoni. 2003. Gender, Genre, and Writing Style in Formal Written Text. Text **23**(3): 321-346.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: University of Chicago Press.

Matthiessen, C. 1995. *Lexicogrammatical Cartography: English Systems*, Tokyo: International Language Sciences Publishers.

Matthiessen, C. (ed.) 2000. The Semantic System of Rhetorical Expansion: Rhetorical Structure Theory revised. The Meaning Potential of Language: Mapping Meaning Systemically.

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

Unsworth, L. (in press/2004) Comparing school science explanations in books and computer-based formats: The role of images, image/text relations and hyperlinks. *International Journal of Instructional Media,* Vol31:3.

Wu, C. 2000. *Modelling Linguistic Resources: A Systemic Functional Approach*. PhD thesis, Dept of Linguistics, Macquarie University, Sydney.