Australian Government
Department of Defence
Defence Science and
Technology Organisation

The Use of Systemic-Functional Linguistics in Automated Text Mining

Astika Kappagoda

Command, Control, Communications and Intelligence Division
Defence Science and Technology Organisation

DSTO-RR-0339

ABSTRACT

Systemic-functional linguistics is a linguistic framework for the analysis of grammatical and semantic information in text, with a potential role in automated text mining. This report outlines essential features of the theory, its application in computational work, and the rationale for use in automated text mining, and develops a grammatical annotation scheme– word functions– to enrich a mixed text corpus of newspaper articles and e-mails, for machine learning of semantically-oriented grammatical patterns. Testing demonstrates high accuracy in predicting word functions in unseen text in co-training with other grammatical information, providing the basis for further grammatical and semantic text processing.

RELEASE LIMITATION

Approved for public release

March 2009

The Use of Systemic-Functional Linguistics in Automated Text Mining

Executive Summary

Using grammatical and semantic patterns as the basis for large-scale text processing has wide potential to improve the quality and speed of information management and analytical tasks in the defence and intelligence domains. It is proposed that a robust linguistic model is needed to support the automation of these tasks, which is achieved by co-training semantic and grammatical information with unstructured text, and that systemic-functional linguistics (SFL) provides a prime means for achieving this. SFL is a linguistic theory that has had a substantial presence in natural language processing work for the past 40 years, with recent developments in rule-based and machine learning (ML)-based text processing. An outline of the theoretical apparatus of SFL is presented, focusing on a detailed treatment of the functional structure of word groups and phrases. This is used to derive a grammatical annotation scheme for the labelling of the functions of single tokens in unstructured text (WFG). A justification for using this scheme is presented, and a method is outlined for the preprocessing of unstructured text and for annotation with the WFG scheme, in order to produce training and testing corpora for a ML system employing the 'conditional random fields' algorithm. It is demonstrated via this system that automated WFG annotation can be achieved with high accuracy, and that such labelling supports the automated recognition of other grammatical information such as chunk labelling. It is proposed that WFG annotation provides a robust semantically-oriented foundation for other kinds of semantically-based text processing, such as information extraction and text categorisation, which are important elements in information management in defence and intelligence tasks.

Authors

Astika Kappagoda
Command, Control, Communications and
Intelligence Division

Dr Astika Kappagoda graduated with an MB BS and BA(Honours) in ancient Greek from the University of Sydney in 1996, a MPhil in classical linguistics from the University of Cambridge in 1998 and a PhD in linguistics from Macquarie University in 2004. After a lectureship in linguistics at the University of Wollongong, he joined the then Command and Control Division of DSTO in 2006. He works in the text processing team of the Intelligence Analysis discipline, investigating and developing schemes to semantically characterise text for automated processing, such as information extraction from unstructured text, text mining and text classification, and more generally investigating linguistic approaches to information management.

Table of Contents

Figures

Tables

Examples

1. Introduction

1.1 Information extraction and information retrieval in the defence and intelligence domains

The extraction of relevant information from free text documents is a highly important area of information management in the defence and intelligence domains. An analyst engaged in defence or intelligence tasks is most often presented with large volumes of information, all or some of which is potentially relevant to the particular task at hand. The volume can exceed the capacity of a single person to assimilate in a timely manner to produce high quality intelligence reporting. Consequently, such information needs to be ordered and managed systematically so that the analyst can detect important elements of information and their relationships with each other. Processed in this way, such information can then be used in intelligence reporting that takes account of most or all of the large volume of information that is available, thereby increasing the quality and accuracy of the reporting, and the quality of defence or intelligence agency decision-making that depends on it.

For an analyst habituated in processing all information sources manually, the use and potential reliance on automated tools can seem daunting. In particular, significant anxiety surrounds the perceived potential for an automated system to 'miss' important information, and this indeed is a major consideration in information processing system development because of the clear potential for this to happen. However, a well-designed system successfully minimises this risk to acceptable levels while also introducing a number of benefits to the quality of assessment and reporting. This becomes even more apparent when one considers the difficulties in manually processing large volumes of information in terms of quality and accuracy.

It follows that the problem of managing, sorting and storing large volumes of information can be handled in an automated way, so that pieces of information can be retrieved easily and related to each other. One way of managing the number and variety of information sources is to sort the sources into categories relevant to the analyst, a task which is central to automated information retrieval. In this way, the analyst can readily identify a relatively manageable subset of documents most relevant to the analytical task, without having to manually examine large numbers of documents that turn out not to be relevant. The categorisation of documents can theoretically be along any set of criteria, whether this be document topic, style, document type, the person / people who have produced the document or a document's intended audience. However, in the final instance, any classification scheme (whether this is manual or automated) ideally should be geared to the needs of the analyst, with the result that the classification allows the analyst to readily identify the documents most relevant to the analytical task.

A second, parallel, method of information management is to automatically analyse text directly and extract from these texts relevant information that can be used for reporting in the current instance, and can then be stored for future analytical needs. This information typically is in the form of entities, events, relationships between entities, relationships between entities

and events, and relationships between events (such as a 'timeline' of a sequence of events). These are central activities in the field of automated information extraction.

1.2 The application of information extraction to information management and retrieval

Information retrieval (IR) can only go so far in systematising a large volume of information for an analyst. Even with a preliminary sorting of information sources, there still may remain a large volume of information sources from which the relevant information needs to taken and stored systematically. Information extraction (IE) is the automated process by which such information is gathered out of unstructured sources to be fed into an information storage system. Unstructured information typically comprises documents or other kinds of meaningful material whose form, content and organisation do not reflect the form, content and organisation of the information storage system into which new information is to be added. This typically encompasses free text documents in written or electronic form, and printed or electronic images typically created manually (such as photographs and diagrams). This is to be distinguished from structured information, which is typically of a form and organisation that can be fed into a database or knowledge representation system (for which human-in-the loop verification may be required), such as information in forms and tables where the form and table structure reflects that of a knowledge representation system structure and processes. Put simply, unstructured information is information that is oriented to comprehension by humans, whereas the form of structured information directly reflects the computer- mediated knowledge representation system for which it is destined. In this way, IE is the process of transforming some or all of a given amount of unstructured information into a structured form, thus rendering the information 'computationally transparent' (Moens 2006: 4-10).

IE has the potential to enhance document retrieval and classification tasks. Document retrieval and document classification, driven by queries into a document storage system, was initially a task that developed in the field of information retrieval (IR). Traditional IR was predominantly driven by keyword matching between query and target documents. This severely limited the ability to have 'flexible' searching grounded on the semantics of both query and documents, because such IR systems were 'naïve' to semantics. On the other hand, IE allows for documents to be semantically labelled and systems can be trained on such labelling, with the result that there can be more precise matching between query and document, allowing for more precise and relevant IR strategies based on semantic criteria (Moens 2006: 13-17).

1.3 Classifying the Enron dataset: a 'test case' for text mining

This report focuses particularly on the grammatical and semantic labelling of unstructured text– in particular, e-mail communication. A machine learning system is trained to recognise particular kinds of grammatical and semantic information, and to correlate this information with text tokens. This machine learning approach is taken as the basis for automated text mining– the detection of patterns in textual data (Weiss, Nitin et al. 2005: 1-12). The particular

approach to labelling semantic and grammatical information is linguistically motivated, through the use of the theory of systemic-functional linguistics (SFL), and in particular, systemic-functional grammar (SFG). In this way, SFL provides the linguistic framework for semantic and grammatical tagging on which other text processing activities can be based.

The application of SFL to enrich unstructured text is intended to have utility for a number of text processing tasks in both IE and IR. One particular task is the classification of e-mails from the Enron dataset into 'official' and 'non-official' types, using text mining strategies informed by the machine learning (ML) of SFL information to provide a semantic basis for the classification. This task is intended to be a demonstration of a concept that can potentially be expanded to a wide variety of information processing tasks within the defence and intelligence communities.

The work performed to date indicates that the labelling of SFL categories at the single word level is a feasible task, and that such labelling enriches unstructured text with semantic information whose patterning is machine learnable and can provide the basis for other kinds of semantic processing. This means that there is wide scope to employ the approach in different contexts. Given that the SFL categories used are applicable to all varieties of language, a number of IR strategies can be applied to a variety of e-mail and informal text datasets. In this way, this work provides the basis for the classification of a large number of documents on meaningful criteria that are relevant to an intelligence analyst. This allows an analyst to better manage large volumes of information.

As well as text categorisation, the application of SFL is also intended to inform projects which lie more within the IE domain. These include named entity and event recognition, intra-document normalisation, entity relationships, and temporal information. Such extracted information, after manual verification, can then be incorporated into a knowledge representation system which can then be readily shared between individuals, groups and organisations. In this way, this work is expected to be highly applicable to information management and processing in analytical tasks, where reporting and interpretation is expected to take account of all the relevant information available in a timely manner.

2. The need to link linguistic structure and linguistic meaning

What is required in any semantically-based text processing system is a model that links linguistic structure to linguistic meaning. Such a model would resemble a person's ability to discern from a document what states or events are being construed through the language, and to what degree two or more documents are similar in terms of their meaning. Simply put, such a cognitive model aims to link the physical manifestation of language (sound and writing) that a human perceives, to what that language means. This linkage lies at the heart of any comprehensive linguistic theory. For this reason, linguistic theory has an important role to play when specifying such a model for a computationally driven system that depends on classifying text and extracting the content of what is said and written.

2.1 The role of linguistic theory in adding value to automated pattern recognition in text

Computational systems are particularly good at detecting patterns in the obvious, physically manifest aspect of language– in the case of e-mail and other electronic documents, patterns of characters and strings of characters. In this way, and with the appropriate algorithms, ML systems are highly effective in detecting patterns of letters or characters, patterns of character combinations and collocations of words. Computational systems have the potential to detect relatively subtle patterns and correlations in a large amount of data which may be beyond the capabilities of the individual language analyst to detect. However, ML systems pick out these correlations in 'ignorance' of what the patterns of characters and words mean.

Linguistic theory– in the case of this task, SFL– has the role of enriching detected words and phrases with grammatical and semantic information, by 'tagging' individual tokens in unstructured text. An ML system can then learn the associations between document tokens and their grammatical and semantic labels in a corpus of documents (the 'training corpus'), thereby building a semantic model of the corpus that is linked to its linguistic forms. Such a model can then be used to infer semantic representations of documents to which the system has not been exposed.

2.2 The use of co-training for grammatical and semantic labelling

ML methods also allow for co-training of tokens with multiple tags. If we assume that a document or corpus is broken up into tokens *token$_1$*, *token$_2$*, *token$_3$*, …, each of these tokens can be assigned one or more tags. Each of the tags assigned to a token comes from a tagset whose members are unique to that tagset and are independent of (or, 'orthogonal to') other tagsets.

Theoretically there is no limit to the number of tags from different tagsets that can be assigned to a token. In this way, the combination of tags with their tokens forms a two-dimensional matrix (Table 1). The development of such a matrix is one of the processes used in text mining, and data mining in general (Weiss, Nitin et al. 2005: 3-6).

Table 1 – Matrix of tokens with multiple tags derived from orthogonal tagsets

Token	Tags				
token$_1$	a$_1$	b$_1$	c$_1$	…	z$_1$
token$_2$	a$_2$	b$_2$	c$_2$	…	z$_2$
token$_3$	a$_3$	b$_3$	c$_3$	…	z$_3$
…	…	…	…	…	…
token$_n$	a$_n$	b$_n$	c$_n$	…	z$_n$

Given the complexity of language features in documents and the multiple kinds of information that can be extracted from text, allowing the labelling of each of the text tokens with multiple orthogonal tags is advantageous for text mining tasks. A given token in a document has a number of linguistic features associated with it, from its constituent characters through to its role in the semantics of the text. From the point of view of text mining, tokens can be labelled with their grammatical role and with what they mean.

2.3 Grammatical and semantic enrichment of the Enron dataset

Such an approach outlined above can be applied to classifying the e-mails of the Enron dataset. In the training of the system, each token of the selected e-mails is labelled with grammatical and semantic information. Specifically, each token (typically, a word or a separate item of punctuation) of the training corpus is manually labelled with its part-of-speech (POS) category tag, a tag that denotes a word's grammatical function in the group or phrase to which it belongs (word function in the group, or WFG), and its location within a group or phrase of a particular type ('chunk' information). These tags are then co-trained with tags that denote the location of the token in a sentence of a particular e-mail class (official, non-official, mixed or uncertain) (Table 2).

*Table 2 – Tagging of tokens in the training corpus (**POS** = part-of-speech; **WFG** = word function in the group; **CHUNK** = group / phrase type; **CAT** = category of sentence with respect to e-mail type)*

Token	Tags			
$token_1$	POS_1	WFG_1	$CHUNK_1$	CAT_1
$token_2$	POS_2	WFG_2	$CHUNK_2$	CAT_2
$token_3$	POS_3	WFG_3	$CHUNK_3$	CAT_3
…	…	…	…	…
$token_n$	POS_n	WFG_n	$CHUNK_n$	CAT_n
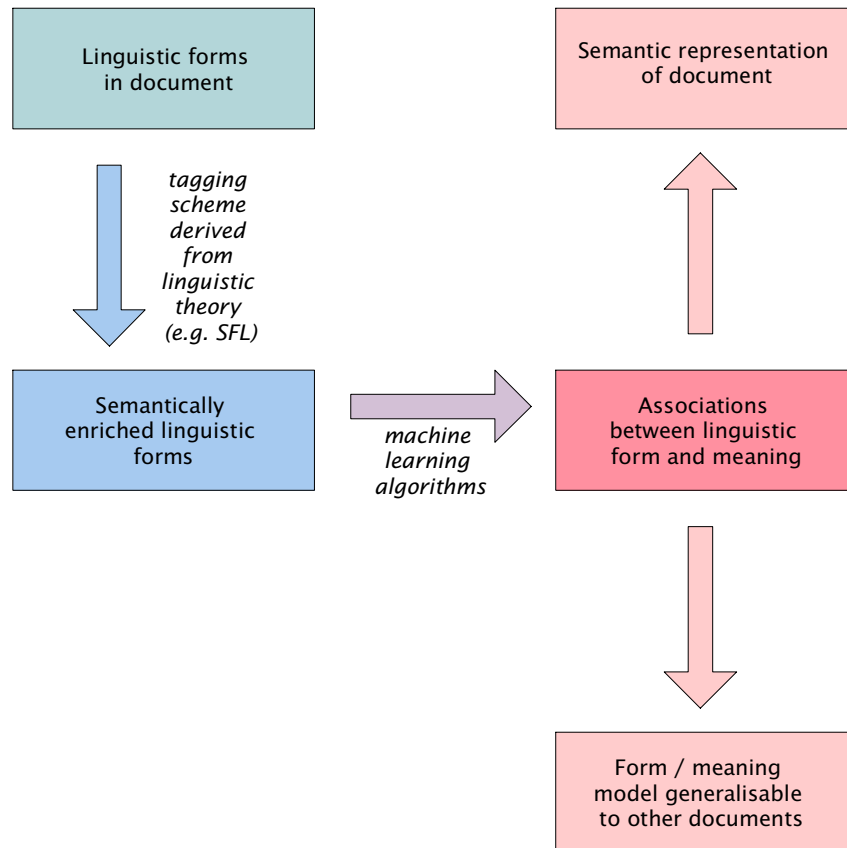
Figure 1 – 'Structure-function model' implicit in training set

The innovation in this task is the use of 'word functions' (via the WFG tagset) as part of the grammatical and semantic enrichment of the dataset for training of the text classification system. The WFG tagset, derived from the theory of SFL, denotes the grammatical function of single words. Used in this context, it aims to strengthen the linkage between the linguistic form of a token and its meaning, thus increasing the chances of successful ML of the semantic content of the tokens. In this way, WFG annotation of the training set provides a useful 'layer' within the model that links linguistic form and meaning (Figure 1).

3. Using SFL in computational contexts

SFL has had a presence as a linguistic theory in computational contexts for a number of decades, and work still continues in this field. The main attractions of SFL from a computational perspective are: the theory's emphasis on meaning in context; the representation of language in terms of system networks; the notion of systems constantly adapting to continuous input and usage; the dependence on actual text utterances to justify the categories of the theory and the structure of the system networks; and the detailed treatment of grammar and its relationship to semantics. This work has proceeded alongside more 'mainstream' language processing work, which has been closely aligned with Chomskyan generative and formalist description and theory. This influence reflects the

influence Chomskyan and generative approaches had on linguistics generally, but also arose because Noam Chomsky's conception of an innate language faculty naturally aligned with people's attempts to encode linguistic knowledge into linguistically 'ignorant' systems in artificial intelligence and NLP (Manning and Schutze 1999: 4-5).

More recently, however, there has been an increasing trend towards implementing approaches which are statistically based, which are based on the study of language in actual use, and which have been less ready to accept a theoretically 'pure' approach, concentrating instead on those linguistic insights which 'work better' (Manning and Schutze 1999: 4-7). The application of SFL in computational contexts, although deriving its approaches significantly from its theoretical apparatus, can be seen as part of this trend.

 Much of the work using SFL to inform tasks in computational linguistics has occurred in places where SFL theory has a presence– predominantly in Europe and Australia. Areas of application include machine translation, text classification, natural language generation, and parsing. (O'Donnell and Bateman 2005) provides an overview of the history of the application of SFL to computational work.

3.1 Computational SFL work on word group functions

On the text processing side, there has been work in both grammatical parsing and text classification on semantic grounds. (Munro 2004) is of particular interest for this report. This outlines an ML system that employs a 'mixture model' using entropy measures (similar to a 'maximum entropy' method– see Munro 2004: 16) in the context of semi-supervised learning of the grammar of word groups (the grammar on which the WFG scheme in this report is based), to infer this word group grammar, and parse unseen text, with about 90% accuracy even with texts of different genre and register. This is meant as a parsing system intended to aid other text processing activities, such as disambiguation of terms in translation and information retrieval. Frequency- and collocation-based approaches have also been employed in the identification of nominal group functions in German (Bohnet, Klatt et al. 2003). However, these approaches are different in orientation to the work covered in this report. The work contained in (Munro 2004) and (Bohnet, Klatt et al. 2003) is oriented towards disambiguation strategies and word group boundary determination, and the creation of an ML representation of text that is labelled in a way consistent with SFL theory. In contrast, this report concerns a tagging system oriented to labelling the function of single words– a task which has similarity to, but is distinct from, the labelling of the immediate constituents of word groups. However, it has been demonstrated by this previous work that the categories of SFG as applied to text are amenable to ML techniques, and that the models generated do have at least some portability to testing and deployment in other domains.

3.2 Computational work on clause-level grammatical functions

Other work has concentrated on clause-level grammar and is still in progress. Initial work on clause-level parsing of texts was achieved in (O'Donnell 1994) and (Souter 1996), but the grammar was rule-based rather than being truly inferred on statistical grounds from a corpus,

and parsing times were prohibitively slow. Parsers had particular difficulty with complex sentences. Part of the reason for this is that SFL has a large number of grammatical categories and a single category can be manifested by a group of words, with the result that large numbers of possible parse trees are generated. These problems stand somewhat in contrast to the implementation of SFL-based natural language generation systems which have had considerable success (where the large number of meaningful categories can be seen as an advantage for language generation).

What is lacking is a sufficient corpus marked up with SFL grammatical formalisms which can then be used to train an ML system to learn a statistically-based SFL grammar for testing and deployment across a variety of domains. SFL grammatical categories are semantically oriented (and therefore suited to semantic text processing) but at the same time sufficiently general to allow domain portability, which is a desired feature (Moens 2006: 35). It therefore makes sense to develop a suitable marked-up corpus to 'machine learn' the grammar. The work in (Honnibal 2004a) and (Honnibal 2004b) develops automated means to convert the syntax trees of already existing corpora (such as the Penn Treebank). If this work is extended further, this may provide a quicker way to develop large-scale training corpora, allowing for more development time of the ML algorithms to successfully induce a robust SFL grammar.

3.3 Text classification

Text classification and sentiment analysis are developing areas in the computational application of SFL. The work in (Whitelaw and Argamon 2004), (Whitelaw, Herke-Couchman et al. 2004) and (Whitelaw and Patrick 2004) explicitly talks of SFL-derived features that were used to classify potentially fraudulent web documents in the 'Scamseek' project. The features selected derive from grammatical and semantic systems developed in SFL, such as conjunction, modality, pronominal determination and comment. These features, both in terms of frequency and their chaining across a text, were taken as salient stylistic features that had a direct relationship to the latent meaning of the documents (this relationship being determined through manual linguistic research) and use of a particular linguistic subsystem, thus achieving a degree of domain independence. SFL-inspired text classification strategies have also been developed using purely semantic information that derives from semantic system networks such as that of appraisal (Martin and Rose 2003 : 22-65; Martin and White 2005). In particular, this has been used to classify movie reviews according to the sentiment (favourable or unfavourable) that they express (Whitelaw, Garg et al. 2005).

Thus there is considerable work that has been done and is underway in the computational implementation of SFL. What is of priority in the text classification domain is the availability of a sufficiently large marked-up corpus of material as a training set for ML, the selection and / or development of suitable ML algorithms that take into account the nature and distribution of SFL categories as applied to text, and an optimisation of the SFL categories to avoid a 'combinatorial explosion' in the ML process. It may very well be the case that the development of a machine-learnt SFL grammar may need to come about through the concurrent training of several processing modules, each of which attend to a select part of the grammar, and whose training models, or the results of testing and deployment, are then combined.

4. Using systemic-functional linguistics to enrich documents

Systemic-functional linguistics (SFL) is a comprehensive linguistic theory that views language as a means for people to make and exchange meanings with each other, and as being heavily informed by the context in which it occurs. In this way, SFL is a theory that aims to account for the kinds of meaning present in language, and how these meanings relate to the context in which people speak, write, hear and read.

4.1 Text and context

SFL is distinguished by the way in which it has a considerable theoretical apparatus for specifying context– in particular, the social context in which people speak. It talks about this context in two ways (Figure 2). Firstly, there is the immediate situation in which people speak– the 'context of situation'. This is comprised of who is talking and the subject matter or state of affairs about which they speak (Field), the social relationship between interactants (Tenor), and the means and style through which they communicate, and how the interaction evolves towards a certain outcome or purpose (Mode) (Halliday and Hasan 1989; Martin 1992: 496-546; Hasan 1999). The *register* of any text is the sum total of meanings which reflect this context of situation.

Secondly, and at the same time, people in the act of communication are thought to be reprising particular conventionalised kinds of communicative activity that are present in the culture of a community– a community's 'way of getting things done', such as telling a story, negotiating the transaction of material goods, mounting an argument, and so on. In other words, there is a 'context of culture' for any text (Martin 1992: 502-503, 546-573). The *genre* of a text is the sum total of the meanings that most directly reflect this context of culture.
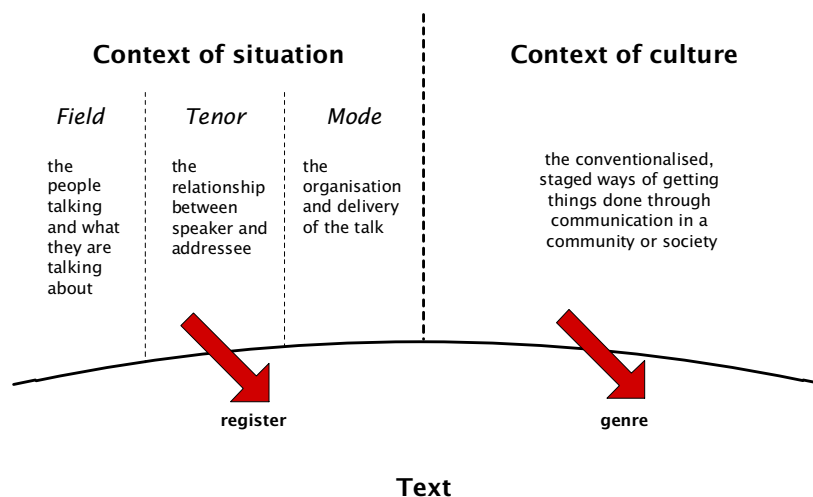
Figure 2 –The relationship between text and context in SFL

9

4.2 Linguistic strata and realisation

Context of situation   **Context of culture**

is realised by   *is realised by*

register   **discourse semantics**   *genre*
(eg. phase analysis; generic structure
potential; move / argument / turn
structure)

Semantics
(systems of meaning)   **relational semantics**
(eg. rhetorical structure; conjunction,
cohesion and coherence;
appraisal/sentiment)

conceptual semantics
(element/entity, figure/event and
sequence/complex event)

is realised by

clause complex configuration
(elements of the complex sentence)

Lexicogrammar   **clause configuration**
(systems of wording)   (elements of the 'simple sentence')

group / phrase configuration
(elements of groups and phrases)

morphological configuration
(lexical and grammatical elements of words)

is realised by

tone group   **rhythm group**
(intonation system)   (prosodic system)

Phonology
(systems of language
sounds)   **syllable / hemisyllable**
(system of sound combinations)

phoneme
(system of available sounds in
the language)

Figure 3 – Strata and realisation in SFL

Like other linguistic theories, SFL proposes that there are several levels (or 'strata') present in language and present in any individual (Figure 3). Each of these levels (semantics,

lexicogrammar, phonology) is characterised by a 'systemic' organisation– that is, each stratum consists of a system of choices from which a speaker selects to create their talking or writing. Thus, in the production of any utterance in a given context of situation and context of culture, a person makes choices at the semantic stratum (choices of meaning). Then, in order to express these meanings, the person makes choices at the lexicogrammatical stratum (choices of wording), which are then in turn expressed through choices at the phonological stratum (choices in the sound system of the language). These interstratal relationships, from semantics to lexicogrammar to phonology, are called realisational relationships, because they bring about the transformation and progression in a language from meaning into an expressible form.

There is a very large amount of research material which justifies this theoretical model of language, and it is beyond the scope of this report to cover this research in detail. However, there are two major differences (among many) between SFL and other linguistic frameworks which are worth pointing out here. The first is that the lexicon of a language is not explicitly labelled in the model as a separate component apart from the grammar, instead being considered part of the 'lexicogrammar'. This is because in SFL, lexis is seen as the endpoints of choice within grammatical systems, and so it makes sense from this perspective to view lexis and the grammatical rules which govern it as a unified system (Hasan 1996).

The second major point of difference worth mentioning here is that there is no stratum in this model labelled as 'pragmatics'. In other language theories, pragmatics is considered to be the domain of 'language in use' by people and societies in context, where these context-specific meanings are manifest in particular features and structures of language (Levinson 1983: ix, 5-11). In SFL this issue is handled as one of 'meaning in context': if one has an appropriate model of context, an appropriate model of semantics and lexicogrammar, and a model of the realisational relationships between the strata, then there is no need to propose pragmatics as a separate stratum of language description.

4.3 Systemic organisation

An important concept in SFL is that a language is organised in terms of systems. This means that at each stratum of a language, the various features possible at that stratum are organised in relation to each other in terms of 'system networks' of choices or options, and that a user of the language makes selections or 'choices' between these various options, navigating through the system network to arrive at terminal points, which represent the final choices that a user makes in producing or understanding an utterance in its context. In SFL, the system network is the central motif in describing the organisation of a language, and in many ways parallels the notion of a system in other fields such as systems engineering, where various parts of a system are said to have interdependent functions in the operation of a system as a whole, and where the parts of a system can be connected to each other by nodes of choice.

An example of a system network describing a certain limited part of the interpersonal lexicogrammar of English is shown in Figure 4. This system network represents the basic grammatical choices that a person makes in understanding or producing a 'speech act'– a statement, question, or a so-called 'minor clause' such as an exclamation or address to a

person. The first choice encountered is between making a full clause or an incomplete clause such as an exclamation. If a full clause is chosen, then there is a choice between the grammatical structure of the indicative (expressions that typically concern information) and the imperative (expressions that involve ordering a person to perform an action). If indicative is selected, another choice is made between the declarative (an expression used to state information) and the interrogative (an expression used to request information). If the interrogative option is selected, then a choice is made between a 'yes / no' interrogative (asking for an answer 'yes' or 'no') and a 'WH-' interrogative (which asks for a specific piece of information).
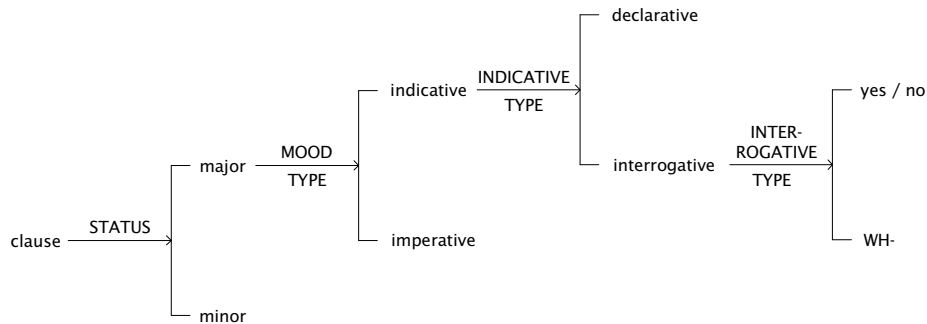
Figure 4 – System network of the system of MOOD in English (after Halliday and Matthiessen 2004: 23)

All of these choices are manifested in the presence, absence, ordering and configuration of the interpersonal grammatical elements of the clause, as the following examples show (Table 3):

Table 3 – Differing interpersonal grammar of declarative, interrogative and imperative clauses

Clause: major: imperative	Get	an application and a brochure	from the Bocconi University!	
	Predicator	**Complement**	**Adjunct**	

Clause: major: indicative: declarative	I	can	get	an application and a brochure	from the Bocconi University
	Subject	**Finite**	**Predicator**	**Complement**	**Adjunct**

Clause: major: indicative: interrogative: yes / no	Can	I	get	an application and a brochure	from the Bocconi University?
	Finite	**Subject**	**Predicator**	**Complement**	**Adjunct**

Clause: major: indicative: interrogative: WH-	Who	can	get	an application and a brochure	from the Bocconi University?
	Wh-Subject	**Finite**	**Predicator**	**Complement**	**Adjunct**

These clauses differ in their interpersonal grammar, in terms of the presence or absence of the grammatical category Finite (the distinction between indicative and imperative clauses) and the relative ordering of the categories Subject and Finite (the distinction between yes / no interrogatives and declaratives). In this way, choices in the system network specify the grammatical structure of the clause.

The above system network accounts for only a fraction of the interpersonal lexicogrammatical system of English. However, the system network is used to describe all of the possible systems in the lexicogrammar of English (interpersonal, experiential or textual), and indeed any kind of linguistic phenomena at any stratum. Thus there are phonological, lexicogrammatical, semantic and contextual system networks used to account for phonological, lexicogrammatical, semantic and contextual features associated with a given instance of speaking or writing. The use of the system network to describe the nature of a language, and language in general, also clarifies the relationship between the strata, because then one can see that given a certain configuration in the context of culture and situation, one makes semantic choices, which are then realised by lexicogrammatical choices, in turn realised by phonological choices.

Systemic organisation in SFL has a number of implications for natural language processing work using this linguistic theory. Firstly, system networks have a relatively simple logical structure to them, using basic operators 'and', 'or', 'if…then', and 're-enter'. This potentially renders them amenable to algorithmic treatment, particularly in the context of natural language generation, and provides a computationally tractable model for the psychological processes that underlie language production and comprehension. In turn, because SFL attempts to model these cognitive processes, this has the potential to improve NLP tasks by making them more 'human-like'. Secondly, system networks have the advantage of making explicit relatively subtle choices that are made in speaking, writing, listening and reading language. This opens up the potential for any computational system using a SFL-based framework to deal with relatively subtle language patterns that may be of interest, particularly subtle grammatical, lexical or semantic patterning that occurs over stretches of

text. This has particular application in the issue of detection of 'veiled talk' and assessing the overall 'tone' of communication.

4.4 Features of the lexicogrammar

4.4.1 Compositionality and constituency

Like other theories of syntax and grammar, SFL proposes that the lexicogrammar (hereafter referred to as LG) has levels of constituency (Figure 5). This is to say that a clause (or 'simple sentence') can be decomposed into its constituent groups and phrases, which can in turn be decomposed into single words. Words themselves can be decomposed into their constituent lexical and grammatical morphemes. In this way, in the LG, there is a hierarchically organised 'rank scale' of linguistic forms. At each constituency level, grammatical functions can be attributed, such as clause functions to the members of a clause, and group functions to members of a group or phrase.
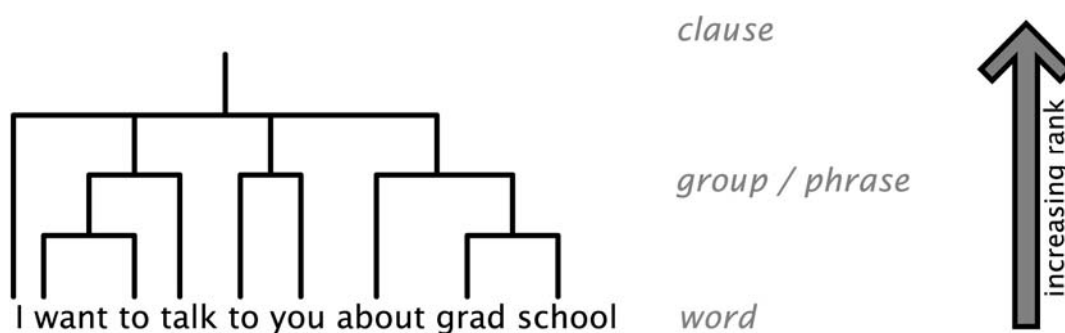
Figure 5 – Constituent structure of the lexicogrammar in SFL

In theory, the concept of constituency can be extended to other linguistic strata. This is particularly true of the phonological stratum, because it is relatively straightforward to decompose syllables into sound clusters, and then into individual language sounds. However, applying constituency as an organising principle at the semantic stratum is highly problematic, as it is difficult to see concept-level semantics as being constitutive of discourse level semantics; another kind of organising principle needs to be invoked. However, at the level of the lexicogrammar, constituency in grammatical structure is relatively easy to see.

4.4.2 Metafunctions

The particular area of interest in SFL for application in text mining is the LG, as LG categories provide the basis for the 'generic' semantic information that is required for portable and

adaptable text mining and classification. The first thing to note is that SFL has three sets of LG categories, each of which represent the three kinds of 'generic' semantic information expressed by the lexical and grammatical structures of the clause. These three categories are referred to as metafunctions.

In any given clause, one of the strands of meaning present conveys the kind of communicative act that is taking place between the speaker and the addressee, such as a statement, demand, exclamation, question, request and so on. This is called the interpersonal metafunction. Another strand of meaning in the clause is to do with the construal of states of affairs in the world– the experiential metafunction. Lastly, another strand of meaning in the clause signals how the information in the clause is organised with respect to other clauses in a text– the textual metafunction. In SFL, each metafunction in the LG is considered as an aspect of the total lexis and grammar of the clause, with its own set of categories. In this way, each clause is said to have three strands of meaning operating simultaneously, and each of these strands can be described through a set of grammatical categories. An example clause from the Enron dataset labelled with these category labels is shown in Table 4.

Table 4 – Trimetafunctional grammatical description of a clause in SFL

Clause	I	want to	talk	to you	about grad school
Interpersonal Metafunction	Subject	Finite: modal: inclination	Predicator	Adjunct	Adjunct
Experiential Metafunction	Participant: Actor	Process: material: behavioural		Participant: Recipient	Circumstance: matter
Textual Metafunction	Theme / Given	Rheme / New			

If one concentrates on the experiential grammatical description of the clause, the grammatical categories model the state of affairs represented by the clause in terms of a process (the kind of event taking place), participants (entities that are directly linked to or involved in the process), and circumstances (elements which contain additional information about the setting or environment in which the process takes place). As such, experiential grammar in SFL is clearly the most relevant set of categories to tasks associated with information extraction (IE). Given that IE often revolves around important tasks such as extracting entities, events and relationships from unstructured text, the categories of experiential grammar closely align with the kinds of information that are typically extracted in the IE process. Further examples of experiential analysis are shown in Example 1.

Demand	was	at a record high	for the month of October	due to strong pick-up sales in North America and Western Europe
Participant: Carrier	**Process: relational & attributive: circumstantial**	**Participant: Attribute**	**Circumstance: time: extent**	**Circumstance: cause**

Can	I	get	an application and a brochure	from the Bocconi University?
Process: material: dispositive	**< Participant: Actor > ***	**Process: material: dispositive**	**Participant: Goal**	**Circumstance: location: point**

The ISO	has obtained	an extension of time
Participant: Actor	**Process: material: dispositive**	**Participant: Range**

Example 1 – Examples of experiential analysis of clauses from the Enron dataset (* the Participant I is said to be 'interrupting' the material Process can get)

4.5 Experiential functions within the group / phrase

In the discussion of metafunctions above, the grammatical labels for the experiential and interpersonal metafunctions denote the function of word groups and phrases in the clauses of which they are constituent. Thus, these labels and categories typically do not apply to single words within the clause in English, except in certain cases such as newspaper headlines where there is a certain amount of abbreviation (Example 2):

Crash	claims	lives
Participant: Actor	**Process: material: dispositive**	**Participant: Range**

Example 2 – Single words fulfilling clause-level grammatical functions

This analysis is not different from that of sentences which are semantically similar but more lexically and grammatically elaborated (Example 3):

The crash	claims	two lives
The car crash	claimed	these lives
The crash of the two vehicles	had claimed	a family of four
The stockmarket crash	will have claimed	a number of victims
Participant: Actor	**Process: material: dispositive**	**Participant: Range**

Example 3 – Clause-level grammatical functions mapping onto immediate clause constituents

The above examples illustrate that experiential functions of the clause properly map onto the immediate constituents of the clause– structures which are at one rank level down from that of the clause, whether those structures be single words or groups of words. Hence they do not consistently map only onto groups, or only onto words.

Many of the immediate constituents of the clause are word groups or phrases. SFL proposes that these word groups have their own internal experiential grammar. Just as the experiential grammar of a clause represents a certain state of affairs in terms of participants, processes and circumstances, each of these participants, processes and circumstances has their own functional grammar that represents their internal characteristics. SFL, similar to other linguistic theories, proposes the following group and phrase types:

- nominal group
- verbal group
- prepositional phrase
- adjectival and adverbial group
- conjunction group
- preposition group

The experiential structure of these will be discussed as they are the most relevant to the functional labelling scheme of word function in the group (WFG). The following characterisation follows, and is consistent with, the descriptions in (Fawcett 2000), (Morley 2004), and (Halliday and Matthiessen 2004). It is important to understand these functions as they form the basis for the annotation scheme that has been devised in this report to support text mining and text classification tasks– the word function in the group (WFG).

4.5.1 Experiential structure of the nominal group

A nominal group typically has the function of representing a participant in the experiential structure of the clause. Structurally, it consists of a noun which is at the 'core' of the nominal group, on which other words and phrases depend, which have functions with respect to the core noun.

The possible functions within the nominal group are as follows:

- Thing (the 'core element')
- Deictic
- Pre-Deictic
- Post-Deictic
- Numerative / Quantifier (simple and extended)
- Epithet
- Classifier
- Qualifier

Each of these will be explained in turn.

4.5.1.1 Deictic and Thing

In many cases in English text, nominal groups consist of a **Thing** plus an element (the **Deictic**) that serves to anchor the nominal group in the text and context in which it occurs (Example 4):

the	crash
Deictic	**Thing**

Example 4 – Deictic and Thing

Here, `crash` is the semantic 'core' of the **nominal group**, the word that most directly expresses the key entity. **Deictics** are typically expressed through words such as definite and indefinite articles or determiners, demonstratives, and possessive determiners (Example 5). These words help to connect the entity to some other entity or event in the surrounding text or context, or 'introduce' that entity into the text for the first time.

a / an	crash(es)
this / that / which(ever) / what(ever)	accident(s)
his / her / its / our / your / their / whose	
no / any / some / all	
each / every / either / neither	
Deictic	**Thing**

Example 5 – Examples of words fulfilling Deictic function

4.5.1.2   Pre-Deictic and Post-Deictic

Pre-Deictics and **Post-Deictics** are words that can occur before and after, respectively, the **Deictic**, and which further describe the relationship between the entity and its surrounding text or context or assumed background values or knowledge, or make the Deictic more 'specific' in its function (Example 6).

all / all of	the / a	same / different / identical / other	sell-
some of / none of		possible / probable / necessary / desired	off
just		normal / ordinary / habitual / typical	
only		partial / complete / entire / whole	
each of		certain / particular / special	
		so-called / reported / alleged / suggested	
Pre-Deictic	**Deictic**	**Post-Deictic**	**Thing**

Example 6 – Pre-Deictic and Post-Deictic

Pre-Deictics serve to 'particularise' what the **Deictic** 'points to' in the text or context. **Post-Deictics** tend to relate the entity represented by the nominal group to certain 'value scales' such as its probability, how often it occurs, how well it is known, whether it is directly known or reported, or qualifies the scope of the **Deictic**. These group-level functions are in many ways homologues of clause-level grammatical functions such as modality, logico-semantic relationships, and reported speech (Halliday and Matthiessen 2004: 317).

4.5.1.3   Numerative and Quantifier

Numerative and **Quantifier** are functions that typically occur after the **Deictic** or **Post-Deictic** (if any) and before the Thing in the nominal group, and serve to express the number or quantity of the Thing (Example 7). The **Numerative** is typically expressed through precise, approximate or relative numbers or ordinals:

Deictic	Numerative	Thing
the	thirty-nine	steps
a	dozen	roses
	three	kings
	about 100 / roughly 100	people
	few	applicants
	just as many	supporters
the	eighteenth	hole
the	last	chocolate

Example 7 – Simple Numeratives

Numeratives can also be 'extended'– that is, they can consist of a group of words which are linked to the Thing that they enumerate by the word 'of' (Example 8):

Deictic	Numerative	Thing
an	unknown number of	civilians

Example 8 – Extended Numerative

The **Quantifier** expresses a quantity of the Thing, including definite, indefinite and relative amounts (Example 9):

Deictic	Quantifier	Thing
a	quarter	pounder
	1g	penicillin
	2kg	tomatoes
	100,000	dollars*
(the)	little / less / least	evidence
(the)	much / more / most	ice-cream

Example 9 – Examples of Quantifier function

*amounts of money, although they appear on the surface to be countable objects, are more sensibly handled as a Quantifier + Thing structure where the money denomination is treated as a substance which can be measured in units (compare 3 dollars to 3 coins).

It should be mentioned that the category of **Quantifier** is usually treated as a subvariety of **Numerative** (Halliday and Matthiessen 2004: 317-318; Morley 2004: 76; (Munro 2004: 28). The distinction is made here because it was felt that for the purposes of IE it would be useful to make a distinction at a relatively low level between numbers of distinct entities and amounts of non-discrete substance.

4.5.1.4   Extended quantification

Some expressions of quantity are expressed through a nominal group in their own right. These are connected to the entity they quantify through the use of the word of (Example 10). Such a structure is handled as a 'pre-quantifier' occurring before the nominal group it quantifies. Both of these have their own internal structure:

a	spoonful	of	the	sugar
Deictic	**Thing**			
Pre-Quantifier			**Deictic**	**Thing**

Example 10 – Extended Quantifier

4.5.1.5   Epithet and Classifier

The **Epithet** and **Classifier** are functions which further describe the Thing in some way, and which occur between the Deictic and Thing (Example 11). The Epithet describes some quality of the Thing which is either objective or subjective:

the	largest	deficit
my	miserable	day
a	more desirable	location
this	sceptred	isle
such	highly coloured	illustrations
that	most interesting	conversation
Deictic	**Epithet**	**Thing**

Example 11 – Examples of Epithet

A feature of many **Epithets** is that they are often subject to intensification, through inflected forms of adjectives (the comparative and superlative) or modification of the adjective with words such as *more*, *most*, and *very*. This turns out to be a very useful test for the presence of the **Epithet,** particularly when distinguishing them from Classifiers. Furthermore, there can be

multiple Epithets within the nominal group, with the effect that several qualities are attributed to a single entity (Example 12):

a	big	red	car
Deictic	**Epithet**	**Epithet**	**Thing**

Example 12 – Multiple Epithets

The **Classifier** has the function of assigning the **Thing** to a particular category or type (Example 13).

the	trade	deficit
that	military	operation
	software	engineering
the	British	government
Deictic	**Classifier**	**Thing**

Example 13 – Examples of Classifier

Epithets and **Classifiers** can coexist within the one nominal group (Example 14):

these	expensive	government	policies
Deictic	**Epithet**	**Classifier**	**Thing**

Example 14 – Collocation of Epithet and Classifier

4.5.1.6 Qualifier

Apart from the **Thing**, the above nominal group functions all have the role of modifying the sense of the Thing in some way, by either adding more information or relating it to the context of what is said or written. All of these functions obligatorily precede the Thing, and thus can be altogether classed as 'pre-modifiers'. The **Qualifier**, however, is the only nominal group function which post-modifies the **Thing**, and thus obligatorily follows the **Thing** in the nominal group. It also provides additional information about the Thing, typically to further define the **Thing** in relation to other entities or events in the surrounding text or context. Some instances follow in Example 15.

Deictic	Thing	Qualifier
the	cargo	below
the	girl	[with the Surrey fringe on top]
the	violence	[in the West Bank]
the	situation	[on the ground]
that	doggie	[in the window]
those	presents	[for the married couple]
an	offer	[[you can't refuse]]
the	man	[[who would be king]]

Example 15 – Examples of Qualifier (see below for explanation of bracketing)

A common characteristic of the **Qualifier** is that it rarely consists of a single word. Instead, one often finds a group or phrase (marked with single brackets, [ ]), typically a prepositional phrase acting as **Qualifier** in a nominal group by being 'rank shifted' into the post-modifying position. One can even have full clauses with their own verb (marked with double brackets, [[ ]]) rank shifted into a nominal group as **Qualifier**– this is the case with the so-called 'defining relative clause' as in The man [[who would be king]] . It follows that many of these **Qualifiers** can in turn be analysed for their own clause and group functions (see section 4.5.5).

4.5.2 Experiential structure of the verbal group

The verbal group is the word group that typically expresses the process in a clause. Like the nominal group, it has a functional core– the Event– which is premodified by other functions that provide additional information about the occurrence, timing and internal structure of the Event. These functions are:

- Finite
- Event
- Auxiliary
- Polarity

Each of these will be discussed in turn.

4.5.2.1 Finite and Event

The **Finite** is the first functional element of the verbal group (Example 16). It carries what is termed the 'primary tense' (past, present or future) of the verbal group.

Verbal Group		Traditional tense designation
will	eat	simple future
has	come	perfect
shall	go	simple future
may	call	modal present
Finite	**Event**	

Example 16 – Examples of Finite

A simple Finite + Event structure is typically shown in the 'simple future' tense of verbal groups.

The **Event**, as mentioned before, is the semantic core of the verbal group, in the sense that it is the element that expresses directly the nature of the process or event. An **Event** by itself does not carry any tense information. However, there are some forms of verbal groups in English which conflate or map the **Finite** and the **Event** onto a single word (Example 17). This typically happens in verbal groups which are traditionally described to have the 'simple present' or 'simple past'. Such single-word verbal groups are labelled simply as **Finite/Event**.

Nominal group	Verbal Group	Traditional tense designation
The cat	saw	simple past
She	came	simple past
The leader	wishes	simple present
They	pass	simple present
	Finite / Event	

Example 17 – Conflated Finite and Event in simple tenses, labelled as Event

Some verbal groups are non-finite– that is, they lack a primary tense which relates to the time of speaking or writing. These verbal groups are typically found in subordinate or embedded clauses in sentences, and in commands:

<u>E-mail</u> the writers

(non-finite verbal group present in imperative)

Our continued focus as one team will be [[<u>to present</u> "one face to the customer"]]
(non-finite verbal group present in embedded clause)

by … <u>doing</u> the right thing for Compaq during this challenging time, || we will continue to win the endorsement of our customers and partners
(non-finite verbal group present in subordinate clause)

These verbal groups lack a **Finite** altogether and are simply labelled as **Event**.

4.5.2.2 Multi-word Events

Frequently, the **Event** in a verbal group consists of a single word as verb. However, the Event can consist of a 'phrasal verb' consisting of a full verb and another word (often a preposition or adverb, or sometimes another verb) which modifies the sense of the verb (Example 18). The sense of this phrasal verb is not readily predictable from the individual senses of the full verb and the modifying word.

Nominal Group	Verbal Group		Nominal Group / Prepositional Phrase	Sense of phrasal verb
I	can	go through	the partnership structure	'explain, discuss'
She	will	find out	the answer	'discover'
They	have to	make do	with biscuits	'be contented'
We	'll	put up with	almost anything	'tolerate'
	Finite	**Event**		

Example 18 – Phrasal verbs

One test to distinguish phrasal verbal groups from verbal groups followed by a preposition is to rearrange the sentence to separate the verb and the preposition. When performed on a sentence with a phrasal verb, there is significant disruption of the sense or grammaticality:

*through the partnership structure I can go

(compare to I can go through the town centre / through the town centre I can go)

*out she will find the answer

*do with biscuits they have to make

*with almost anything we'll put up / *up with almost anything we'll put

4.5.2.3 Auxiliary

The **Auxiliary** can occur one or more times in the verbal group. They are typically 'small' verbs, which have the function of expressing the secondary tense and the aspect (that is, the internal time structure) of the **Event** (Example 19). When they occur, they occur between the **Finite** (if present) and the **Event**.

Verbal Group			Traditional tense designation
will	be	doing	future progressive
have	been	following	present perfect progressive
are	going to	help	'simple future' / future in present / incipient future progressive
may	have	missed	modal present perfect
Finite	**Auxiliary**	**Event**	

Example 19 – Examples of Auxiliary function

4.5.2.4 Polarity

The function of **Polarity** is to negate the existence of the whole event denoted by the verbal group, usually expressed by the word 'not' (Example 20).

26

				Traditional tense designation
has	not	been	seen	negated present perfect
is	not		coming	negated present progressive
will	not	have	sought	negated future perfect
Finite	**Polarity**	**Auxiliary**	**Event**	

Example 20 – Polarity

4.5.3 Adjectival, adverbial, and conjunction groups

Adjectival groups typically have the function of assigning a quality or attribute to an entity in a clause. Adverbial groups typically have the function of expressing a circumstance in the clause, providing further information about the 'happening' in the clause. In many ways, their internal functional structure resembles that of the nominal group, with an adjective or adverb acting as the core of the group. In each case, this core can be premodified with one or more words that modify the degree of the adjective or adverb, with the function **Modifier** (Example 21):

In this way, both adjectival and adverbial groups have a common structure. This is consistent with the view that both categories can be viewed as subtypes of a 'quality group' (Fawcett 2000: 206-207).

very	accurate	
totally	hard	
hardly	compliant	*Adjectival Groups*
altogether	incomprehensible	
utterly	stupid	
fairly	convincingly	
reasonably	quietly	*Adverbial Groups*
somewhat	happily	
Modifier	**Head**	

Example 21 – Modifier in adjectival and adverbial groups

4.5.3.1 Adjectival groups

Adjectives can be the sole constituents of their respective groups, but can be intensified through the regular morphological system of English to express degree (*-er*, *-est*), such as *kinder* (from *kind*), and *greatest* (from *great*). As they are single words, they are typically labelled as (adjectival) **Head**.

Adjectival groups can also contain a definite article functioning as **Deictic** (Example 22):

the		greatest
the	very	best
a	bit	worse
Deictic	**Modifier**	**Head**

Example 22 – Presence of Deictic in adjectival group

4.5.3.2 Adverbial groups

Adverbial groups, as noted before, have the function of providing additional information about a happening or a state of affairs. This information can be of two types. The first kind is of 'experientially oriented' information independent of the speaker. The second kind relates to 'interpersonally oriented' information, which relates the information of the clause to the actual act of speaking or to the speaker's estimation of the likelihood or occurrence of the event, or the speaker's judgement or comment on the event with respect to their value system– what is often called 'appraisal' in SFL (Martin and Rose 2003: 22-65). Some examples of both types follow in Table 5.

Table 5 – Experiential and interpersonal adverbs

Experientially oriented adverbs	Interpersonally oriented adverbs / adverbial groups		
quickly, forcefully, there, here	tense related		later, soon, now, imminently, just now, earlier, ago
	aspectual		still, continually, gradually, immediately, just (eg. *he's just arriving*), just about, almost, nearly, hardly
	modal	likelihood / ability	possibly, probably, conceivably
		obligation	necessarily, obligatorily
		frequency	once, sometimes, occasionally, often, frequently, always
		usuality	usually, rarely, typically, customarily
		inclination	rather, preferentially
	appraisal		sadly, unfortunately, beautifully, grotesquely, unexpectedly, harmoniously

4.5.4 Conjunction groups and particles

Conjunctions (words which link groups and clauses together into a sequence) most often occur as single words. However, a group of words can also function in totality as a conjunction:

- but also
- even if

- just as

Particles are exclamations and other small words which occur in the clause but do not take part in its lexicogrammatical structure. These are often words which occur as a result of 'mistakes' or pauses in speech production, or are often present in conversation turns to start, maintain or conclude the conversational flow between interactants:

- ah, um, er (non-words)
- hi, hello, hey, cheers, bye, well, right, yeah, no (single words)

Single word abbreviations often used in electronically mediated text (such as LOL, IMHO, LMAO, AFAIK, among others, 'emoticons' and single word expletives) are also labelled as particles.

4.5.5 Prepositional phrases

Prepositional phrases typically express a circumstance in the clause. They typically consist of a preposition followed by an element which the preposition 'governs'. This element is often a rank shifted or embedded nominal group (or clause in certain cases). In functional terms, prepositions are often seen as 'verb-like' (and many prepositions in English indeed do derive originally from full verbs). For this reason, prepositions are often said to 'govern' the words, groups or phrases to which they are attached, and therefore prepositional phrases are often seen to have a 'predicator-complement' structure akin to that of verbs and what they govern (Halliday and Matthiessen 2004: 359-361) (Example 23).

Prepositional Phrase		Structure of Prepositional Complement
for	corruption	noun
under	[the boardwalk]	rankshifted nominal group
by	[the sea]	
after	[the recession]	
in front of	[the house]	
towards	[the holiday season]	
due to	[the backlog of orders]	rankshifted nominal group with post-Deictic / extended Quantifier element
before	[the start of the competition]	
for	[those [[who are interested]]]	rankshifted nominal group containing embedded clause as Qualifier
towards	[[advancing our own interests]]	rankshifted embedded clause
Prepositional Predicator	**Prepositional Complement**	

Example 23 – Examples of functional configuration of prepositional phrase

It should be noted that the **Prepositional Predicator** does not necessarily map onto a single word. Some **Prepositional Predicators**, (such as due to, because of, for the sake of, to the side of and thanks to, among others) are in fact 'preposition groups'. These prepositional groups, like adjectival, adverbial and conjunction groups, can have another word which modifies their sense (largely due to, primarily because of, only for the sake of, just to the side of, no thanks to). These words can be labelled **Modifier**.

5. The WFG annotation scheme

The word function in the group (WFG) annotation scheme is heavily based on the functions within word groups and phrases as outlined in Section 4.5. It is a tagging scheme that aims to semantically label individual words with information that denotes the function of that single word in the word group in which it occurs. In this way, WFG annotation provides a functional counterpart to parts-of-speech annotation for single words. Table 6 shows the list of tags

employed in this scheme, together with their correlation with the word group functions in Section 4.5.

Table 6 – Table of word functions in the group (WFG) tags

Group / Phrase Class	Tag	SFL Word Function	Refer to Section	Example
Nominal Group	**PRD**	Pre-Deictic	4.5.1.2	**all** the children
	DT	Deictic	4.5.1.1	**the** cat
	WHD	Wh-Deictic	4.5.1.1	**Which**/**what** car will you drive? (expecting the answer to refer to a particular car, eg. *I'll drive this car over here*)
	POD	Post-Deictic	4.5.1.2	the **very** problem the **following** issues the **above** argument no **particular** problem
	NM	Numerative	4.5.1.3	those **two** trains
	QN	Quantifier*	4.5.1.3	a **quarter** pounder
	EP	Epithet	4.5.1.5	a **beautiful** day
	CL	Classifier	4.5.1.5	the **Siamese** cat
	WHC	Wh-Classifier	4.5.1.5	**What** car will you buy? (expecting the answer 'I'll buy this kind of car', eg. *I'll buy a Golf*)
	TH	Thing	4.5.1.1	the Siamese **cat**
	WHT	Wh-Thing	4.5.1.1	**What** is the problem?

Group / Phrase Class	Tag	SFL Word Function	Refer to Section	Example
				they would **rather** go
	ASPC	Aspectual Adjunct	4.5.3, 4.5.3.2	they were **nearly** there; they were **about** to go; they are **still** there
	WHA	Wh-Adverb	4.5.3	**How** will you go there? **Why** will you stop?
Conjunction Group	CE	Conjunction Group Element (coordinating or subordinating)	4.5.4	if/CE only/CE, rather/CE than/CE
Prepositional Phrase	PR	Preposition / 'Predicator'	4.5.5	**on** the table by the seaside
Existent	EX			**there** is a price
Particle	PART	Particle	4.5.4	**well**, it's possible

5.1 Preprocessing of text for WFG annotation

For WFG annotation, texts need to be rendered in a form that allows for effective functional grammatical annotation. There are two aspects to such preprocessing– tokenisation and grammatical analysis– as usually occurs with any linguistic annotation of text for computational purposes. Tokenisation aims to transform a unitary text or corpus of texts into discrete items that can be individually labelled with items from the annotation scheme. The grammatical analysis is required for linguistically appropriate and consistent labelling of the tokens.

5.1.1 Tokenisation

Whitespaces are generally taken as being token boundaries in text. Punctuation marks are also taken as separate tokens regardless of the presence or absence of whitespace around them, with the following exceptions (Table 7):

Table 7 – Exceptions to tokenisation of punctuation

Exceptions to tokenisation	Examples
hyphens (single dashes) which are part of a single word	topsy-turvy Hay-Roe
punctuation which forms part of an email address or URL	http://www.dsto.defence.gov.au steve@mac.com
apostrophes which represent the missing letters in a contraction	doesn't shan't haven't
emoticons	:-) :-S
Telephone numbers	08-8252-6700

Each single word is taken as one token. However, words which are contractions are split into separate elements with their distinct grammatical functions (Table 8):

Table 8 – Treatment of contractions

Contraction Type	Original Word	Full Form	After Tokenisation
contracted negation	haven't	have not	have / n't
	won't	will not	wo / n't
	can't / cannot	can not	ca / n't can / not
contracted possession	Julia's	'belonging to Julia'	Julia / 's
	James'	'belonging to James'	James / '
contracted future tense	they'll	they will	they / 'll
	Ken'll	Ken will	Ken / 'll

Contraction Type	Original Word	Full Form	After Tokenisation
contracted past tense	we'd	we had	we / 'd
	I've	I have	I / 've
contracted modals	she'd	she would (also, *she had*)	she / 'd

5.1.2 Grammatical analysis

Grammatical analysis of the text is an essential part of accurate and consistent linguistic tagging, as it is for any manual linguistic annotation no matter what the annotation scheme employed. This is a process that is informally and implicitly done during any annotation; however it is a process worth making explicit. The recognition of the grammatical functions of single words presupposes that the following steps of analysis have been undertaken:

- recognition of sentences within the text or corpus
- division of text or corpus into individual clauses (± recognition of embedded / rankshifted clauses)
- division of clauses into constituent groups and phrases (± recognition of embedded / rankshifted groups and phrases)

This process should result in the identification of the smallest groups and phrases whose immediate constituents are single words. This then allows subsequent annotation of the single words with their WFG category labels. The following table of examples (Example 24) illustrates the intended output of the tokenisation and analysis preprocessing steps.

Original Text	Resultant groups and phrases	List of tokens
Two apartments in [the area] have been very successful. ‖	Two apartments in the area have been very successful	Two / apartments / in / the / area / have / been / very / successful / . /
The property would < probably > be completed 18 months out. ‖	The property probably would be completed 18 months out	The / property / would / probably / be / completed / 18 / months / out / . /

Original Text	Resultant groups and phrases	List of tokens
Also, Dave knows || that my husband is coming with me || but I don't know || if Andrea knows. ||	Also Dave knows that my husband is coming with me but I don't know if Andrea knows	Also / , / Dave / knows / that / my / husband / is / coming / with / me / but / I / do / n't / know / if / Andrea / knows / .
"The end state is [[to destroy the al-Qa'ida influences [in this province] || and eliminate their threat [against the people]]] ," said US commander General Mick Bednarek.	The end state is to destroy the al-Qa'ida influences in this province and eliminate their threat against the people said US commander General Mick Bednarek	"/ The / end / state / is / to / destroy / the / al-Qa'ida / influences / in / this / province / and / eliminate / their / threat / against / the / people / , / " / said / US / commander / General / Mick / Bednarek / . /

Example 24 – Intended output of grammatical analysis and tokenisation for WFG annotation

5.1.3 Differences between WFG annotation and standard SFL analysis of word groups and phrases

The major difference between WFG annotation and standard SFL word group analysis is that the WFG scheme is applied exclusively to single words. Standard SFL analysis of word groups and phrases labels immediate functional constituents of the group or phrase, which are often single words but (especially in the case of the Qualifier) often are groups or phrases themselves. In contrast, pre-processing for WFG annotation involves breaking these groups and phrases down further till one arrives at groups and phrases whose immediate constituents are single words, which are then labelled with the WFG functions. This illustrated in the following table (Table 9).

Table 9 – Differences between WFG annotation and SFL group / phrase analysis

	the	al-Qa'ida	influences	in this province		
SFL Analysis	**Deictic**	**Classifier**	**Thing**	**Qualifier**		
				in	this province	
				in	this	province
WFG Annotation	**DT**	**CL**	**TH**	**PR**	**DT**	**TH**

In this way, the WFG scheme should be seen as an *adaptation* of SFL group / phrase analysis to provide single word-level grammatical information. This is so that the scheme provides a 'functional counterpart' to the structural information embodied in POS annotation. This in turn provides 'structure and function' modelling to grammatical analysis and annotation in order to improve information extraction (Figure 6).

		the	al-Qa'ida	influences	in	this	province
function	*SFL Clause Constituent Function*	**Participant: Goal**					
function	*SFL Group Function*	**Deictic**	**Classifier**	**Thing**	**Qualifier**		
structure	*Chunk Annotation*	**NP**			**PP**	**NP**	
function	*WFG Annotation*	**DT**	**CL**	**TH**	**PR**	**DT**	**TH**
structure	*POS Annotation*	**DT**	**NNP**	**NNS**	**IN**	**DT**	**NN**

Figure 6 – A 'structure-function' model using WFG annotation of a nominal group in the sentence 'the al-Qa'ida influences in this province'

It can be seen that this form of text enrichment provides a relatively robust model to link linguistic form and linguistic meaning to support IE tasks.

5.1.3.1 Tensed Auxiliary and Finite

The other point of difference in the WFG scheme is the tagging of tenses of the Finite, Auxiliary and Event in the verbal group (Table 10). The scheme as it stands separately tags types of Finites, Auxiliaries and Events according to the tense that each of these elements may carry.

Table 10 – WFG scheme of 'tensed' Finite, Auxiliary and Event

WFG Tag	Meaning	Example
FNP	Finite, present tense	**has** been going
FND	Finite, past tense	**had** been going
FNF	Finite, future tense	**will** be going
AU	Auxiliary, unmodified	will **be** e-mailed
AUP	Auxiliary, present tense	is **being** sought
AUD	Auxiliary, past tense	has **been** sought
AUF	Auxiliary, future tense	is **going** to eat
EV	Event, unmodified	will have to **go**
EVP	Event, present tense	they are **going**
EVD	Event, past tense	they had **gone**

This scheme follows the SFL treatment of grammatical tense (see Halliday and Matthiessen 2004: 337-351), where the overall tense and voice of a verbal group is built through individual tense forms of the Finite, Auxiliary and Event. It is quite possible to not make distinctions between the various forms of Finites and Auxiliaries and simply use tags such as FN and AU. This would have the advantage of reducing the total number of labels in the WFG scheme, and so potentially make ML processes more efficient and accurate. However, it was felt that it was important to denote tense information in the WFG scheme since temporal information extraction from unstructured text is an important aspect of IE overall. A WFG scheme that explicitly identifies tense information has the potential to make temporal information extraction better grounded.

It is acknowledged that the tense information in the WFG scheme has considerable overlap with the tense information in the Penn Treebank POS scheme, as shown in the following table (Table 11).

DSTO-RR-0339

Table 11 – Correlates between tense information in the Penn Treebank POS scheme and the WFG scheme

WFG Tag	Penn Treeebank POS Correlate
FNP	VBZ, VBP
FND	VBD
FNF	VBZ, VBP
AU	VB
AUP	VBG
AUD	VBN
AUF	VBP
EV	VB
EVP	VBZ, VBP, VBG
EVD	VBD, VBN

Two arguments can be made to support this apparent overlap. Firstly, the POS system is geared towards the tagging of structural tense– that is, verb tense that has a morphological correlate (eg. saw has a past tense morphology). The WFG scheme, on the other hand, is intended to label tense from a non-structural point of view– that is, what tense is meant in the context of the group and clause. This is particularly the case with modal finites in verbal groups, whose morphology can be indeterminate with respect to tense, but clearly do carry some form of temporal information as the following examples show (Example 25):

Example	Modal Finite	Sense	Inferred Tense
I would go to the cinema every Tuesday	would	'I used to go'	past
		'I am willing to go'	present / future
he may go to the cinema	may	'he is allowed to go'	present
		'he will possibly go'	future

Example 25 – Inferred tense in modal Finites

40

5.1.3.2 *The labelling of* to *in non-finite verbal forms*

In the analysis of word groups and phrases, the infinitive forms of verbs (to go, to play, to e-mail) are labelled as a unitary Event. Because the WFG scheme annotates single words, it uses the non-finite marker (NFM) to label the word to in these non-finite forms. In addition, in SFL, the future progressive auxiliary going to is labelled as a single Auxiliary. In the WFG scheme, going is labelled as AUF (future tense Auxiliary) and to as NFM. This is a somewhat artificial separation, but necessary to maintain consistency in the labelling.

5.1.3.3 *The labelling of possessives*

As noted before, the morpheme denoting possession ('s) is tokenised separately from the word to which it is attached. This means that it needs to have a WFG label assigned to it. In this context, the possessive is understood to denote a relation of possession between an entity to which the possessive is attached and another entity which follows the possessive. As such, the possessive is regarded as a premodifier connecting a particular entity to a particular context, or as having a determinative modifying function (Morley 2004: 76), and so is regarded as a post-Deictic (WFG label POD). Below is an example of the labelling of the possessive (Example 26).

Nominal Group	the	state	's	energy	mess
WFG Label	DT	TH	POD	CL	TH
SFL Analysis	Deictic	Classifier		Classifier	Thing

Example 26 – WFG labelling of the possessive

5.1.3.4 *Tagging of separately tokenised punctuation and list items*

In SFL analysis, separately tokenised punctuation and list items are not formally labelled as they are not regarded as embodying or participating in grammatical structure. In the WFG scheme, these are labelled as 'out' (O).

6. The value of enrichment with WFG annotation for automated text mining

As the above sections show (Sections 4.5, 5), the WFG annotation scheme is heavily based on SFL-derived concepts and tools concerning the analysis of groups and phrases in the clause. Tagging a corpus with the WFG tagset enriches the corpus with word-level grammatical functions that are oriented to the semantics of the groups and phrases in a clause. This means that the WFG scheme provides enrichment of the corpus with grammatical information

internal to the linguistic entities that denote people, things, events and relations– the fundamental basis of IE. In this way, the WFG system provides grammatical support to recognition of relevant information from unstructured text through ML by an IE system. The rationale for expecting that WFG tagging improves IE in this way can be made on two grounds: from the machine learning perspective and the linguistic perspective.

6.1 The ability to co-train multiple layers of information

During training of an ML-based text classification system, the ML algorithm essentially determines in probabilistic fashion the strength of association between the tokens of the training corpus and the tags which are assigned to them, and (depending on the 'training window' specified in the algorithm) the likelihood of particular sequences of tags and their tokens in the corpus. The most commonly used tagging systems used in such a scenario are the POS system and the 'chunk' labelling system which identifies types of phrases and their extent.

A critical consideration in this process is the potential number of token types in the corpus which the ML algorithm may encounter. It is rather hard to determine how many different words there are in English– not least because of methodological problems as to what counts as a word– but it has been estimated that there are at least 250,000 distinct words in the English language as a whole and most likely more when specialised areas of knowledge are included (www.askoxford.com). Similar figures would be expected for other major regional or international languages. Any text corpus, particularly if it is relatively small (less than 100,000 separate tokens), is only likely to contain a fraction of this number of tokens, but even so the number of different word tokens in a text increases with text length, and a lot of these words are low or single frequency (Baayen 2001: 1-23). This can present a potential problem for ML processes, as there would be a number of tokens and possible token sequences and combinations to learn, and many kinds of associations between tags and tokens. In contrast, the number of tags in a tagging system is dramatically smaller than this, and so the number of tag sequences and tags combinations is easier for an ML system to learn.

In the Penn Treebank POS system, there are 36 different types of tags (Santorini 1990: 6-7). In the WFG tagset, referring to Table 6 above, there are 38 different tag types. Given each system has a relatively small number of tags compared to the number of token types in a corpus, a tokenised corpus can be co-trained with POS and WFG tags. In such a setup, the ML algorithm can learn the associations between POS and WFG tags, and the possible combinations and sequences of tags that are learnt are also likely to be relatively small. The result of such co-training is that the training corpus is represented by means of the sequences, combinations and associations of the WFG and POS tags, and thus a word-by-word grammatical representation of unstructured text is created during training.

	Token	POS	WFG
	One	**CD**	**NM**
	participant	**NN**	**TH**
	,	**,**	**O**
	who	**WDT**	**WHD**
	agreed	**VBD**	**EVD**
	to	**TO**	**NFM**
	speak	**VB**	**EV**
	on	**IN**	**PR**
	the	**DT**	**DT**
	condition	**NN**	**TH**
	he	**PRP**	**TH**
	not	**RB**	**PL**
	be	**VB**	**AU**
	identified	**VBN**	**EVD**
	,	**,**	**O**
	said	**VBD**	**EVD**
	the	**DT**	**DT**
	meeting	**NN**	**TH**
	appeared	**VBD**	**EVD**
	to	**TO**	**NFM**
	be	**VB**	**AU**
	geared	**VBN**	**EVD**
	toward	**IN**	**PR**
	getting	**VBG**	**EVP**
	participants	**NNS**	**TH**
	to	**TO**	**NFM**
	support	**VB**	**EV**
	Lay	**NNP**	**TH**
	`s	**POS**	**POD**
	vision	**NN**	**TH**
	and	**CC**	**CE**
	then	**CC**	**CE**
	champion	**VB**	**EV**

	it	**PRP**	**TH**
	to	**TO**	**PR**
	officials	**NNS**	**TH**
	who	**WDT**	**WHD**
	are	**VBP**	**FNP**
	trying	**VBG**	**AUP**
	to	**TO**	**NFM**
	solve	**VB**	**EV**
	the	**DT**	**DT**
	state	**NN**	**TH**
	`s	**POS**	**POD**
	energy	**NN**	**CL**
	mess	**NN**	**TH**
	.	**.**	**O**
Total Number	47	19	17

Example 27 – Example of sentence from the Enron corpus tagged with POS and WFG tags

This grammatical representation of the text involves fewer types of tags in each tagset; in case of each of the POS and WFG tags in the above example (Example 27), the number is less than half than the number of different text tokens, with the consequence that a grammatical model involving relatively few elements can be built by ML algorithms.

Such a model has three advantages. Firstly, it is a model that is simpler to build, and therefore potentially more robust, since a POS-WFG model has less dependence on specific tokens. Secondly, such a model provides support to the kind of 'structure-function' modelling that explicitly links the form of a text with what it means. Thirdly, because of the relative independence from text tokens, the generated model has the potential to be ported to different texts, and potentially texts in different domains.

It may be argued that there may be some limitations to the portability of such a learnt model to other corpora or other domains. It may indeed be possible that different corpora and different domains may contain different styles of writing, with the consequence that this kind of model derived from training may not appropriately model other kinds of texts. However, it can be said that the information that is used to construct the model is grammatical and low-level semantic information, combinations of which are likely to be predictable in relatively well-formed sentences. Hence it is argued that there is at least some scope for portability between different domains.

6.1.1 Support for future IE work

The other reason for using WFG to enrich text corpora is to support SFL-derived lexical, grammatical and semantic tagging schemata that will be developed in the future. Of particular interest is future tagging of immediate clausal constituents (corresponding to what is called 'transitivity analysis' in SFL work), as this most clearly relates to the categories of information processed in IE tasks. It would be advantageous to be able to co-train an ML system with WFG tags and any tagging scheme that grammatically labels these clausal constituents, as this would reinforce the learning of these categories and thereby make any trained models more robust in testing and deployment. Thus it is a useful exercise to determine whether WFG tags can be successfully co-trained with POS tags, so that its viability as a 'grammatical base' on which to build further semantic annotation can be determined.

6.2 Known linguistic patterns underlying entities, events, and relationships

From the SFL point of view, texts of various types differ not only with respect to their different semantics (that is, different texts have different meanings) but also differ with respect to the lexicogrammar of the clauses in the text. This is a fundamental assumption that practitioners of SFL analysis make when interpreting the results of text linguistic analysis, and in particular lexicogrammatical analysis. Particular features of the lexical and grammatical choices are therefore seen as significant to what makes a text what it is, and what the clauses of that text mean. Therefore, a SFL approach to looking at how language construes the world assumes that classes of events, entities and relationships can be characterised as having particular semantic, lexical and grammatical features that may be relatively unique to individual classes. The picture is most often not as clear cut as this: categories can have overlapping linguistic features, be they semantic, lexical and grammatical. However, a SFL analyst will often look for linguistic features that will have useful discriminative value, and this includes lexical and grammatical features. Such features can be picked out by qualitative methods, where the analyst uses his/her judgement and interpretative abilities to determine salient features. Alternatively, one can use quantitative methods by examining the frequency and distribution of features across a corpus of text or between different sets of texts, and assigning clusters or categories on this basis. It is presumed that most ML approaches to text analysis and classification aim to model such frequencies and distributions, and then predict the classification or analysis of unseen text on the basis of the probabilistic strength of association between features in previously 'learnt' texts.

However, it is important in this report to outline the 'qualitative' grammatical features that may serve as useful discriminative markers of particular ways of representing events and states of affairs. A focus on grammatical, rather than lexical, features is particularly appropriate in this regard. Firstly, a focus on grammatical features lends itself well to making distinctions between texts of different 'styles', since differences in style can be partly attributable to differences in grammatical usage.

Secondly, a focus on grammar can, at least partially, circumvent the problem of 'veiled speech', where in order to deceive a potential 'eavesdropper' on the real content of a

communication, the writer or speaker is likely to make overt lexical substitutions while leaving the underlying grammatical structure relatively unchanged. For example, if one were to produce the sentence

I will give you the flowers at your office at 2 p.m. today

this could potentially be a case of veiled speech, where the word *flowers* may act as a substitute for another entity whose identity the speaker or writer wants to conceal, for example *gun*. However, the underlying grammatical structure is relatively unchanged (Example 28).

I	will give	you	the flowers	at your office	at 2 p.m.	today
Actor	**Material Process**	**Recipient**	**Goal**	**Circum-stance: point: location**	**Circum-stance: point: temporal**	**Circum-stance: point: temporal**

I	will give	you	the gun	at your office	at 2 p.m.	today
Actor	**Material Process**	**Recipient**	**Goal**	**Circum-stance: point: location**	**Circum-stance: point: temporal**	**Circum-stance: point: temporal**

Example 28 – Veiled speech: different lexis, same grammar

An analysis of the grammar in both cases reveals that there is an act of transferring goods at a certain location and time, and this would constitute important information in a text from an intelligence point of view. A system that relies on lexical features alone may put inappropriate value on the presence of the word flowers and throw out potentially valuable information. This strategy of focusing on the grammar rather than the lexis certainly has the potential to pick up a lot of 'false positive' results if used as the sole strategy for detecting veiled talk, and as such this approach tends to 'err on the side of caution' when picking out text segments of interest. However it is hoped that this is used in conjunction with other text processing strategies to result in an appropriately constrained selection of text segments from a large corpus of material for further manual examination, without leaving out information of potential value.

6.2.1 SFL findings on the grammar of scientific and technical writing

The Enron dataset contains numerous e-mails that are about the business of the company. Many of these e-mails deal with the running of a large corporation and the myriad activities

that occur in large corporations– recruitment, further education, market reporting, management activities, initiation of new ventures, commentary on company performance, scheduling of meetings and minutes of meetings, and many other kinds of activity. The particular domain areas such talk ranges across includes economics and trading and energy markets, each with their own specialised language to describe phenomena, and through this language individual and collective decisions are made to maintain or change Enron's business activities. Such 'bureaucratic' language therefore consists of assimilating or creating specialised knowledge in one or more domains, and then using that knowledge to inform and implement company actions. Thus, in an 'official' e-mail, we may expect to see the grammatical manifestation of such language.

Much study of bureaucratic language focuses more on the semantic stratum and what the semantics says about the ideology of the organisation. However, in SFL-inspired approaches, attention is paid to the grammatical features of organisational discourse, such as nominalisation and grammatical metaphor (Iedema 2003, reviewed in Langlotz 2007). These kinds of grammatical features are similar to those of scientific writing, about which there is a considerable amount of SFL research (Halliday and Martin 1993; Martin and Veel 1998; Halliday 2004). This should not be surprising, as many large companies depend on a professionally oriented body of specialised knowledge in a particular domain, such as economics and commodity trading. Therefore, in 'official' e-mails, one might expect language whose grammar in a broad sense resembles that of scientific writing.

6.2.1.1 Multiple use of Classifier function

One particular feature that can be detected directly through WFG information is the presence of complex nominal groups with one or more Classifiers. The building of systematic knowledge in the sciences and in bureaucracies depends on that knowledge being able to be categorised in relation to other pieces of knowledge, and this necessarily involves some form of classification of entities and events. Some examples of this are shown below (Example 29).

our	Trakya	payment	issue
Deictic	**Classifier**	**Classifier**	**Thing**

LNG	regas	facility
Classifier	**Classifier**	**Thing**

the	new	Turkish	power	trading	company
			Classifier	Thing	
Deictic	**Post-Deictic**	**Classifier**	**Classifier**		**Thing**

a	work	order	solutions	company

	Classifier	Thing		Thing
	Classifier		Thing	
Deictic	**Classifier**			**Thing**

the	follow-ing	new	or	[[revised]]	ISO	operat-ing	proced-ures
		ad-jective	conjunct-ion	verb (embed-ded)			
Deictic	**Post-Deictic**	**Epithet**			**Classi-fier**	**Classi-fier**	**Thing**

Example 29 – Examples of nominal groups with use of the Classifier

It follows that many entities– abstract or concrete– are classified in this way within these 'official' texts. The WFG scheme explicitly labels such functional elements, whereas the POS scheme is likely to only label these words as an adjective or noun without any indication of their function (Example 30):

	our	Trakya	payment	issue
WFG	**DT**	**CL**	**CL**	**TH**
POS	**DT**	**NNP**	**NN**	**NN**

	LNG	regas	facility
WFG	**CL**	**CL**	**TH**
POS	**NN**	**NN**	**NN**

	the	new	Turkish	power	trading	company
WFG	**DT**	**POD**	**CL**	**CL**	**TH**	**TH**
POS	**DT**	**JJ**	**JJP**	**NN**	**NN**	**NN**

	a	work	order	solutions	company
WFG	**DT**	**CL**	**TH**	**TH**	**TH**
POS	**DT**	**NN**	**NN**	**NNS**	**NN**

	the	follow-ing	new	or	[[revised]]	ISO	operat-ing	procedures

WFG	DT	POD	AJ	CE	EVD	CL	CL	TH
POS	DT	JJ	JJ	CC	VBN	NNP	JJ	NNS

Example 30 – Comparative WFG and POS tagging of complex nominal groups

Looking for Classifiers in this way is most directly helpful in relation to the Mode of 'official' e-mails (see section 4.1), as the use of Classifiers gives valuable clues as to how information is 'packaged' in the text.

6.2.1.2 Complex nominal groups

Quite apart from the relatively frequent use of Classifiers in nominal groups, the nominal groups in scientific writing, and by extension bureaucratic writing, are complex. This means that in any one nominal group there are multiple experiential functions in the nominal group, and many of these functions are expressed by structures which may have multiple embedding and rankshift (Example 31).

a transition from GBL extranet to CGBX e-Procurement

a	transition	from	GBL	extranet	to	CGBX	e-Procurement
DT	TH	PR	CL	TH	PR	TH	TH
Deictic	Thing	Qualifier			Qualifier		

sufficient supplies at or below its announced purchase price ceiling

sufficient	supplies	at	or	be-low	its	an-nounced	pur-chase	price	ceil-ing
POD	TH	PR	CE	PR	DT	POD	CL	TH	TH
		preposition group			nominal group				
							Classi-fier	Thing	
					De-ictic	Post-Deictic	Classifier		Thing
		Predicator			Complement				
Deictic	Thing	Qualifier							

49

consequences to the market of a return to command and control

con-se-quen-ces	to	the	mark-et	of	a	re-turn	to	com-man-d	and	con-trol
TH	**PR**	**DT**	**TH**	**PR**	**DT**	**TH**	**PR**	**TH**	**CE**	**TH**
	pre-posi-tion	nominal group		pre-posi-tion	nominal group		pre-posi-tion	nominal group complex		
							Predi-cator	**Complement**		
					De-ic-tic	**Thing**	**Qualifier**			
	Predi-cator	**Comple-ment**		**Predi-cator**	**Complement**					
Thing	**Qualifier**			**Qualifier**						

Example 31 – Examples of complex nominal groups

At the level of WFG tagging, this complexity is only partially conveyed, because the WFG scheme (like other annotation schemes) applies labels only once to each word, effectively 'ignoring' the multiple levels of rank shift involved in such complex groups. However, if WFG labelling is co-trained with other kinds of grammatical information, such as chunk labelling or a scheme that labels the functional constituents of the clause (such as the functions found in Table 4, section 4.4.2 above), it is possible that a single chunk or clause function label becomes associated with multiple WFG tags, some of which (such as DT) appear to be repeated within the group or phrase. Therefore it is possible that such associations can be learnt in order to pick up on these complex groups as makers of 'official' e-mails.

The reason why these groups are so complex is that in much scientific and bureaucratic talk there is a priority to condense information as much as possible, to make that information systematic and ordered, and to 'objectivise' and make it impersonal (Martin 1993a; Martin 1993b). It often takes several clauses to unpack such groups adequately (Example 32):

Original	Unpacked
a transition from GBL extranet to CGBX e-Procurement	Things will move in stages from the GBL extranet into CGBX e-Procurement.
sufficient supplies at or below its announced purchase price ceiling	Things will be supplied sufficiently, and they will be bought at the same or less than the amount which was announced. They should not be bought from a greater amount than this.
consequences to the market of a return to command and control	People used to buy and sell while they were being directed and controlled. The market will be affected if this happens again.

Example 32 – The unpacking of complex nominal groups

There are regular relationships between the 'packed' and 'unpacked' forms, but it is beyond the scope of this report to explore these in detail (for a summary of them for scientific writing, refer to (Halliday 2004: 41)). But it is useful to know that such groups are complex, and that it is useful to know what kinds of information can be extracted from these structures. Such complex groups are a useful indicator of scientific or professionally-oriented texts, and this complexity can be captured through the use of the WFG tags and their association with POS and chunk labelling.

7. Automatic tagging of the Enron corpus with WFG tags

In order for the WFG system to be useful in a computational setting, it must be determined whether a corpus can be feasibly annotated both manually and automatically.

7.1 Manual tagging of the training corpus

Properly conducted, a proposed linguistic tagging scheme involves a team of researchers who develop a provisional tagging scheme based on a linguistic theory. Each member applies the tagset manually, with feedback between the different members of the team to refine the tagset further, to check for errors and to determine and improve inter-annotator agreement. Refinement of the tagset also needs to achieve the appropriate trade-off between the richness of the category set and the optimal tagset size for more effective ML processing. The aim of this activity is to produce a tagset which reflects the categories of the linguistic theory on which it is based, which can be applied to text consistently, which ideally does not require

extended training of annotators, which results in sufficiently high inter-annotator agreement, and which is relatively ready to be incorporated into a training corpus.

Limited resources meant that the current WFG annotation of the corpus has not gone through all of the required quality control processes in the development of the tagset and the tagging of the corpus. This is largely due to the fact that application of the WFG tagset requires some degree of linguistic expertise, and that there was only one person available in the team to do this. Despite this limitation, the same feedback cycle of tagset development, corpus tagging and tagset refinement was followed to produce the most consistent and useful tagset possible. It is possible to say that the tagset system has been developed and applied with a high degree of consistency that will facilitate a successful ML process.

The lack of availability of a team of annotators in the quality control process is an important issue. Ideally, one would want a linguistic tagging scheme and approach that reflects not just the judgement of one person, but that of a number of people, given that language and its use is not just an individual psychological phenomenon, but a social one as well. This means that a linguistic annotation scheme has to pick out the kinds of language features that a number of speakers of the language can recognise consistently, and not just the potentially idiosyncratic judgements of a single person; this would result in a scheme that labels the features of a language which are readily recognised by at least a subsection of a language's speech community. Automated tagging is ideally assessed against such a 'community' standard, and one should not necessarily aim for a system that labels perfectly consistently, but rather one whose output matches the same variabilities of performance as inter-annotator variation.

7.2 Preparation of the training and testing corpora

All stages of the preparation of the training and testing corpora were undertaken by a single person with a full-time salaried position in the research team at the DSTO. The training corpus consists of a mixture of e-mails from the Enron dataset and newspaper articles, with a total of 69751 separate tokens with 4882 line breaks between separate sentences. The e-mails in the corpus were prepared by removal of any metadata and header information, and any formatting such as strings of repeating characters around signature blocks (but preserving the content of the signature block itself). Newspaper articles also had their formatting removed. The corpus was then tokenised according to the principles set out in section 5.1 using an automated tagger developed in-house, with manual checking of the output. Punctuation with variable form (quotation marks and apostrophes) were manually normalised to a single variant.

The texts that were used in the construction of the training corpus are set out in Table 12.

Table 12 – List of contents of training corpus

Enron Dataset	Newspaper Articles
\lay-k\inbox\19,29,41, 51, 245, 264, 285, 413,550, 864	albrechtsen_happiness
\skilling-j\sent\2, 39, 68, 115, 129, 350, 358, 364, 395, 412, 421, 455, 456, 467, 581	colebatch_trade
\lavorato-j\sent\10, 194, 302, 332	gottliebsen_woodside
\dasovich-j\edwards_air_force\1-14	guardian_education
\badeer-r\california\1-70	australian_northkoreabomb
\causholli-m\sent_items\1-127, 149-230, 236-249, 273, 292, 352	periclesspeech.translation
	thetimes_shanghai
	times_anorexia
	times_englandcricket
	economist_eusummit
	economist_schwarzenegger
	globeandmail_lebanon
	guardian_zimbabwe
	theaust_arrowhead
	uktimes_mob

The corpus so formulated resulted in a substantial majority of tokens deriving from the e-mails (estimated about 70%). The reason for interspersing the Enron texts with newspaper articles was so that the training model was sufficiently generalisable. That is, it was hoped that with this mixture the ML-generated model would not overfit the Enron dataset, but would achieve accuracy in tagging well-formed non-email texts, thus achieving at least some portability between domains.

The testing corpus also consisted of a combination of Enron e-mails and newspaper articles, with a total of 3035 tokens and 163 line breaks between sentences, resulting in about 50% of tokens coming from the Enron dataset. The preprocessing of this corpus was the same as for the training corpus. The result is a testing corpus that again has a heavy e-mail component but allows one to see whether training has been generalised to other domains.

The list of texts in the testing corpus is below (Table 13).

Table 13 – List of contents of testing corpus

Enron Dataset	Newspaper Articles
carson_18	scannerdarkly_review
carson_72	
lay_mayblackmore	
shults_tradingsimulation	

A validation corpus constructed entirely of newspaper articles (list of articles in Table 14) was also constructed for testing in order to determine to what degree system accuracy was maintained during testing on non e-mail texts.

Table 14 – List of contents of validation corpus

Validation Corpus Texts
ABC090403
ABC140403
Australian 170304
Australian 180903
Australian 240403
Australian 311002

7.3 Machine learning strategy

Since WFG annotation provides information which is the functional counterpart of POS information, both the POS and the WFG tagsets were used to label each token in the training and testing corpus, and so both these sets of tags were co-trained with the tokens to create the training model. The reason for the creation of this model was so that one could determine whether the WFG tagset could be automatically labelled with a high degree of accuracy.

The machine learning algorithm employed was that of 'Conditional Random Fields' (CRF), which is essentially a probabilistic model for sequential data (Lafferty, McCallum et al. 2001; Moens 2006: 114-117). This was implemented by the CRF++ toolkit (Kudo 2006) which was used to build the training models and produce predicted values in testing for the particular tagset being tested. This allowed an accuracy count for tagging to be determined, calculated by the number of tokens for which predicted value matched the actual value, divided by the total number of tokens, expressed as a percentage.

The CRF++ toolkit makes use of a template which allows the user to specify the 'window' of tokens and tags around the current token input and associated tags, in the form of n-grams

(Figure 7). It also allows the construction of n-grams to enable the algorithm to learn on the basis of sequences of tokens and tags and associations between tokens and different tag types.

```
U02:%x[-2,0]
U03:%x[-1,0]
U04:%x[0,0]          token unigrams
U05:%x[1,0]
U06:%x[2,0]

U11:%x[-2,0]/%x[-1,0]
U12:%x[-1,0]/%x[0,0]     bigram sequences of tokens
U13:%x[0,0]/%x[1,0]
U14:%x[1,0]/%x[2,0]

U24:%x[-2,1]
U25:%x[-1,1]
U26:%x[0,1]          tag₁ unigrams
U27:%x[1,1]
U28:%x[2,1]

U32:%x[-2,1]/%x[-1,1]
U33:%x[-1,1]/%x[0,1]
U34:%x[0,1]/%x[1,1]      bigram sequences of tag₁
U35:%x[1,1]/%x[2,1]

U47:%x[-2,1]/%x[-1,1]/%x[0,1]
U48:%x[-1,1]/%x[0,1]/%x[1,1]    trigram sequences of tag₁
U49:%x[0,1]/%x[1,1]/%x[2,1]

U112:%x[-2,1]/%x[-2,0]
U113:%x[-1,1]/%x[-1,0]   cross-correlation
U114:%x[0,1]/%x[0,0]     between token and tag₁
U115:%x[1,1]/%x[1,0]
U116:%x[2,1]/%x[2,0]
```
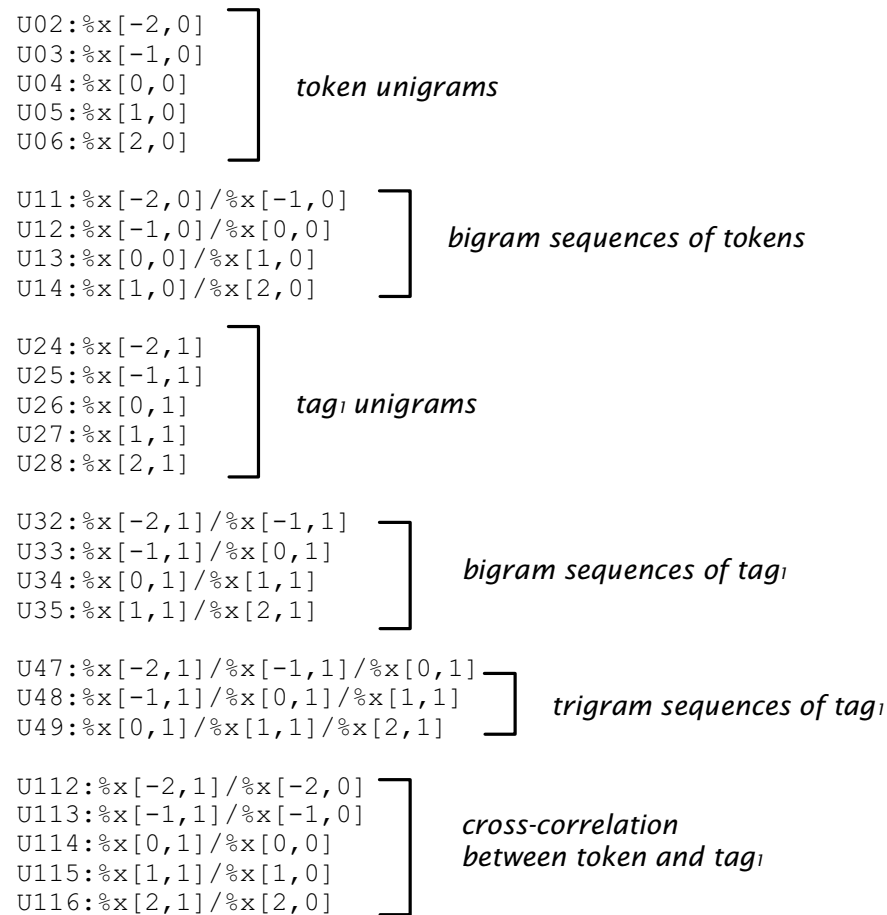
Figure 7 – Sample template (modelled on templates provided in Kudo 2006) used in CRF++ demonstrating types of n-grams (window size -2 to +2 with cross-correlation, co-training tokens in column 0 and POS tags in column 1 to predict WFG tags in column 2)

In this way, several training runs were performed, where the window size was altered between 2- and 4-grams around a current n-gram with both bigrams and trigrams representing sequences of tokens and tags within the window. This was done both with and without cross- correlations between tokens and POS tag types. All of the training models that resulted were tested with both the testing corpus described above, and a validation corpus which consisted entirely of newspaper articles. The predicted results from testing were tested for their accuracy.

7.4 Results of testing

7.4.1 Training and testing with POS and WFG tags

The results of the accuracy testing for WFG labelling were as follows (Table 15):

Table 15 – Results of co-training POS and WFG tags

Co-trained tag	Tag trained	Template window size	Cross-correl-ation	Pro-cessor threads	Training time (s)	Training Model Size (KB)	Testing corpus	Accuracy %
pos	wfg	-2 to +2	no	2	1647.91	49204	3k_wfg	94.63
pos	wfg	-2 to +2	yes	2	1842.51	60521	3k_wfg	94.63
pos	wfg	-3 to +3	no	2	2003.94	69604	3k_wfg	93.97
pos	wfg	-2 to +2	yes	2	1842.51	60521	validation-test_wfg	94.7

The relatively high accuracy rate (in the low to mid 90's in percentage terms) demonstrates that the WFG scheme can be machine learnt successfully, and in particular in co-training with the POS tags. It also appears that accuracy is improved by almost 0.5% with a smaller training window, with no significant difference observed with cross-column correlation, indicating that WFG training using CRF is more dependent on learning sequences of tags rather than their relationships to the POS system. It should also be observed that the system performed slightly better with the validation testing corpus, which may indicate that the system is more biased towards the labelling of well-formed text. However, larger datasets need to be constructed in order to bear out the above findings.

7.4.2 Co-training with chunk labelling

It is also important to see whether co-training with POS and WFG improves the accuracy of other kinds of grammatical tagging, such as chunk labelling. This was also tested using the testing and validation corpora. The results of examining the accuracy of chunk labelling with and without WFG co-training were as follows (Table 16):

Table 16 – Labelling of chunks with and without WFG co-training

Co-trained tag	Tag trained	Template window size	Cross-correl-ation	Proces-sor threads	Training time (s)	Training Model Size (KB)	Testing corpus	Accuracy %
pos/wfg	chunk	-2 to +2	no	2	466.34	29741	3k_chunk	96.7
pos	chunk	-2 to +2	no	2	370.39	29146	3k_poschk	96.04
pos/wfg	chunk	-2 to +2	yes	2	472.02	38486	3k_chunk	96.8

pos	chunk	-2 to +2	yes	2	404.58	35962	3k_poschk	95.75
pos/wfg	chunk	-3 to +3	no	2	620.11	43408	3k_chunk	96.64
pos	chunk	-3 to +3	no	2	450.75	40452	3k_poschk	95.42
pos/wfg	chunk	-3 to +3	yes	2	674.99	54624	3k_chunk	96.74
pos	chunk	-3 to +3	yes	2	526.81	49724	3k_poschk	95.48
pos/wfg	chunk	-4 to +4	no	2	696.04	56439	3k_chunk	96.6
pos	chunk	-4 to +4	no	2	591.76	52434	3k_poschk	95.35
pos/wfg	chunk	-4 to +4	yes	2	778.82	71681	3k_chunk	96.57
pos	chunk	-4 to +4	yes	2	660.85	64459	3k_poschk	95.32
pos	chunk	-2 to +2	no	2	370.39	29146	validation-test_ poschk	94.23
pos/wfg	chunk	-2 to +2	yes	2	472.02	38486	validation-test_ chunk	94.41
pos	chunk	-2 to +2	yes	2	404.58	35962	validation-test_ poschk	94.13
pos/wfg	chunk	-3 to +3	no	2	620.11	43408	validation-test_ chunk	no result*
pos	chunk	-3 to +3	no	2	450.75	40452	validation-test_ poschk	93.85
pos/wfg	chunk	-3 to +3	yes	2	674.99	54624	validation-test_ chunk	94.32
pos	chunk	-3 to +3	yes	2	526.81	49724	validation-test_ poschk	93.79
pos	chunk	-4 to +4	no	2	591.76	52434	validation-test_ poschk	93.92
pos/wfg	chunk	-4 to +4	yes	2	778.82	71681	validation-test_ chunk	94.38
pos	chunk	-4 to +4	yes	2	660.85	64459	validation-test_ poschk	93.85

(* CRF++ subject to repeated crashing– cause unknown)

In general, when WFG tags are co-trained with chunk tags, it results in a small but significant improvement in the accuracy of chunk labelling over an already high accuracy count when

WFG tags are not co-trained. The presence of cross-correlation improves the accuracy of chunk labelling, and increases the effect of co-training with WFG tags. Testing on the validation corpus resulted in a similar pattern, but the accuracy scores were overall depressed by 2%. In all cases, a smaller window size for training improves accuracy. These results indicate that the chunk patterns may be genre-dependent, but that WFG provides significant grammatical information for the identification and categorisation of groups and phrases (the basic units on which much of IE depends). Furthermore, ML of chunk patterns is considerably supported by taking into account correlations between the chunk labelling and WFG information.

8. Conclusion

The report has aimed to demonstrate that a certain part of systemic-functional grammar– the grammar of groups and phrases in clauses, and the grammatical role of single words seen from a functional viewpoint– can be implemented in an annotation scheme to usefully support the automatic identification of elements of unstructured text that are important and relevant for text classification and entity and event detection. It does so for two reasons. Firstly, the scheme provides relatively 'generic' semantic information that is the functional counterpart of the structural information contained in parts-of-speech, and which acts as a 'semantic bedrock' on which more complex linguistic entities are built. Secondly, the generic nature of the information increases the chance of portability of the system between domains, which is a desired property of an IE system in defence and intelligence contexts.

Some recommendations can be made with respect both to the WFG scheme, and to the use of SFL within IE more generally. There is scope for the WFG scheme to be refined and improved, in terms of the number of categories in the tagset and how the tagset is manually applied. The number of categories can be further adjusted with a view to reducing their number in order to make ML strategies more efficient. In this report, the number of categories is generous, because it was felt that some of them denoted information that would be useful for future work, such as the categorisation of immediate clause constituents and temporal information. Given that the development of the tagset was an individual effort, a team approach and peer review are important in refining the scheme further, ensuring consistency in application, and developing a workflow for manual annotation of unstructured text in larger volumes.

As mentioned before, the WFG scheme covers only a small part of the grammar of clauses. Other tagsets to cover clause-level grammar need to be developed for the purposes of IE, as these categories very directly reflect the semantic entities that IE aims to extract from unstructured text– entities, events and relationships. This has been attempted before (see, for instance, O'Donnell 1994), but automatic parsing for these categories has often been slow. This could be attributable to the large number of categories which can give rise to an extraordinarily large number of combinatorial possibilities in the ML process, thus reducing ML efficiency (Bateman 2006, personal communication). Future work needs to be done to determine how best to implement clause-level grammar into a workable tagset scheme that can be readily learnt by an ML system. This is in addition to the variety of semantic

phenomena described by SFL that has great applicability in IE and IR, such as appraisal (a form of sentiment analysis).

9. Bibliography

Bateman, J. (2006). University of Bremen.

Baayen, R. H. (2001). Word frequency distributions. Dordrecht, Boston, London; Kluwer Academic Publishers.

Bohnet, B., S. Klatt, et al. (2003). A bootstrapping approach to automatic annotation of functional information to adjectives with an application to German. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2003).

Fawcett, R. (2000). A theory of syntax for systemic functional linguistics. Amsterdam, Philadelphia, PA, John Benjamins.

Halliday, M. A. K. (2004). Language and knowledge: the 'unpacking' of text. The language of science. M. A. K. Halliday and J. J. Webster. London, New York, Continuum.

Halliday, M. A. K. (2004). The language of science. London ; New York, Continuum.

Halliday, M. A. K. and R. Hasan (1989). Language, context, and text : aspects of language in a social-semiotic perspective. Geelong, Vic., Deakin University Press.

Halliday, M. A. K. and J. R. Martin (1993). Writing science : literacy and discursive power. Pittsburgh, University of Pittsburgh Press.

Halliday, M. A. K. and C. M. I. M. Matthiessen (2004). An introduction to functional grammar. London, Arnold.

Hasan, R. (1996). The grammarian's dream : lexis as most delicate grammar. Ways of saying, ways of meaning : selected papers of Ruqaiya Hasan. G. Williams, D. Butt and C. Cloran. London ; New York, Cassell: Open linguistics series.: 73-103.

Hasan, R. (1999). Speaking with reference to context. Text and context in functional linguistics. M. Ghadessy. Amsterdam ; Philadelphia, J. Benjamins: 219-328.

Honnibal, M. (2004). Converting the Penn Treebank to systemic functional grammar. Proceedings of the Australian Language Technology Workshop, Macquarie University, Sydney, Australian Speech Science & Technology Association Inc.

Honnibal, M. (2004). Design, creation and use of a systemic functional grammar annotated corpus. Department of Linguistics, Division of Linguistics and Psychology, Macquarie University. **BA (Hons) Thesis**.

Iedema, R. (2003). Discourses of post-bureaucratic organization. Amsterdam; Philadelphia, PA, John Benjamins.

Kudo, T. (2006). "CRF++: yet another CRF toolkit." Retrieved 8 September 2006, from http://chasen.org/~taku/software/CRF++/.

Lafferty, J., A. McCallum, et al. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, San Francisco, Morgan Kaufman.

Langlotz, A. (2007). "Book review: Discourses of post-bureaucratic organization." Discourse & Communication **1**(1): 115-117.

Levinson, S. C. (1983). Pragmatics. Cambridge, Cambridge University Press.

Manning, C. D. and H. Schutze (1999). Foundations of statistical natural language processing. Cambridge, Mass., MIT Press.

Martin, J. R. (1992). English text : system and structure. Philadelphia, John Benjamins Pub. Co.

Martin, J. R. (1993). Life as a noun: arresting the universe in science and humanities. Writing science : literacy and discursive power. M. A. K. Halliday and J. R. Martin. Pittsburgh, University of Pittsburgh Press: xiii, 283.

Martin, J. R. (1993). Technicality and abstraction: language for the creation of specialised texts. Writing science : literacy and discursive power. M. A. K. Halliday and J. R. Martin. Pittsburgh, University of Pittsburgh Press.

Martin, J. R. and D. Rose (2003). Working with discourse : meaning beyond the clause. London, Continuum.

Martin, J. R. and R. Veel (1998). Reading science : critical and functional perspectives on discourses of science. London, New York, Routledge.

Martin, J. R. and P. R. R. White (2005). The language of evaluation: appraisal in English. London, Palgrave.

Moens, M.-F. (2006). Information extraction: algorithms and prospects in a retrieval context. Dordrecht, Springer.

Morley, G. D. (2004). Explorations in functional syntax : a new framework for lexicogrammatical analysis. Oakville, CT, Equinox Publishing.

Munro, R. (2004). Towards the computational inference and application of a functional grammar. Department of English and School of Information Technologies. Sydney, University of Sydney. **Bachelor of Arts / Bachelor of Science (Hons) Thesis**.

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

O'Donnell, M. and J. Bateman (2005). SFL in computational contexts: a contemporary history. Continuing discourse on language: a functional perspective. R. Hasan, C. M. I. M. Matthiessen and J. Webster. London; Oakville, CT, Equinox. **Volume 1:** 343-382.

Santorini, B. (1990). "Part-Of-Speech Tagging Guidelines for the Penn Treebank Project." 3rd revision. Retrieved 7 September, 2006, from http://www.cis.upenn.edu/~treebank/.

Souter, D. C. (1996). A corpus-trained parser parser for systemic-functional syntax. School of Computer Studies, University of Leeds. **PhD Thesis**.

Weiss, S. M., I. Nitin, et al. (2005). Text mining: predictive methods for analyzing unstructured information. New York, Springer Science+Business Media.

Whitelaw, C. and S. Argamon (2004). Systemic functional features in stylistic text classification. AAAI Fall Symposium on Style and Meaning in Language, Art, Music and Design.

Whitelaw, C., N. Garg, et al. (2005). <u>Using appraisal groups for sentiment analysis</u>. 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany.

Whitelaw, C., M. Herke-Couchman, et al. (2004). <u>Identifying interpersonal distance using systemic features</u>. Proceedings of AAAI Fall Workshop on Exploring Attitude and Affect in Text: Theories and Applications, AAAI Press.

Whitelaw, C. and J. Patrick (2004). <u>Selecting systemic features for text classification</u>. Proceedings of Australian Language Technology Workshop, Macquarie University, Sydney, Australian Speech Science & Technology Association Inc.

www.askoxford.com. (22 September 2007). "How many words are there in the English language?" Retrieved 14 November, 2007, from http://www.askoxford.com/asktheexperts/faq/aboutenglish/numberwords?view=uk.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA		1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE The Use of Systemic-Functional Linguistics in Automated Text Mining	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document　　　　　(U) Title　　　　　　　(U) Abstract　　　　　(U)

4. AUTHOR(S) Astika Kappagoda	5. CORPORATE AUTHOR DSTO Defence Science and Technology Organisation PO Box 1500 Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER DSTO-RR-0339	6b. AR NUMBER AR-014-419	6c. TYPE OF REPORT Research Report	7. DOCUMENT DATE March 2009

8. FILE NUMBER 2009/1016253/1	9. TASK NUMBER INT 07/020	10. TASK SPONSOR ASCP and EXEC DIR CTSTC	11. NO. OF PAGES 82	12. NO. OF REFERENCES 38

13. URL on the World Wide Web http://www.dsto.defence.gov.au/corporate/reports/DSTO-RR-0339.pdf	14. RELEASE AUTHORITY Chief, Command, Control, Communications and Intelligence Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS	No

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.htm

systemic-functional linguistics, text mining, information extraction, machine learning, text categorisation

19. ABSTRACT
Systemic-functional linguistics is a linguistic framework for the analysis of grammatical and semantic information in text, with a potential role in automated text mining. This report outlines essential features of the theory, its application in computational work, and the rationale for use in automated text mining, and develops a grammatical annotation scheme- word functions- to enrich a mixed text corpus of newspaper articles and e-mails, for machine learning of semantically-oriented grammatical patterns. Testing demonstrates high accuracy in predicting word functions in unseen text in co-training with other grammatical information, providing the basis for further grammatical and semantic text processing.