Genre Identification

June Murphy

B.A. (Mod.) Computer Science, Linguistics and a Language
Final Year Project March 2010
Supervisor: Dr. Zachary Lyons

School of Computer Science and Statistics

O’Reilly Institute, Trinity College, Dublin 2, Ireland
DECLARATION

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or at any other university.

___________________________  ______________________
Name:                                      Date:


PERMISSION TO LEND

I agree that the library and other agents of the college may lend or copy this dissertation upon request.

____________________________________
Name:

____________________________________
Date:
Acknowledgements:

I would like to thank all of the teaching and other staff and my fellow students at Trinity College who have contributed to such a large extent in making the past few years such an interesting experience.

In particular I would like to thank my fellow CSLL class mates and the course director Carl Vogel for all their help and support. I would also like to thank my supervisor Zach Lyons for his help and encouragement in making this dissertation possible.

I would also like to thank my friends who have all been so patient and supportive during this time.

With special thanks to my parents and family for their encouragement, for believing in me and for always being there for me.
Quote:

High thoughts must have high language – Aristophanes (450 BC – 388 BC)
# Contents

List of Tables and Figures 6

Abstract 7

Chapter 1.
1.1 Introduction 8
1.2 Discourse Analysis 9

Chapter 2.
2.1 Defining Genre 11
2.2 Definitions 16
2.3 Register Variation 18

Chapter 3.
3.1 Corpus linguistics 21
3.2 Various methodologies 23
3.3 Proposed methodology 24

Chapter 4.
4.1 Qualitative Analysis 27
4.2 Quantitative Analysis 29

Chapter 5.
5.1 Conclusion 51
5.2 Future Work 54

Bibliography 56

Appendices:
I Information Regarding the Corpora 60
II Analysis Software 62
III Tagset 67
IV Custom Programme Code 69
V Linguistic Features in Detail 88
VI Weighted Values 91
List of Tables and Figures

Table 1. List of the top fifty keywords 31
Table 2. List of each part-of-speech and the number of occurrences 34
Table 3. Ratios of features between the corpora 36
Table 4. Distribution of the different verb forms among the corpora 38
Table 5. List of the incidence of place adverbials 41
Table 6. List of the incidence of time adverbials 42
Table 7. List of each punctuation mark and frequency of occurrence 47
Figure 1. Distribution of prepositions among the corpora 44
Figure 2. Distribution of nominalisations among the corpora 45
Figure 3a. Distribution of punctuation marks for English 47
Figure 3b. Distribution of punctuation marks for Geography 48
Figure 3c. Distribution of punctuation marks for Mathematics 48
Figure 4. Distribution of all of the features for all of the corpora 50
Abstract

This study set out to identify characteristics of different genres of school-based corpora, specifically the three post-primary curricular subjects, English, Geography and Mathematics as presented in subject textbooks. The identification of genre was problematised by the multitude of definitions of what constitutes genre in the existing literature and this led the researcher to devote a lot of time and space to defining terms used in this research in order to be consistent in approach. It also meant that a number of different research methodologies and instruments had to be investigated in order to find those suitable for the task in hand.

It was decided to perform a discourse analysis of the corpora compiled from the post-primary curricular subjects, English, Geography and Mathematics as presented in subject textbooks. Both qualitative and quantitative means were used to analyse the data. The former involved an appraisal of the various corpora by hand (and eye) while the latter involved the development of a custom programme and the exploration and use of various other tools that were freely available in order to carry out the analysis. Analyses of the data led to certain conclusions being made. The first is that while school-based genres do share many common features they also exhibit very particular traits. It was originally thought that it might be possible to place the different subject areas along a continuum, with Mathematics and English at either extreme and Geography at some point in between. Initially, from an impressionistic viewpoint and also following the qualitative analysis, this seemed to be a possibility. However, during the quantitative analysis of the linguistic features that had been extracted from the texts it was discovered that the corpora for each of the three subjects varied in different ways regarding the different features being examined. This led to a further observation that automatic means of genre identification would benefit from hand-editing on the part of the researcher. Consideration was also given to the impact these linguistic features might have on education, both for teachers and students alike.
Chapter 1

1.1 Introduction:

In a previous project I set out to identify and demonstrate that there is no single feature that differentiates between spoken and written language, in French. Rather, it was shown that linguistic features within a discourse change according to whatever register is being used. Further, taking a typical spoken genre such as face-to-face conversation, and a typical written genre such as academic prose, it was shown that all other genres lie somewhere in between these two extremes as a continuum rather than as discrete categories. These findings echo the extensive work done by Biber (1988, 1995) with the English language.

My findings led to the conclusion that it might be interesting during my final year project to delve deeper into the whole area of corpus linguistics as a means of discourse analysis by focusing on genre itself, and more specifically, on its identification. This present study sets out to investigate the characteristics of different types of school-based genres, and then to consider the impact these might have on education. The exact genres it focuses on are English, Geography and Mathematics. The textbooks used were all taken from the current Irish Junior Certificate syllabus. These are taken from post-primary syllabi, and have been scanned from textbooks as part of the English Language Support Programme, which provides online language support primarily for post-primary students whose first language is not English. This programme is itself part of the Trinity Immigration Initiative, a research programme on diversity, integration and policy. This corpus is unique in that it consists of some of the actual documents currently in use in the Irish education system. The complete corpus presently stands at 5.6 million words and is still growing as more texts are added.

Furthermore, the meaning, origins, development and current status of genre will be investigated plus an exposition of the key terms and definitions upon which the study will rely. Previous relevant research will be considered, this latter also in respect of investigating the various methodologies available for the automatic identification of genre. Finally a qualitative and a quantitative analysis of several known genres will be
undertaken in order to identify the linguistic features, as well as any other relevant features, typical of each. This will be done with reference to topical and non-topical descriptors including the register, vocabulary and key-words used, and in fact the whole bundle of features typical of each will be considered. Since various methods will be used to carry out the analysis, a description of the methods available and of those used will be given as well as an evaluation of the methods themselves.

1.2 Discourse Analysis:

Discourse analysis provides a means of analysing discourse according to its constituent parts. The discourse itself can be any type of occurrence of language, whether spoken, written or signed and whether fabricated or occurring naturally. The constituent parts might pertain to linguistic features as varied as presentation style, syntactic structure, grammatical features, speech acts, intonation units, etc, with the type, level and depth of the analysis being driven by the requirements of the researcher. Corpus linguistics, text linguistics and computational linguistics are all part of discourse analysis. “Corpus linguistics studies are generally considered to be a type of discourse analysis because they describe the use of linguistic forms in context.” (Biber 2007:2).

This present study will undertake a discourse analysis of several texts taken from a particular corpus and, as mentioned earlier, this analysis will make use of both qualitative and quantitative methods. It must be noted however that how a corpus is interrogated and what it is compared to will determine the results.

In their definition of discourse analysis, Schiffrin et al (2001) identified three general categories as being:

1) the study of language use,

2) the study of linguistic use ‘beyond the sentence’, and

3) the study of social practices and ideological assumptions that are associated with language and/or communication.
From that point of view the quantitative part of the study will fall into the first category since it concerns typical distributional patterns as identified by computational means; while the qualitative part of it will fall into the second category since it is concerned with linguistic description based on detailed analysis of individual texts.

This corpus itself is unique in that it consists of some of the actual documents currently in use in the Irish education system. The complete corpus presently stands at 5.6 million words and is still growing as more texts are added. These are taken from post-primary syllabi, and have been scanned from textbooks as part of the English Language Support Programme, which provides online language support primarily for post-primary students whose first language is not English. This programme is itself part of the Trinity Immigration Initiative, a research programme on diversity, integration and policy (See Appendix I for a full description of the corpus).
Chapter 2

2.1 Defining Genre:

The definition of genre turns out to be something quite difficult to pin down. ‘Form, content and purpose’ seems to be a good general description. Indeed there are many definitions of genre. One source cites genre as "a class or category of artistic endeavour having a particular form, content, technique, or the like." According to the Oxford English Dictionary, genre refers to "a particular style or category of works of art; especially a type of literary work characterized by a particular form, style, or purpose."

In his book “Genre Analysis” (1990), John Swales claims genre to be “[a]…distinctive category of discourse of any type, spoken or written…” (Swales 1990:33). It remains to be explored further what it is about the discourse that makes it distinctive and identifiable as such a particular category. He also added that: “If there were only minor differences among genres, there would be little need for genre analysis as a theoretical activity separable from discourse analysis.” (1990:61).

Discourse itself serves a social function and is used for effect in order to achieve rhetorical goals. The identification of genre as being a relevant factor in discourse can be traced back to its use in ancient Greek rhetoric, where different styles serving different purposes were identified, along with sub-genres of each. It was only when quantitative studies began, which analysed language with the explicit purpose of identifying features, that these features started to become associated with functional language varieties, or more specifically, with registers. However, where linguistics is concerned, register has always played an important role. And it is only in recent times that linguists are being forced to look towards genre as it becomes more and more important in terms of text structure studies. (Swales 1990).

Traditionally, genre studies focused on particular works, these being mainly literary, which were analysed individually in order to identify the linguistic features which made

---

them members of a particular genre, and perhaps in that analysis itself, the genre was also identified and defined. Swales (1990) makes a further observation that genres are the property of discourse communities serving a socio-rhetorical purpose based on common goals. Its members possess both personal interest and familiarity with the genres to be used in furthering these goals, and therefore they are no longer the property of the individual alone. The genres are thus considered to be classes of communicative events, consisting of texts and their processing (coding/decoding) procedures, with these being bound by their communicative purpose. According to Hymes, “genres often coincide with speech events”. (Hymes 1974:61.) This notion that genre is identified with and relative to concerned communities seems to be quite a commonly held view. For example, the idea of conceptualisation of a discourse community is also expounded by Swales (1990), with these communities (i) containing common public goals, (ii) encouraging inter-communication among members, (iii) providing information and feedback, (iv) utilising one or more genres, (v) having an acquired lexis, and (vi) whose members have relevant content and expertise. He also mentions Herzberg's 'cluster of ideas' with (i) language use as a form of behaviour with (ii) discourse maintaining and extending the group's knowledge and (iii) discourse being epistemic or constitutive of the groups knowledge. (Herzberg 1986). The difference between a speech community and a discourse community is that in the former the linguistic behaviour is based on social factors while for the latter it has more to do with function, for example with a special interest group.

Simply put, discourse is classified according to certain linguistic features it displays; these being affected by the linguistic function of the discourse, and this determines the genre to which it belongs. While genre may serve a classificatory function, Swales also suggests identifying particular genres as 'ideal types' and then measuring how other genres deviate from this norm, in much the same way as Biber did with register variation. Continual transgressions of this norm is in fact how 'new' genres are created. Genre can also serve a clarificatory purpose. Fowler (1982) states that genre provides a communication system which implies guidelines for writers and readers alike. This means that writers know what is required when they are writing for a particular genre, and likewise readers know what to expect when they choose from such a genre.
According to Fowler, this enables clarity of purpose. However, the distinction between genres is not always as clear-cut and well-defined as one might like, and there can be disagreement as to what genre a work belongs, and often hybrids appear which take the particular traits of several genres and combine them to form another. Though as already implied, this could be merely a manifestation of genre evolution. So, while the term genre had its roots in rhetoric and then in literature, there is a much wider variety of classifications available these days, from 'press conference' to 'memo' to 'blog'. But care needs to be taken that the term is not applied according to some arbitrary formula, a concern these days as automatic means of genre identification are being developed, which need to be based on sturdy definitions and descriptions.

Another approach sees genre as a form which remains unchanged while it is its role in society that changes. Martin (1985) remarks that: “Genres are how things get done when language is used to accomplish them. They range from literary to far from literary forms: poems, narratives, expositions, lectures, seminars, recipes, manuals, appointment making, service encounters, news broadcasts and so on” (Martin 1985:250). He goes on to say that the way the register variables of field, tenor and mode can be combined is constrained by genre and that they are a system for accomplishing social purposes.

According to an International Pragmatics Association "survey of research in progress" (Nuyts 1988), language varieties can be divided into register, dialect, argot, slang, and jargon. Register is further divided into areas such as: aviation language, journalese, legalise, religious language, scientific language, technical language, mythical language, etc. It also differentiates discourse types between conversation and text. Included in the former are: adult-child, classroom, interview, dinner, meeting, narrative, and courtroom; and in the latter: advert, comic strip, essay, joke, legal text, letter, literature, message, monologue, narrative, obituary, report, summary, etc. So it can be seen that the distinction between register and genre tends to be quite abstract and vague as regards these supposedly discrete categories.

Consideration must be made as to how this fuzziness might aid or impede attempts to classify according to genre. In particular, if criteria are to be defined for the automatic
detection of genre, which is becoming increasingly important in this age of technology with more and more of our literary, scientific and technical knowledge being made available for access on the internet, then these criteria need to be robust and well-defined.

Nowadays though, genre seems to be evolving and the criteria used to distinguish one from the other seem to be becoming more complex. Perhaps this is just a case of misclassification, however. Maybe the form of the documents is new and revolutionary, for example on web-sites, blogs and the like, while their actual content remains the same as ever. There is a multi-modal view of documents which sees them as comprising more than just words. This would appear to be quite different from that which would have been considered to be the traditional approach of linguists to studies of this kind. One might be inclined to think of linguists being more concerned with "...exceedingly complex structured semiotic artefacts." (Bateman 2008:8). But in fact, documents, depending on their nature, have always contained illustrations and graphs and consisted of various layouts to achieve a certain impact, so that hasn't really changed. What has changed and continues to do so, is the medium through which they are delivered and made available.

According to Kessler et al. (1997:1): "Genre is necessarily a heterogeneous classificatory principle, which is based among other things on the way a text was created, the way it is distributed, the register of language it uses, and the kind of audience it is addressed to." This seems to be extending the definition of genre to take other factors into account which might indeed be affecting not only the form of the texts but also perhaps being themselves affected by the level of generality to be encountered on the internet, due to its transmission through electronic media. However it could be argued that the kind of audience for which it is intended will determine the register, which will also become a consideration when it is being created and distributed. The same factors define the genre, but new types of genres are evolving, while still others are becoming obsolete. But at the end of the day it does not change what genre actually means, although it might seem as if its definition is being moulded to suit the individual's purposes. But as mentioned earlier, one must bear in mind that genre
serves more than just a classificatory purpose, from the other angle it might be viewed as being clarifactory. And while arbitrary categorisations are to be avoided, so is rigid prescriptivism.

Santini differentiates between topical descriptors and non-topical descriptors when discussing document classification. The former comprise topic, content, subject matter and domain, while the latter consists of genre, register, style, sentiment/opinion, readability and generalisation. She goes on to describe how the non-topical descriptors are used as a means of automatic classification of web documents. With regard to the topical descriptors, this project will focus on selected pieces of subject matter within a particular domain, thereby analysing them in terms of content and topic. Of the non-topical descriptors, register variation demands special consideration, not only because there has been a lot of research in this area, particularly by Biber, but also because of its relative importance as a means of identifying the linguistic features which shed light on the genre itself. Register variation also has an impact on the style, sentiment/opinion, readability and generalisation of any corpus. In fact, in terms too of topic, content, subject matter and domain, register variation will also be influential. For example, if the domain was childhood, then the subject matter, content and topic would be quite different if discussed in a casual conversation or in a government policy paper. Small wonder then that these terms are difficult to distinguish and separate since there is such overlap between the different areas, and they depend on each other to such a large extent.

To obtain an overview of the definition of genre, it is worth considering some points found to be relevant by Swales (1990:44-45). These are: (i) a distrust of classification and prescriptivism, (ii) a sense of their importance for distinguishing past and present, (iii) a recognition of their place within discourse communities, (iv) an emphasis on communicative purpose and social actions, (v) an interest in generic structure which provides the rationale, and (vi) an understanding of the generative capacity of genres to establish rhetorical goals and to further their accomplishments. His final definition is

---

3 http://sites.google.com/site/marinasantiniacademicsite/marina_santini
that genre consists of a class of communicative events, sharing a communicative purpose, as defined by a discourse community. Content and style are constrained by the rationale for the genre. It is narrowly focused on rhetorical action and often prototypical.

As also pointed out by Lee (2001), there is a lot of overlap. But once a working definition can be arrived at, without being too rigid, in order to find a definition that works as a classificatory principle.

In some of the contexts above, the term 'texts' may seem to imply exclusively the written word, but this is not the case. However, for this study and because of the quantitative methods involved, written texts are of prime importance as inputs. If speech were being used this would necessitate the transcription of such using relevant conventions, in order to convert them into a written form. The qualitative analysis will take the form of analysing particular texts from the corpus in terms of layout and content as well as considering the actual linguistic features of the text itself.

2.2 Definitions:

This section undertakes to define the terms used in this project, in order to avoid ambiguity or misunderstanding. The terms 'corpus', 'text' and to a lesser extent 'document', are used interchangeably throughout. They all refer to a distinct body of discourse being used for the purpose of analysis. ‘Discourse’ can refer to both written and spoken forms. The term 'subject' is often used with regard to each distinct unit of discourse within the school-based genres under investigation in this project. These 'subjects' are English, Mathematics and Geography as defined by the Irish Junior Certificate syllabus that they are taken from. The term 'genre' itself is taken to mean the general class to which a piece of discourse belongs, this class having been identified by taking into account a variety of features, focusing mainly on the linguistic features as identified through the register used.

The non-topical descriptors suggested by Santini are topic, content, subject matter and domain. Of the non-topical descriptors, register is really the key, with genre, style, sentiment/opinion, readability and generalisation all being greatly influenced by it.
Register is the type of language used in the discourse which is dependent to a large extent on the audience to which it is addressed, as reflected by the social relationships between the addressee and the addressor as well as the social situation. The purpose for which it is intended is also important since the context will affect the level of care to be invested in the communication. Style will depend largely on the register used, with slight modifications within that, depending on the individual’s preferences. For example, the style of a dinner conversation might be more or less formal. Again, sentiment/opinion is related to the register used, since the addressor’s attitudes are reflected in the type of language used, which in turn depend on the circumstances. As regards readability, this category would need to include spoken language, so it will be considered to include transcripts. Again, how ‘readable’ it is will depend on the register used, reflecting its purpose, and the level of understanding of the addressee. For instance, a face-to-face conversation using a lot of slang and a medical report directed at that profession might equally present problems to the uninitiated. The term generalisation is assumed to refer to the level of generalisation within a discourse. This would imply that it might range from the very specific, using quite technical terms, to very broad, non-specific terms. All in all these criteria for differentiating between documents seem rather vague and obscure in terms of how they might actually be implemented or applied.

The statement: “exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content and intended audience” (Swales 1990:58) seems to present other features to be examined in relation to identifying genre, but these too are subject to the comments made above regarding the features selected by Santini as being important.

Still more definitions are suggested and explored by Lee (2001), but this time in a more systematic and useful fashion. In his article entitled “Genres, Registers, Text Types, Domains and Styles”, he attempts to clarify some of the issues. He starts by identifying genre as being based on external criteria such as audience, purpose and activity type. And then by distinguishing it from text-type which he claims is based on internal linguistic criteria such as grammatical co-occurrence features. These are further
compared to the traditional four-part rhetorical categories of narrative, description, exposition and argumentation. But in the end he admits that there is not a great difference between the two and that they are often used interchangeably. He continues by distinguishing between genre, register and style. As proposed previously in this article, genre serves cultural and social purposes. Register on the other hand is based on the situation or the immediate context. And style comes down to the individual’s use of language. If language can be considered in terms of form, function or meaning, so too can register and genre be considered as two very different ways of looking at the same item. Lee finally admits that “genre is never just about situated linguistic patterns (register), functional co-occurrences of linguistic features (text types), or subject fields (domain) and it is not even simply about text-structural/discoursal features…” (Lee 2001:52) He goes on to argue that it is about all of these things and that however messy this might seem, it still serves a purpose.

A further element of genre that is considered important in this study is form, which basically takes into account the appearance of the genre in terms of presentation and layout, being also influenced by the mode of communication.

As stated previously, there is a lot of overlap between these terms, so this is merely an attempt to demystify them, albeit to a limited extent.

There is often a lack of clarity between register and genre but this can be cleared by realising that register is a functional language variety characterised by groupings of linguistic features together with situational features. (Gregory and Carroll 1978.) Genres on the other hand are realised through registers. (Martin 1985.) Register occurs at the level of vocabulary and syntax while genre happens at the level of discourse structure. (Couture 1986.)

### 2.3 Register Variation:

“A register is the constellation of features (lexical and grammatical) that characterise particular uses of language. (Halliday & Hassan 1989, Martin 1992).” (Schleppegrell 2001:431). However, non-linguistic factors greatly influence the choice of register, in terms of the function that these linguistic features will serve. That is to say that the
need to communicate with different people, at different times and in different ways has resulted in well-defined linguistic differences developing. This is referred to as 'the functional dimension' by Biber (1995), since the function of the communication motivates the language choices to be made. The relationship between register and genre is thus that genres are functional varieties with these functions being identifiable by the occurrence of particular features. Biber (1988) identified sixteen grammatical categories and carried out extensive research into how these different categories manifest themselves in different types of discourse. These different types of discourse had been identified previously as specific genres. However this was not a comprehensive comparison of all genres, although it did encompass quite a number of them, as taken from the LOB (written word) and the London-Lund (spoken word) corpora. Therefore the identification of linguistic features within a corpus can give us clues as to the register being used which will have a bearing on the genre to which the corpus belongs.

In the various studies carried out by Biber, he identified particular linguistic features that tend to co-occur in texts, these having been identified by computational means. He initially identified three primary dimensions of linguistic variation and gave these the following tentative labels:

(i) Interactive versus Edited Text
(ii) Abstract versus Situated Content
(iii) Reported versus Immediate Style

The first dimension is characterised by features such as questions, first and second person pronouns versus word length and vocabulary richness. The second takes into account features such as nominalisations and passives versus time and place adverbials. The third dimension compares features such as past tense versus present tense verbs. He underlines the complexity of the relations between genres, and states that “no single dimension adequately captures the similarities and differences among genres“ (Biber 1988:57), that a multidimensional approach is needed. This implies that one needs to take all of the different factors into account when interpreting the results.
Some criticism has arisen with regard to how genres are identified as such for inclusion in the major corpora mentioned above and also in the BNC. Lee (2001) argues that genre is about many things, from register to text-type to domain and text-structural/discoursal features, and that however difficult this might make classification, that such classification still serves a purpose. The solution he suggests is to take a prototypical approach and to create an index allowing access to genres according to key elements.
Chapter 3

3.1 Corpus Linguistics:

Corpus Linguistics is the empirical study of language use where the data for such a study are taken from corpora. A corpus is defined as naturally occurring language texts gathered together in a collection. (Sinclair 1991). Thus they are composed of texts taken from language in everyday use and therefore considered to be representative of the real world. It could be argued that the textbooks used as the basis of the corpora under consideration in this study are not particularly natural since they are in fact artificially produced. It could also be argued that the fact that they are used as exemplars in education since the foundation of the education system itself means that they have evolved into objects representative of everyday use. Quantitative techniques allow the observation of trends, and further analysis of the nature of these trends enables conclusions to be drawn about the nature of the text itself. These techniques also allow inter-textual comparisons to be made, which can be very useful. These texts can be investigated from a phonological, morphological, lexico-grammatical, discoursal, etc., perspective, in order to determine linguistic features relating to register, genre, and the like.

The beauty of corpus linguistics, as pointed out by Sinclair (1998), lies in the fact that it allows us to access, more objectively, the subliminal patterns which run through a language than could ever be considered through introspection and intuition alone. Some of these techniques are currently in use to perform computer based tasks such as information retrieval, question answering, summarisation, part-of-speech tagging, genre recognition, anaphora resolution, syntactic dependency extraction, clustering, collocations, and terminology extraction. (Stuart 2005). The nature, structure and design of the corpora themselves will rely largely on their purpose and what they are meant to represent. It follows that in order to make generalisations about the language then the corpus being used must be representative of the target population. Encoding often makes corpora more useful and Stuart (2005) distinguishes four levels of document and text mark-up, as follows:
1) general document mark-up – this includes genre identification and bibliographic description of the document with details of author, title, publication date, journal name, author affiliation, etc.

2) general textual and structural mark-up – this involves the structural units of text such as volume, chapter etc., down to paragraph level, and also quotations, footnotes, headings, subheadings, tables, figures, graphs, etc.

3) contextual and linguistic annotation – includes discourse annotation (anaphora resolution, cohesive devices), pragmatic annotation (speech act type) and semantic annotation (semantic category of word).

4) language dependent annotation and mark-up for sentence-level structures – these include sentences; words; abbreviations, names, dates, etc.; morphological information; syntactic information (part-of-speech tagging); and prosodic annotation.

Since it is known that many disciplines are characterised by their own particular terminology, it is often considered useful to identify these terms. This is done by first building a corpus of domain knowledge and then using statistical and linguistic analysis to extract the relevant information. This task itself is a form of knowledge discovery in texts; that is “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in unstructured data”. (Karanikas et al. 2005:2). The procedure to be followed is first to design the corpus. This is achieved by determining the optimal number of texts so that they include everything necessary to be representative of the knowledge domain. Since many disciplines are characterised by their own particular terminology, it is useful to identify these terms. Secondly, the criteria for eliminating non-technical terms must be established. This removes function or grammar words and then high-frequency terms can be selected. Statistical comparison of terms found in a general text and those found in a specialist text can also be used for this purpose, or to make a 'stop word list'. The corpus is then ready to be analysed.
3.2 Various Methodologies:

The methodologies which can be used for genre identification are quite varied. These range from the quantitative to the qualitative. The quantitative methods are generally automatic, usually involving the tagging and parsing of texts, followed by statistical analysis of the various features identified. The effectiveness of these methods is often considered in relation to how ‘computationally expensive’ they are. Basically this means how complex they are and the processing resources required. The qualitative methods are usually a manual analysis of a specific piece of discourse.

The automatic methods available to search or otherwise exploit corpora in order to identify the inherent knowledge are many and varied. From indexing tools to search and retrieval tools, the identification of concordances and collocations, through other statistical and quantitative tools that identify parts-of-speech and thus patterns within texts. Some of these methods are investigated and discussed below, and some of them will be used in the analysis. The main criteria for differentiating between the various methods are how shallow or how deep they are and this is usually reflected in their computational cost. However, that is not to say that the more costly methods will be more effective.

Lustrek (2007) separates automatic genre identification methods into three classes. The first class uses traditional methods where a set of features is extracted for treatment by a classification algorithm. The second class is character based where the documents are modelled as sequences of characters. The third class uses visual methods, typically on scanned documents where the documents are represented as bitmaps. (Lustrek 2007). The feature extraction methods within the traditional method are considered interesting as far as this study is concerned. These include the identification of surface, structural and presentation features. Surface features include: function words; genre-specific words, phrases and punctuation marks; classes of words or phrases; vocabulary richness; all words and punctuation marks; word length; sentence complexity; document length. Structural features include: parts-of-speech; phrases; verb tense; sentence types (e.g. declarative, imperative and question
sentences); parser specific. Presentation features include the token type, graphical elements, links and other html tags.

Stylometric approaches are discussed by Stamatos et al (2000). These include vocabulary richness and common word frequencies. As the name suggests, the latter takes into account the frequencies of the occurrence of the most frequent words in a corpus. Typically the top fifty words are used. This method involves minimal computational cost and is independent of the language and the domain. Other methods mentioned by them are style markers (i.e. countable linguistic features). These include structural markers (e.g. noun count), lexical markers (e.g. 'it' count), and token-level markers (e.g. words per sentence average, type/token ratio, etc). They also mention character level markers (e.g. punctuation mark counts) and derivative markers (e.g. ratios and variation measures derived from lexical and character level markers), neither of which require a tagged or parsed text.

Santini (2004) discusses different criteria for the automatic classification of documents. These include topic (text categorization, information filtering), author (authorship attribution), and genre (identification/detection/classification). Function words, n-grams of parts-of-speech, and machine learning techniques as well as the identification of syntactic features are mentioned in this paper. Of the latter she writes that they have often been neglected because they were previously considered to be computationally expensive.

This gives a general idea of the amount of variety and overlap amongst the methods available and proposed by the different researchers.

3.3 Proposed Methodologies:

As mentioned previously, the methods used in this study are both quantitative and qualitative. These are outlined and discussed below.

The subjects that have been chosen are English, Geography and Mathematics, since these represent a broad cross-section of the subjects available in the Irish Junior Certificate syllabus. Register variation has been shown to be something continuous
rather than discrete. And different genres can be shown to contain linguistic and other features typical of the genre itself. That is to say they can be identified by the bundle of features typical of each. It follows then that sub-genres may be identified by traits particular of that type, too. So if we were to consider “…school-based genres [which] exhibit many common register features.” (Schleppegrell 2001:432), in terms of register variation in general, then we might be able to place the different academic styles within this along a continuum, on the basis of the linguistic and other features identified. The English and Mathematics texts might be considered as representative of subjects at the extreme ends of this continuum, with Geography somewhere in the middle. This is based purely on impressionistic factors, at least for the moment. Following a cursory glance at the textbooks, it can be seen that indeed the English textbooks contain a lot of words and not much else, while the Mathematics textbooks contain a lot of mathematical calculations and a lot less words. The Geography textbooks contain more of a balance between words, diagrams, charts, maps, etc.

Thus the corpus comprises the complete set of textbooks currently in use for these three subjects. In order to work with the corpus, sections were isolated as follows: English only, Mathematics only, Geography only, the corpus excluding English, the corpus excluding Mathematics, and the corpus excluding Geography. Please see Appendix I for exact information on the corpus.

The qualitative analysis will involve a review of the physical appearance of the three subject areas in terms of layout and content. This will take into account the use of non-linguistic features, such as maps, graphs, charts, diagrams, etc. Any striking linguistic features will also be identified and discussed. A complete qualitative analysis is discussed in the section entitled ’Qualitative Analysis’ in Chapter 4.1.

The methods proposed for the quantitative analysis include the following: generation of word-lists and key-word identification, concordances and collocations as provided by the WordSmith programme. The texts will also be tagged and parsed using the TreeTagger programme. Surface and structural features will be extracted and analysed using a custom programme and the ’R’ statistical software, with the help of spreadsheets. Further features will be identified and statistically analysed (including
factorial analysis) using Hyperbase, a programme designed specifically for corpus linguistic purposes. This will result in the added advantage of enabling the assessment of the effectiveness of the various methods to be considered, in terms of computational cost and time.
Chapter 4

4.1 Qualitative Analysis:

The qualitative analysis consists of a review of the physical appearance of the three subject areas in terms of layout and content. This will take into account the use of non-linguistic features, such as maps, graphs, charts, diagrams, etc. Any striking linguistic features will be also be identified and discussed.

The language used at this level, that is, aimed at fourteen and fifteen year-olds preparing for the Irish Junior Certificate examination may not appear to be overly scholarly at first glance. However, one should bear in mind that these textbooks are directed at those in their early teens, who would have had very little exposure to academic writing, if any. In fact, it seems to be preparing the students for more academic levels. For instance, one sentence in the Geography text reads:

(1) “Carbon dioxide is given off (released).”

And the Mathematics text starts to introduce other ways of asking the same question in more academic terms. For example:

(2) “This question is asking ‘where does the curve meet the x axis?’ ‘Find the values of x for which $2x^2-3x-5 = 0$’ is another way of asking the same question.”

To start with the Mathematics textbook, the language used is very concise and to the point, factual, characterised by a description of the subject matter, usually explained using a diagram and a formula. This is followed by worked examples, that is, step-by-step instructions on how to calculate a solution. Exercises and solutions are given, so that students can check if they are correct. Another noticeable feature is that of the presence of theorems which are very rigid proofs, set out in a very strict manner. The language is very particular, and probably quite academic, since key-words are very specific and precise. However, these same key-words are repeated throughout. In fact the text is not very wordy: it uses calculations, formulae and symbols mainly. To a certain extent, this is the language of maths, and one must understand the fundamentals in order to be able to follow what is going on. One striking feature is an
attempt to keep the language within the scope of the students, or at least to keep them interested by using terms and settings with which they might be familiar. For example:

(3) “Hours spent watching tv”, “Mars compared to Twix and Snickers”, “swimming pool attendance” or “trees in an orchard”

Again, graphs, charts and diagrams are used as is the technique of containing important information within a box, in order to highlight it.

Most noticeable is the use of the imperative (e.g. ‘Find the equation’) or the infinitive (‘To find: ’) as well as terms such as:

(4) “You will find…” or “We find…”

There also seems to be an over abundance of nouns, and one would suspect a lack of vocabulary richness. There’s a high incidence of propositions, which may indicate abstractness or a means of compacting information.

To continue with the Geography corpus, this seems to be a very broad subject area. Even the list of contents gives some clues as to what to expect. It is split between the physical, social and economic aspects of the world, and these areas are further divided into topics such as ‘the human habitat’, 'population' and 'rich and poor regions'. The text is structured to teach by describing physical, social and economic phenomena as a series of events, one following from another, and building on the information already imparted. So the vast reach of this subject would suggest a rich vocabulary and perhaps varying styles. The written text is further exemplified by diagrams, graphs, charts, sketches, instructions, cartoons, photographs and maps. Therefore it is a multi-modal text, relying on various strategies other than just language in order to communicate its knowledge and information. It is quite specific and fine-grained, going into a lot of detail and elaboration. As shown, for example, in the following sentence:

(5) “Limestone is formed from the remains of tiny sea creatures, shells, fish skeletons and coral.”
It includes case studies, and one such, pertaining to manufacturing industries and the computer industry in particular, even goes as far as to describe computer components. This is far reaching, indeed.

Geography seems to use the present tense quite often, perhaps in order to remove the focus from time events. It also appears to be very factual through the use of many nouns and also very descriptive through the use of many adjectives. Place names are quite common, as would be expected.

Next the English text whose subject matter is quite broad, since it consists of practically everything to do with the English language, including the meta-language itself. A glance at the list of contents shows just how broad this subject matter is. Each section will have features characteristic of its type, with its own style and register. These sections are: reading, personal writing, functional writing, media studies, drama, poetry, and fiction. This broad scope puts a question on whether English can be treated as a genre by itself, or should it be subdivided between these individual elements, the sub-genres? In general the texts are designed to teach the student how to read and understand different kinds of writing techniques. They give instructions on how this might be achieved. Sample texts from all sections are followed by a summary and a commentary. Possible sample questions are given, and a successful possible answer is outlined. The styles, or indeed genres, found in the variety of writing included range through academic, factual, literary, narrative, articles and reports. Some show or try to provoke quite abstract responses from the reader and encourage conceptualisation; others are factual requiring no imagination; and others are extremely imaginative. At a glance, the use of descriptions is very noticeable in the fiction section, and this would be identified through the use of adjectives. The poetry section makes use of first and second person pronouns showing a highly involved style. There are very few graphical or visual elements.

4.2 Quantitative Analysis:

The quantitative analysis is achieved by the use of computational linguistic means. These include the generation of word-lists and key-word identification, concordances
and collocations as provided by the WordSmith programme. The texts will also be
tagged and parsed using the TreeTagger programme. Syntactic features will be
extracted and analysed using a custom programme and spreadsheets with some help
from the 'R' statistical software. Further structural features will be identified and
statistically analysed (including factorial analysis) using Hyperbase, a programme
designed specifically for corpus linguistic analysis purposes.

Using the WordSmith programme, texts from the entire corpus comprising the three
chosen subject areas were analysed. Wordlists were first produced for each of the
subject areas only, and then a list of keywords was identified by comparing the corpus
for each subject with the entire corpus. This was also done by comparing each
individual corpus with the entire corpus excluding itself.

A preliminary examination of the key-words for the corpora revealed the following facts,
as summarised in Table 1.
<table>
<thead>
<tr>
<th>Ranking</th>
<th>English</th>
<th>Geography</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HE</td>
<td>THESE</td>
<td>#</td>
</tr>
<tr>
<td>2</td>
<td>YOU</td>
<td>AREAS</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>HIS</td>
<td>SUCH</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>WAS</td>
<td>CITY</td>
<td>CALCULATE</td>
</tr>
<tr>
<td>5</td>
<td>WHAT</td>
<td>COUNTRIES</td>
<td>FIND</td>
</tr>
<tr>
<td>6</td>
<td>IT</td>
<td>POPULATION</td>
<td>NUMBER</td>
</tr>
<tr>
<td>7</td>
<td>STORY</td>
<td>MANY</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>POEM</td>
<td>TOWNS</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>SHE</td>
<td>IN</td>
<td>ANGLE</td>
</tr>
<tr>
<td>10</td>
<td>CHARACTER</td>
<td>IDENTIFY</td>
<td>L</td>
</tr>
<tr>
<td>11</td>
<td>ABOUT</td>
<td>AND</td>
<td>Y</td>
</tr>
<tr>
<td>12</td>
<td>THINK</td>
<td>MAP</td>
<td>TRIANGLE</td>
</tr>
<tr>
<td>13</td>
<td>HIM</td>
<td>PEOPLE</td>
<td>ANGLES</td>
</tr>
<tr>
<td>14</td>
<td>LIKE</td>
<td>STUDY</td>
<td>LINE</td>
</tr>
<tr>
<td>15</td>
<td>MY</td>
<td>CITIES</td>
<td>M</td>
</tr>
<tr>
<td>16</td>
<td>HER</td>
<td>ARE</td>
<td>EQUATION</td>
</tr>
<tr>
<td>17</td>
<td>WRITER</td>
<td>SOME</td>
<td>DIAGRAM</td>
</tr>
<tr>
<td>18</td>
<td>THE</td>
<td>SETTLEMENT</td>
<td>E</td>
</tr>
<tr>
<td>19</td>
<td>US</td>
<td>IRELAND</td>
<td>C</td>
</tr>
<tr>
<td>20</td>
<td>BUT</td>
<td>LARGE</td>
<td>P</td>
</tr>
<tr>
<td>21</td>
<td>ME</td>
<td>LAND</td>
<td>IMAGE</td>
</tr>
<tr>
<td>22</td>
<td>HOW</td>
<td>NEW</td>
<td>H</td>
</tr>
<tr>
<td>23</td>
<td>WORDS</td>
<td>TOWN</td>
<td>TAX</td>
</tr>
<tr>
<td>24</td>
<td>READ</td>
<td>INDUSTRY</td>
<td>PROVE</td>
</tr>
<tr>
<td>25</td>
<td>HAD</td>
<td>PER</td>
<td>CM</td>
</tr>
<tr>
<td>26</td>
<td>YOUR</td>
<td>FACILITIES</td>
<td>SIDES</td>
</tr>
<tr>
<td>27</td>
<td>SEE</td>
<td>EXPLAIN</td>
<td>N</td>
</tr>
<tr>
<td>28</td>
<td>SAID</td>
<td>REGION</td>
<td>SYMMETRY</td>
</tr>
<tr>
<td>29</td>
<td>POET</td>
<td>INDUSTRIAL</td>
<td>CD</td>
</tr>
<tr>
<td>30</td>
<td>GOING</td>
<td>RIVER</td>
<td>VALUE</td>
</tr>
<tr>
<td>31</td>
<td>WRITING</td>
<td>SETTLEMENTS</td>
<td>Q</td>
</tr>
<tr>
<td>32</td>
<td>UP</td>
<td>PHOTOGRAPH</td>
<td>CIRCLE</td>
</tr>
<tr>
<td>33</td>
<td>DRAMA</td>
<td>DUBLIN</td>
<td>F</td>
</tr>
<tr>
<td>34</td>
<td>LOOK</td>
<td>WOMEN</td>
<td>SOLUTION</td>
</tr>
<tr>
<td>35</td>
<td>DOES</td>
<td>DEVELOPED</td>
<td>UNDER</td>
</tr>
<tr>
<td>36</td>
<td>OUT</td>
<td>URBAN</td>
<td>EQUAL</td>
</tr>
<tr>
<td>37</td>
<td>CHARACTERS</td>
<td>WATER</td>
<td>SOLVE</td>
</tr>
<tr>
<td>38</td>
<td>SOMETHING</td>
<td>LOCATED</td>
<td>LENGTH</td>
</tr>
<tr>
<td>39</td>
<td>WOULD</td>
<td>EVIDENCE</td>
<td>EM</td>
</tr>
<tr>
<td>40</td>
<td>PASSAGE</td>
<td>THEIR</td>
<td>CORRECT</td>
</tr>
<tr>
<td>41</td>
<td>DO</td>
<td>SOUTH</td>
<td>POINT</td>
</tr>
<tr>
<td>42</td>
<td>READER</td>
<td>MOST</td>
<td>NUMBERS</td>
</tr>
<tr>
<td>43</td>
<td>ANSWER</td>
<td>OIL</td>
<td>SIN</td>
</tr>
<tr>
<td>44</td>
<td>TO</td>
<td>CENT</td>
<td>T</td>
</tr>
<tr>
<td>45</td>
<td>JUST</td>
<td>BUILT</td>
<td>EVALUATE</td>
</tr>
<tr>
<td>46</td>
<td>WHEN</td>
<td>DEVELOPMENT</td>
<td>TERMS</td>
</tr>
<tr>
<td>47</td>
<td>MAKE</td>
<td>WORK</td>
<td>AXIS</td>
</tr>
<tr>
<td>48</td>
<td>BACK</td>
<td>SERVICES</td>
<td>COST</td>
</tr>
<tr>
<td>49</td>
<td>WITH</td>
<td>HOUSING</td>
<td>GRAPH</td>
</tr>
<tr>
<td>50</td>
<td>DID</td>
<td>HAS</td>
<td>DRAW</td>
</tr>
</tbody>
</table>

**Table 1.** List of the top fifty key-words for each corpus.
This was followed by an investigation into concordance, collocations and clusters for particular words that seemed to be of particular interest. These words were picked at random from the top five keywords. For example, the top five keywords from the Mathematics corpus are '#', 'X', 'B', 'CALCULATE' and 'FIND', thus the relevance of 'find' seemed to demand attention, while 'calculate' seemed fairly predictable. Likewise with the Geography corpus where the top five words are: 'THESE', 'AREAS', 'SUCH', 'CITIES' and 'COUNTRIES', respectively, the other words seemed self-explanatory while 'such' seemed to require investigation. On the contrary, in the English corpus where the top five keywords are: 'HE', 'YOU', 'HIS', 'WAS' and 'WHAT', since these all seemed fairly ordinary then 'you' was selected at random. The following analysis is quite superficial since it considers one word from each corpus only, to give an example of the level of analysis that is available through these means.

To continue with the analysis then, the word ‘find’ in the Mathematics corpus appears mainly in the imperative and the infinitive. It collocates with: the, of, I, a, to, x, II, and, value, image. It clusters with: _ the image, the image of, to _ the, the value of, _ the value, II _ the, I _ the, _ the area, equation to _, solve the equation, etc. Thus it is used primarily in instructions.

Next the word ‘such’ in the Geography corpus tends to be found mainly idiomatically but also adjectivally and adverbially. It collocates with: as, and, in, the, of, areas, are, to, a; and it clusters with: _ as the, activities _ as, areas _ as, facilities _ as, cities _ as, _ as this, _ as Dublin, etc. Therefore it is used primarily in descriptions.

And the word 'you' in the English corpus occurs mainly in narratives, with instructions or questions, and in poetry. It collocates with: you, the, to, a, what, have, do, of, are, will; and it clusters with: do _ think, what do _ think, _ have read, to help _, _ think the, asking _ to, what _ have, is asking _, _ have studied. So it is used in a variety of ways, within the constraints of it being a pronoun.

Further analysis was undertaken using a custom programme to identify and quantify the parts-of-speech that had previously been annotated by the TreeTagger, for each of the corpora. The results of this are shown in Table 2. which lists each part-of-speech
alongside the number of times it occurred in each corpus. Striking differences can be easily identified when the figures are laid out, side-by-side, in this way.
<table>
<thead>
<tr>
<th>Part-of-speech:</th>
<th>English</th>
<th>Geog</th>
<th>Maths</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>9227</td>
<td>9524</td>
<td>11764</td>
<td>NN</td>
</tr>
<tr>
<td>Nouns, plural</td>
<td>2832</td>
<td>5701</td>
<td>2485</td>
<td>NNS</td>
</tr>
<tr>
<td>Nouns, proper</td>
<td>2975</td>
<td>4019</td>
<td>5297</td>
<td>NP</td>
</tr>
<tr>
<td>Nouns, proper, plural</td>
<td>49</td>
<td>53</td>
<td>5</td>
<td>NPS</td>
</tr>
<tr>
<td>Sub-total:</td>
<td>15083</td>
<td>19297</td>
<td>19551</td>
<td></td>
</tr>
<tr>
<td>Adjectives</td>
<td>2949</td>
<td>4985</td>
<td>3553</td>
<td>JJ</td>
</tr>
<tr>
<td>Adjectives, comparative</td>
<td>90</td>
<td>249</td>
<td>124</td>
<td>JJR</td>
</tr>
<tr>
<td>Adjectives, superlative</td>
<td>65</td>
<td>205</td>
<td>116</td>
<td>JJS</td>
</tr>
<tr>
<td>Sub-total:</td>
<td>3104</td>
<td>5439</td>
<td>3793</td>
<td></td>
</tr>
<tr>
<td>Possessive pronouns</td>
<td>3508</td>
<td>835</td>
<td>718</td>
<td>PP</td>
</tr>
<tr>
<td>Possessive endings</td>
<td>1020</td>
<td>537</td>
<td>237</td>
<td>PP$</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>405</td>
<td>265</td>
<td>182</td>
<td>POS</td>
</tr>
<tr>
<td>Prepositions</td>
<td>61</td>
<td>43</td>
<td>54</td>
<td>PDT</td>
</tr>
<tr>
<td>Sub-total:</td>
<td>4994</td>
<td>1680</td>
<td>1191</td>
<td></td>
</tr>
<tr>
<td>Adverbs</td>
<td>2409</td>
<td>1604</td>
<td>1044</td>
<td>RB</td>
</tr>
<tr>
<td>Adverbs, comparative</td>
<td>56</td>
<td>82</td>
<td>19</td>
<td>RBR</td>
</tr>
<tr>
<td>Adverbs, superlative</td>
<td>25</td>
<td>52</td>
<td>5</td>
<td>RBS</td>
</tr>
<tr>
<td>Sub-total:</td>
<td>2878</td>
<td>1828</td>
<td>1165</td>
<td></td>
</tr>
<tr>
<td>Modals</td>
<td>646</td>
<td>354</td>
<td>231</td>
<td>MD</td>
</tr>
<tr>
<td>Verbs, base form</td>
<td>2523</td>
<td>1620</td>
<td>1539</td>
<td>VB+VH+V</td>
</tr>
<tr>
<td>Verbs, past tense</td>
<td>2386</td>
<td>829</td>
<td>373</td>
<td>VBD+VHD+VWD</td>
</tr>
<tr>
<td>Verbs, gerund or present participle</td>
<td>1135</td>
<td>517</td>
<td>397</td>
<td>VBG+VHG+VVG</td>
</tr>
<tr>
<td>Verbs, past participle</td>
<td>1066</td>
<td>1425</td>
<td>809</td>
<td>VBN+VHN+VNN</td>
</tr>
<tr>
<td>Verbs, non 3rd person singular, present</td>
<td>1659</td>
<td>1604</td>
<td>990</td>
<td>VBP+VHP+WP</td>
</tr>
<tr>
<td>Verbs, 3rd person singular, present</td>
<td>1824</td>
<td>1269</td>
<td>1392</td>
<td>VBZ+VHZ+VVZ</td>
</tr>
<tr>
<td>Sub-total:</td>
<td>11239</td>
<td>7618</td>
<td>5731</td>
<td></td>
</tr>
<tr>
<td>Wh-determiners</td>
<td>212</td>
<td>325</td>
<td>162</td>
<td>WDT</td>
</tr>
<tr>
<td>Wh-pronouns</td>
<td>593</td>
<td>280</td>
<td>84</td>
<td>WP</td>
</tr>
<tr>
<td>Possessive wh-pronouns</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>WPS</td>
</tr>
<tr>
<td>Wh-adverbs</td>
<td>527</td>
<td>307</td>
<td>230</td>
<td>WRB</td>
</tr>
<tr>
<td>Sub-total:</td>
<td>1337</td>
<td>912</td>
<td>478</td>
<td></td>
</tr>
<tr>
<td>Determiners</td>
<td>6205</td>
<td>5096</td>
<td>5915</td>
<td>DT</td>
</tr>
<tr>
<td>Prepositions</td>
<td>5034</td>
<td>6321</td>
<td>5086</td>
<td>IN</td>
</tr>
<tr>
<td>Coordinators</td>
<td>1676</td>
<td>1920</td>
<td>1204</td>
<td>CC</td>
</tr>
<tr>
<td>Existential there</td>
<td>118</td>
<td>91</td>
<td>37</td>
<td>EX</td>
</tr>
<tr>
<td>Tns</td>
<td>1248</td>
<td>1113</td>
<td>832</td>
<td>TO</td>
</tr>
<tr>
<td>Foreign words</td>
<td>21</td>
<td>19</td>
<td>33</td>
<td>FW</td>
</tr>
<tr>
<td>List items</td>
<td>310</td>
<td>9</td>
<td>414</td>
<td>LS</td>
</tr>
<tr>
<td>Symbols</td>
<td>139</td>
<td>163</td>
<td>3748</td>
<td>SYM</td>
</tr>
<tr>
<td>Cardinals</td>
<td>1076</td>
<td>1963</td>
<td>4915</td>
<td>CD</td>
</tr>
<tr>
<td>Interjections</td>
<td>28</td>
<td>3</td>
<td>31</td>
<td>UH</td>
</tr>
<tr>
<td>Totals:</td>
<td>54490</td>
<td>53472</td>
<td>54124</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** List of each part-of-speech alongside the number of occurrences identified in each corpus, using the custom programme.
The correlation between the different parts of speech was investigated by reading the comma separated value (.csv) file of the above table into R. It was found that the English text correlated 0.910% with the Geography text showing a p-value of $1.488 \times 10^{-14}$. It correlated 0.855% with the Mathematics text with a p-value of $3.052 \times 10^{-11}$. And the correlation between the Geography text and the Mathematics text was 0.888% with a p-value of $5.016 \times 10^{-13}$. So they appeared quite similar at this high, overall level. A correlation test for all features was executed and those with high values were further tested to check if significant or not. For example, the relationship between the use of nouns and base verbs was shown to have a negative correlation of -0.999. Although the correlation coefficient is not equal to 0, this does not mean that these variables are dependent. However since it is nearer to -1 than not, this does suggest a negative linear relationship. This means that as one increases, the other decreases, so the relationship is inverse. Also the probability of obtaining a result like this is good since the p-value is 0.015 and less than the limit of 0.05. Next, the use of pronouns and base verbs was considered. This showed a positive correlation of +0.999 with a p-value of 0.021, so this was considered to be fairly significant, showing a positive linear relationship between these two features. Other features showing possible relationships between them were found to be nouns and pronouns with a correlation of -0.998 and a p-value of 0.036, and adverbs and present tense verbs with a correlation of -0.998 and a p-value of 0.039. In general high negative and positive correlations between some of the features were observed, but other than the above, the p-values were rarely lower than 0.05 which ruled out any major significance. Results are discussed further when discussing individual linguistic features.

Ratios were also calculated to highlight trends between the corpora and these are presented in Table 3. below, and analysed and discussed in the analysis a little further on which deals with the main linguistic features in turn and in detail. These were also compared to the weighted ratios taken from Hyperbase (see appendix VI for these values) and agreed with these to a very large extent. This shows that the abridged corpora used with the custom programme remained quite true to the original corpora and also that the methods for tagging and parsing also held over the two systems.
### Table 3. Ratios of features between the corpora.

<table>
<thead>
<tr>
<th>Part-of-speech</th>
<th>English</th>
<th>Geog</th>
<th>Maths</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>0.28</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>adjectives</td>
<td>0.25</td>
<td>0.44</td>
<td>0.31</td>
</tr>
<tr>
<td>pronouns</td>
<td>0.63</td>
<td>0.21</td>
<td>0.15</td>
</tr>
<tr>
<td>adverbs</td>
<td>0.49</td>
<td>0.31</td>
<td>0.20</td>
</tr>
<tr>
<td>verbs</td>
<td>0.48</td>
<td>0.33</td>
<td>0.25</td>
</tr>
<tr>
<td>wh-elements</td>
<td>0.49</td>
<td>0.33</td>
<td>0.18</td>
</tr>
<tr>
<td>determiners</td>
<td>0.36</td>
<td>0.30</td>
<td>0.34</td>
</tr>
<tr>
<td>prepositions</td>
<td>0.31</td>
<td>0.38</td>
<td>0.31</td>
</tr>
<tr>
<td>coordinators</td>
<td>0.35</td>
<td>0.40</td>
<td>0.25</td>
</tr>
<tr>
<td>ex-theres</td>
<td>0.48</td>
<td>0.37</td>
<td>0.15</td>
</tr>
<tr>
<td>tos</td>
<td>0.39</td>
<td>0.35</td>
<td>0.26</td>
</tr>
<tr>
<td>foreign words</td>
<td>0.29</td>
<td>0.26</td>
<td>0.45</td>
</tr>
<tr>
<td>numerals</td>
<td>0.10</td>
<td>0.18</td>
<td>0.72</td>
</tr>
<tr>
<td>interjections</td>
<td>0.45</td>
<td>0.05</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Ratios were then calculated to highlight trends within the individual corpora, and those higher than 0.10 were considered to be significant. In the English text these were found to be: simple nouns, total nouns, total pronouns, total verbs, determiners and prepositions. And in the Geography text: simple nouns, plural nouns, total nouns, simple adjectives, total adjectives, total verbs, determiners and prepositions. And finally in the Mathematics text: simple nouns, proper nouns, total nouns, total verbs, determiners, prepositions and symbols. The significance of these is discussed below when discussing the linguistic features in detail.

Spreadsheets were used to calculate rankings so that any striking features could be clearly identified. A baseline ranking was calculated to show where the different features stood in relation to one another. For example, it was found that the overall top five features were: 1.) nouns, 2.) determiners, 3.) prepositions, 4.) proper nouns and 5.) adjectives. Then a baseline rank difference was calculated for each subject area, which clearly showed the features that had the same ranking as the baseline, since the
difference for these from the baseline was zero. These could then be ignored since they did not deviate from the norm. By the same token, those showing high positive values indicating that they were over-represented in that particular text were easily identifiable. Conversely, those showing high negative values indicating that they were under-represented in that particular text were also very apparent. For example, symbols stood at -12 in the English table and the Geography table, while they were at +8 in the Mathematics table. List items stood at -10 in the Geography table while cardinals stood at -9 in the English table. These stood at +7 and +2 respectively in the Mathematics table. This result is hardly surprising, given the nature of the texts.

Personal pronouns, wh-pronouns and past tense verbs were under-represented in the Mathematics table with values of -7, -6 and -4 respectively. While in the English table personal pronouns and past tense verbs were over-represented with values of 5 each. In the Geography table the values of wh-determiners at +5, past participle verbs at +4, plural nouns, coordinators, to's, comparative and superlative adjectives all at +3 are also quite significant. So too is the value of -6 for personal pronouns. A similar exercise was carried out to find the baseline rank difference comparing the overall ranking with the rankings excluding each subject area in turn. These were also quite revealing, and corroborate the findings above. The results of these rankings are supported and further analysed and explained in the following analysis which deals with the main linguistic features in a sequential manner.

The use of verbs is quite striking at first glance showing a distribution of 11239 for English, 7618 for Geography and 5731 for Mathematics. In order to truly appreciate the impact of verbs, it is necessary to look at the finer detail, in the context of tense and aspect marking. This finer detail is displayed in Table 4, below.
### Table 4. Distribution of the different verb forms among the corpora.

For instance, past-tense forms are usually taken to be a surface marker of narrative. These occur 2386 times in the English text, 829 times in the Geography text and 373 times in the Mathematics text. As such, these would seem representative of expectations, as one would expect a more narrative style in the English text, with much less in both the Geography text and also the Mathematics text. Perfect aspect forms have also been associated with descriptive texts, since they mark past actions relevant to the present. The counts for these are 1066, 1425 and 809 for English, Geography

<table>
<thead>
<tr>
<th>Verbs:</th>
<th>English</th>
<th>Geog</th>
<th>Maths</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verbs, base form</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be</td>
<td>204</td>
<td>151</td>
<td>164</td>
</tr>
<tr>
<td>have</td>
<td>86</td>
<td>41</td>
<td>25</td>
</tr>
<tr>
<td>other</td>
<td>2233</td>
<td>1428</td>
<td>1350</td>
</tr>
<tr>
<td>sub-total</td>
<td>2523</td>
<td>1620</td>
<td>1539</td>
</tr>
<tr>
<td><strong>Verbs, past tense</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be</td>
<td>648</td>
<td>340</td>
<td>121</td>
</tr>
<tr>
<td>have</td>
<td>155</td>
<td>52</td>
<td>10</td>
</tr>
<tr>
<td>other</td>
<td>1583</td>
<td>437</td>
<td>242</td>
</tr>
<tr>
<td>sub-total</td>
<td>2386</td>
<td>829</td>
<td>373</td>
</tr>
<tr>
<td><strong>Verbs, ger/pr p</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be</td>
<td>30</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>have</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>other</td>
<td>1098</td>
<td>511</td>
<td>392</td>
</tr>
<tr>
<td>sub-total</td>
<td>1135</td>
<td>517</td>
<td>397</td>
</tr>
<tr>
<td><strong>Verbs, past p</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be</td>
<td>67</td>
<td>59</td>
<td>12</td>
</tr>
<tr>
<td>have</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>other</td>
<td>995</td>
<td>1363</td>
<td>795</td>
</tr>
<tr>
<td>sub-total</td>
<td>1066</td>
<td>1425</td>
<td>809</td>
</tr>
<tr>
<td><strong>Verbs, x3ps pres</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be</td>
<td>397</td>
<td>664</td>
<td>349</td>
</tr>
<tr>
<td>have</td>
<td>169</td>
<td>189</td>
<td>24</td>
</tr>
<tr>
<td>other</td>
<td>1093</td>
<td>751</td>
<td>617</td>
</tr>
<tr>
<td>sub-total</td>
<td>1659</td>
<td>1604</td>
<td>990</td>
</tr>
<tr>
<td><strong>Verbs, 3ps pres</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be</td>
<td>968</td>
<td>609</td>
<td>937</td>
</tr>
<tr>
<td>have</td>
<td>97</td>
<td>196</td>
<td>59</td>
</tr>
<tr>
<td>other</td>
<td>759</td>
<td>464</td>
<td>396</td>
</tr>
<tr>
<td>sub-total</td>
<td>1824</td>
<td>1269</td>
<td>1392</td>
</tr>
</tbody>
</table>
and Mathematics respectively, so these show the Geography text to be more descriptive than the other two. Next, the use of present tense verbs (including both third and non-third person singular plus present participle) amounts to 4618 for the English, 3390 for the Geography and 2779 for the Mathematics so a gradual difference between them. The higher the incidence of these is indicative of a more immediate relevance.

Remarkable too is the high incidence of third person singular verbs in the English corpora with 1824, as compared to 1392 in the Mathematics corpus and 1269 in the Geography corpus. Since their usage tends to place the focus on the immediate and, in academic styles, to remove the focus from time events, this demonstrates a focus on the present. It will be shown that the Geography text tends to focus on time events using many time adverbials, which is not surprising given the nature of that subject area. The findings above supports this. Weber (1985) notes that cognitive verbs typically occur in the present tense, so maybe this further explains why they are higher in the Mathematics corpus than the Geography corpus, albeit slightly. Tests were carried out to find the distribution of a random selection of cognitive verbs between the corpora to see if this hypothesis could be upheld. The list comprised: judge, justify, support, decide, verify, calculate, find, measure, construct, modify, compare, prove, present and define. It was found that these do indeed occur mainly in the Mathematics corpus, however the counts are 262 for calculate, 333 for find, and 121 for prove, so restricted to terms generally associated with that subject area.

There is a surprisingly high number of nouns in the Mathematics and Geography corpora (with counts of 19551 and 19297 respectively), quite a bit higher than that of the English corpus (with a count of 15083), including an elevated number of proper nouns (5302, 4072 and 3024 respectively). This is partly explained by the fact that both Mathematics and Geography use nouns in general for descriptions, and proper nouns in particular for demonstrations, or when phrasing questions. For example, 'the population of', the density of', 'find the length of the perimeter', 'what is the area', 'if Mary has x', 'John has y', etc. These findings are supported by the Hyperbase data set showing 40% for Mathematics, 33% for Geography and 27% for English. Nominal forms are generally associated with a high (abstract) informational focus, as opposed to the narrative focus obtained through the use of verbs, so perhaps these results are not so
surprising after all. One would expect Mathematics to be more abstract than English, particularly when taking into account the fact that the English corpus contains texts from a variety of styles, and thus includes a mixture of registers. In fact these results complement the findings of the verb distribution amongst the corpora, showing an inverse ordering.

The use of place and time adverbials are a means of bringing the focus on the physical environment at the present time and are therefore considered to show the level of involvement and how situated is the text. Place adverbials are very high in the Geography text at 319 with English at 196 and Mathematics at 148. This is not surprising really since the Geography text will be very concerned with words such as north, south, east, west and now without these actually being a feature that is used to situate the text. Table 5. shows a list of the incidence of place adverbials. This list has been compiled from a list suggested by Quirk et al. (1985:514) as further outlined in Appendix V.
<table>
<thead>
<tr>
<th>Place adverbials</th>
<th>English</th>
<th>Geog</th>
<th>Maths</th>
</tr>
</thead>
<tbody>
<tr>
<td>aboard</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>above</td>
<td>25</td>
<td>33</td>
<td>54</td>
</tr>
<tr>
<td>abroad</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>across</td>
<td>21</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>ahead</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>alongside</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>around</td>
<td>42</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>ashore</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>astern</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>away</td>
<td>27</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>behind</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>below</td>
<td>9</td>
<td>6</td>
<td>32</td>
</tr>
<tr>
<td>beneath</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>beside</td>
<td>12</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>downhill</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>downstairs</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>downstream</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>east</td>
<td>1</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>far</td>
<td>4</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>hereabouts</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>indoors</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>inland</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>inshore</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>inside</td>
<td>19</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>locally</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>near</td>
<td>9</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>nearby</td>
<td>0</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>north</td>
<td>0</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>nowhere</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>outdoors</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>outside</td>
<td>5</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>overboard</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>overland</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>overseas</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>south</td>
<td>0</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>underfoot</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>underground</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>underneath</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>uphill</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>upstairs</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>upstream</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>west</td>
<td>0</td>
<td>43</td>
<td>2</td>
</tr>
<tr>
<td>totals</td>
<td>196</td>
<td>319</td>
<td>148</td>
</tr>
</tbody>
</table>

Table 5. List of the incidence of place adverbials identified by the custom programme.
The occurrence of time adverbials in the English text is at 125 closely followed by the Geography text with 101 and then the Mathematics text with only 39. These figures are perhaps more predictable since they show the English text as being more situated followed by the Geography text while the Mathematics text is at the other extreme. Table 6. shows a list of the incidence of time adverbials. This list has been compiled from a list suggested by Quirk et al. (1985:526) as further outlined in Appendix V.

<table>
<thead>
<tr>
<th>Time adverbials</th>
<th>English</th>
<th>Geog</th>
<th>Maths</th>
</tr>
</thead>
<tbody>
<tr>
<td>afterwards</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>again</td>
<td>30</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>earlier</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>early</td>
<td>3</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>eventually</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>formerly</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>immediately</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>initially</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>instantly</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>late</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>lately</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>later</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>momentarily</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>now</td>
<td>36</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>nowadays</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>once</td>
<td>8</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>originally</td>
<td>0</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>presently</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>previously</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>recently</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>shortly</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>simultaneously</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>soon</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>subsequently</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>today</td>
<td>9</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>tomorrow</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tonight</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>yesterday</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>totals</td>
<td>125</td>
<td>101</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 6. List of the incidence of time adverbials identified by the custom programme.
Adjectives are a device used to “expand and elaborate the information presented in a text.” (Biber 1988:237). They occur 5439 times in the Geography text, 3793 times in the Mathematics text and 3104 times in the English text. Therefore we can conclude that the Geography text is much more informative than the other two. This is because it uses adjectives to compact information into sentences, using them to describe phenomena in great detail.

Adverbs are also supposed to serve a similar function to adjectives, but the counts of 2878 for English, 1828 for Geography and 1165 for Mathematics would seem to contradict the findings above. Biber (1988) notes that while adjectives, adverbs and prepositions all elaborate information, the type of information being elaborated may not be the same in each case. Perhaps the difference arises since Geography might use adjectives to describe nouns (and it is more nominal), while English uses adverbs to modify the verbs (and it is more verbal).

Surface features include function words such as prepositions, pronouns and articles. While function words of themselves might have little meaning, they are indicative of the grammatical relationships between words and often reflect the attitude of the speaker or writer. (Lustrek 2007).

The use of prepositions in the Geography corpus is high at 6321, with the Mathematics at 5086 and the English at 5034. Prepositions are “an important device for packing high amounts of information into academic nominal discourse” (Biber 1988:237) and would therefore indicate the level of ‘scholarliness’ of the corpus. It is interesting to note the high count in both the Geography and Mathematics corpora as they attempt to compact descriptions with as much information as possible. Equally interesting is the fact that the Mathematics text has a higher count than might be expected, especially when compared with English which one would expect to be much more complex. This feature supports the findings of the distribution of adjectives. Perhaps this could be due to the nature of the English corpus, containing texts from quite different areas of the syllabus, from narratives through adverts to poems and fiction. Perhaps also the high rate in the Mathematics corpus is due to the same prepositions being used again and again. This is supported by results from Hyperbase where the complete list of prepositions, as taken
from Quirk et al. (1985:665-667), were input individually. These show a high concentration of different words around the English corpus, with less variety at the Geography corpus and very few different words at the Mathematics corpus. The latter are *opposite, plus, minus, onto*, with *per* and *of* veering towards the Geography and English corpora. See Figure 1. for the distribution of prepositions among the corpora.

![Figure 1. Distribution of prepositions among the corpora.](image)

Biber (1986) found a tendency for prepositions to co-occur with nominalisations and passives in academic prose, so it would have been interesting to test for these other features too. **Passivisation** is a complex linguistic structure, containing as it does a combination of auxiliary verb, optional adverbs or noun/pronoun with a past participial verb and possibly a *by*. Unfortunately, the programs being used in this study cannot handle such complex structures, and to find them by hand would be too time-consuming.
A test for nominalisations was carried out however in Hyperbase, using words ending in particular forms as outlined in Appendix V. Figure 2. shows the results of these in more detail. Like prepositions, nominalised forms are used to integrate information and their function is often interpreted as conveying abstract as opposed to situated information. It is perhaps surprising then to note that they hover around the Geography and English corpora and are almost absent from the Mathematics corpus. When considered together with the results found from testing the incidence of nouns previously, these results seem to be somewhat contradictory.

Figure 2. The distribution of nominalisations among the corpora.

The use of pronouns is very high in the English corpus at 4994, while the Geography and the Mathematics corpora contain 1680 and 1191, respectively. This can be interpreted as meaning that the English corpus has a much less formal style than the other two. Personal pronoun counts included in these are 3508, 835 and 718 respectively. This is probably quite predictable since pronouns in general are indicative
of the informational load and show less precision in referential identification, according to Kroch and Hindle (1982) and Brainerd (1972). For example, first and second person pronouns show an interpersonal focus while third person pronouns “mark relatively inexact reference to persons outside of the immediate interaction” (Biber 1988:225).

Again, using the Hyperbase data, it can be shown that their, they, themselves and them occur nearer the Geography corpus while she, her, his, he, him, herself, and himself cluster nearer the English corpus. The most generalised pronoun, it, tends to be used inexplicitly, under time constraints and with a non-informational focus, according to Chafe and Danielewicz, and Biber. An investigation into the use of it shows occurrences of 605, 210 and 160 in the English, Geography and Mathematics corpora. This is probably a good reflection of what might be expected from these types of texts, and from the picture that is beginning to form from the results so far.

The use of determiners, including articles, was found to be 6205, 5915 and 5096 for the English, Mathematics and Geography corpora, respectively. Determiners, while contributing little to meaning might be considered as being detached from the topic but not from the style. (Lustrek 2007). This could be interpreted as meaning that while they do not affect what's being said, they do affect how it's said, again coming back to the idea of register. So English and Mathematics are quite high, closely followed by Geography. Therefore not a huge significance is demonstrated in these results.

Punctuation marks have been used in shallow based approaches to genre identification and these are presented below in Table 6. This table shows clearly the frequencies found to occur in each category, as determined by the custom programme. These are further demonstrated using graphs, to allow the easy identification of any trends. See Figures 3a, 3b and 3c. The counts are 9845 in the Mathematics corpus, 8548 in the English corpus and 5900 in the Geography corpus. Again these don't seem to be very significant when the natures of the texts involved are taken into account. For example, the high rate of full stops must be equivalent to decimal points in the Mathematics corpus, and the colons and parenthesis would also be quite predictable for this text.
Table 7. List of each punctuation mark alongside the number of occurrences identified in each corpus, using the custom programme.

<table>
<thead>
<tr>
<th>Punctuation mark:</th>
<th>English</th>
<th>Geog</th>
<th>Maths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question marks</td>
<td>759</td>
<td>360</td>
<td>113</td>
</tr>
<tr>
<td>Exclamation marks</td>
<td>182</td>
<td>45</td>
<td>38</td>
</tr>
<tr>
<td>Colons</td>
<td>443</td>
<td>321</td>
<td>630</td>
</tr>
<tr>
<td>Semi-colons</td>
<td>83</td>
<td>62</td>
<td>81</td>
</tr>
<tr>
<td>Commas</td>
<td>2370</td>
<td>1614</td>
<td>1931</td>
</tr>
<tr>
<td>Full stops</td>
<td>3814</td>
<td>2636</td>
<td>2713</td>
</tr>
<tr>
<td>Apostrophes</td>
<td>550</td>
<td>109</td>
<td>304</td>
</tr>
<tr>
<td>Double quotes</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Opened parenthesis</td>
<td>172</td>
<td>370</td>
<td>1907</td>
</tr>
<tr>
<td>Closed parenthesis</td>
<td>175</td>
<td>383</td>
<td>2128</td>
</tr>
<tr>
<td>Totals</td>
<td>8548</td>
<td>5900</td>
<td>9845</td>
</tr>
</tbody>
</table>

Figure 3a. Distribution of punctuation marks in the English corpus.
Figure 3b. Distribution of punctuation marks in the Geography corpus.

Figure 3c. Distribution of punctuation marks in the Mathematics corpus.

Legend for Figures 3a., 3b. and 3c.:

- QM = Question Marks
- XC = Exclamation Marks
- CN = Colons
- SC = Semi-colons
- CA = Commas
- FS = Full Stops
- AP = Apostrophes
- OP = Opened Parentheses
- CP = Closed Parenthesis
- DQ = Double Quotes – these were excluded since they were at zero.
A test was also carried out to quantify the average word length as this is often thought to give an indication of the complexity of the vocabulary used, as longer words are often more specific and defining. But with the counts at three for English, four for Geography and three for Mathematics one might conclude that there was very little difference between them. However, given the level at which these textbooks are directed, these findings are perhaps significant.

The original intention was to use Hyperbase for more specific functions that were considered to be more accessible in it than in the custom programme. However, due to problems encountered with Hyperbase (as outlined in Appendix II) it was decided to develop the custom programme and to use other resources to try to match its functionality. The results were to be further confirmed or refuted by comparison with the weighted values found using Hyperbase. Since this was no longer felt to be a very reliable option however, other methods were explored and used. Nonetheless, the graph below (Figure 4.) showing the distribution of all of the features for all of the corpora does seem to represent the facts as outlined and discussed previously. And they have also been supported when compared with ratios derived for the main linguistic features as presented in Table 3.
Figure 4. Distribution of all of the features for all of the corpora.
Chapter 5

5.1 Conclusion:

At the outset, an attempt was being made to arrive at a clear-cut definition for the term genre. This has turned out to be unexpectedly difficult, as a lot of the terms used seem to be interchangeable and indeed there tends to be quite some overlap between the different areas. However, for the purposes of this project it has been necessary to adopt an appropriate understanding of the term and to be consistent in its application throughout. Thus, genre is considered to serve not only a classificatory but also a clarificatory purpose. It is identifiable through both topical and non-topical descriptors. That is to say, what it contains and what it's about (topical) and the style of delivery as expressed through the language used (non-topical). So, equally important are topic, subject, content and domain within the topical, while the most pertinent among the non-topical would seem to be register as identified through the linguistic facets used and affecting the other areas of style, readability, generalisation and, to a certain extent, sentiment/opinion. Genre is therefore a classification system based on the features exhibited by a discourse.

Various methods were adopted in order to undertake the analysis, and this provided an opportunity to compare the effectiveness and computational cost of these. WordSmith produced very interesting word lists and allowed the generation of key-word lists, as well as providing concordancing, collocations and clustering information, which are all very useful as aids to qualitative analysis. However, these were quite general and did not allow the automatic generation of useful trends and statistics to any great extent. The TreeTagger is freely available and very fast at processing information. However, one major drawback is that if a word category is not known, it is randomly tagged with one, and often nouns and adjectives can be mis-represented if there are a lot of unknowns in the text file being processed. This problem can be overcome to a certain extent by using training data. The custom programme depended on this tagged data so its results were affected by this problem. However it was quite a simple programme, with uncomplicated code and it executed very effectively and efficiently. It was also highly extensible and flexible. The data generated by the custom programme however was
quite limited due to the size of the corpora. It was also relatively general because it was quantified only according to the forty eight part-of-speech tags, so this would have excluded certain areas from investigation. It was therefore decided to extend the programme to generate figures for time and place adverbials and punctuation marks, and a facility was added to allow the occurrence of particular words to be searched for as required. Another function allowed average word length to be calculated. The input file could also be determined dynamically, which allowed for greater flexibility. The output from this system could be generated directly to the console or in some cases to a ‘comma separated values’ (.csv) file for use in R. This was in order to be able to perform statistical analysis and generate graphical representations of the data. Spreadsheets were also used extensively to represent the data in tables and to perform various other calculations, including data for use in rank-ordering tests. It was thought and hoped that the Hyperbase programme would offer much more scope in its functionality than the custom programme. Unfortunately problems were encountered when trying to upload data as it seemed quite sensitive to anomalies and constantly went out of bounds just before completion. Notwithstanding that it uses TreeTagger for its tagging and lemmatising function, which worked fine on the same corpora when carried out as a stand-alone activity. This meant that the abridged corpora could not be uploaded and thus the database created using the original, unevenly distributed corpora had to be used. Therefore there were concerns about the data not being evenly distributed amongst the three corpora. This problem was addressed by weighting the results accordingly (see Appendix VI) and bearing in mind, when looking at graphical distributions, that the English corpus contained twice as much data as the other two. In fact, the problems encountered in trying to use Hyperbase led to a substantial increase in the amount of time and effort in developing the custom programme. This also resulted in less dependence on the Hyperbase system. In effect, the methods used were fairly cheap in terms of time and computational cost but taken altogether have facilitated extensive analysis work to be undertaken.

The results of the analysis, while being constructive and allowing salient features to be identified, demands the question as to whether automatic means are sufficient for any kind of analysis. They certainly highlight trends and give direction, but the final word
must be given to the value of intuition and manual intervention, even to a very small degree. For instance, hand-editing could be used to distinguish some of the finer detail. Concessions must be made however to the fact that recent developments as regards word sense disambiguation and the semantic web show that automatic means are being developed and improved all the time.

The results are reported in relation to the dimensions of linguistic variation that have been identified. These dimensions are characterised by particular linguistic features that co-occur in texts. The primary dimensions are:

(i) Interactive versus Edited Text
(ii) Abstract versus Situated Content
(iii) Reported versus Immediate Style

The results show the Mathematics corpus to be very abstract followed closely by the Geography corpus. The English corpus does not share this feature, tending much more towards the situated. (Deduced from the distribution of nouns.) The English text also shows a highly narrative style, followed by Geography and with Mathematics at the other end of the spectrum. (Deduced from the distribution of past tense verbs.) The Geography text shows a very descriptive style followed by English and again with Mathematics at the other extreme. (Deduced from the distribution of perfect tense verbs.) The level of immediacy as indicated by the distribution of present tense verbs tends also in this direction in general. They take similar positions with regard to the level of involvement, except that the Geography is located much nearer to the Mathematics end of the scale. (Deduced from the distribution of pronouns and 'it' count.) The Geography text is at the other end of the spectrum, giving it a more reported style. (Deduced from the distribution of third person present tense verbs.) The Geography text displays quite an academic style with the Mathematics and then the English lying a bit further away. (Deduced from the distribution of prepositions.) The distribution of place and time adverbials further support English as being situated and involved, followed by Geography although these findings will be influenced to a certain extent by the nature of that subject area. Thus one can conclude that the English text
shows a high level of involvement, is very situated with a narrative and immediate style and is fairly descriptive. It is not very academic. Geography is abstract with a reported style. It is very descriptive, very slightly involved and showing medium levels of both narrative and immediacy. It could be said to be quite academic. Mathematics on the other hand is abstract and does not appear immediate or involved. It does not use narrative or descriptive styles and is quite academic. While these results when taken in their totality might seem to support the hypothesis of being able to place them on a continuum with English and Mathematics at the extremes and Geography in the middle, this now seems too simplistic an interpretation, since there are much more complex factors at work here.

5.2 Future work:

This project opens up the whole area of the language of schooling. Schleppegrell (2001) remarks that “School based genres exhibit many common register features. These are compared with features that occur more frequently in the interactional discourse with which all children are more familiar. The ability to respond to linguistic expectations for academic registers depends on sociolinguistic skills that develop through social experience.” (Schleppegrell 2001:433). Therefore it might be interesting to analyse the interactional discourse of these school children in order to better understand the degree to which their sociolinguistic skills are actually suited to the classroom, making them more or less open to education. While it is recognised that the register used among school children will be different to that used in the classroom during teaching, which in turn will be different to that used in the school textbooks themselves, it might be worth comparing all three to better understand how much adaptation is required to switch from one to the other. It might also be interesting to consider the attitudes of the school children themselves to the fact of having to switch. Do they realise that a different register might set the scene for better learning, both in terms of a change of attitude in order to be receptive to knowledge and also to the fact that some registers are better than others in maximising the information to be imparted? These facts could improve their openness to the whole learning process.
Having examined the terms 'genre', 'text type', 'register', 'domain', and 'style', Lee (2001) considers the usefulness of these terms for researchers, or for teachers in the context of classroom unanimity. He suggests that it is worthwhile to start by looking at genres (as categories of texts) and to finish by generalising about the existence of registers (in terms of linguistic characteristics). These linguistic features can then be used to source materials useful for teaching and learning activities alike.

Finally, to quote Cranny-Francis (1993): “A concept of genre allows the critic or analyst to explore [the] complex relationships in which a text is involved, relationships which ultimately relate back to what a text means. This is because what a text says and how it says it cannot be separated: this is fundamental to our notion of genre. Because of this, genre provides the link between text and context; between the formal and semantic properties of texts; between the text and the intertextual, disciplinary and technological practices in which it is embedded.” (111-113).
Bibliography:


Appendix I: Information regarding the corpora.

This corpus itself is unique in that it consists of some of the actual documents currently in use in the Irish education system. The complete corpus presently stands at 5.6 million words and is still growing as more texts are added. These are taken from post-primary syllabi, and have been scanned from textbooks as part of the English Language Support Programme, which provides online language support primarily for post-primary students whose first language is not English. This programme is itself part of the Trinity Immigration Initiative, a research programme on diversity, integration and policy.

As mentioned previously, the subjects chosen for the project were English, Geography and Mathematics. The textbooks used were all taken from the current Irish Junior Certificate syllabus. The textbooks included are as follows:


Geography: Junior Cert Rapid Revision Geography, New GEO JC Geography, The Human Planet, Exam Papers (Edco).

Mathematics: Concise Maths, New Concise Maths, Shortcuts to Success, Exam Papers and Solutions.

The textbooks were scanned and then converted to plain text. Because of this, problems were sometimes caused in that characters were occasionally misinterpreted. This was not considered to be significant in terms of accuracy. So a corpus was compiled for each of the subject areas individually. The texts were then combined to form a text file comprising all three subjects. They were further modified to compile texts of English and Geography; English and Mathematics; and Geography and English for comparison purposes. The final corpora of English, Geography and Mathematics contained 426474, 186346, and 265789 words respectively. These corpora were generally used with the TreeTagger and the WordSmith programs. Because of problems uploading the texts into the Hyperbase programme for treatment, they were further modified and abridged versions were compiled. It was decided to take a random
sample from each original plain text file in order to compile these abridged versions, for input to the Hyperbase programme. The Mathematics and Geography texts were further modified manually, to remove ambiguous areas, or those which it was feared were causing the programme to abort the upload. This was not considered to be compromising to the integrity of the text itself, as the programme is more concerned with words and syntax rather than numbers and algebra! The texts were further modified to eliminate the occurrences of angle brackets, which caused problems when being tagged. (See Appendix II in section concerning TreeTagger.) Since the programme it was hoped to use requires four texts as input, the abridged Mathematics, Geography and English texts were combined to form a fourth. The final abridged corpora for English, Geography and Mathematics contain 54484, 53535, and 54088 words respectively. Unfortunately, due to persistent problems when trying to upload to the programme, Hyperbase, hopes to use the four corpora as outlined just above were annulled. As it now stands, the abridged corpora of English, Geography and Mathematics were used consistently with the TreeTagger for quantification by the custom programme. However, as regards Hyperbase, only the original database containing three English texts and two texts each for Geography and Mathematics as prepared by Etienne Brunet is available for use. This database is therefore quite uneven in its representation, as discussed further in Appendix II. Therefore it will not form part of the main analysis and will only be referred to for supplemental information or for comparative purposes, using the weighted data as presented in Appendix VI.
Appendix II: Analysis Software.

As stated in Appendix I, because the textbooks were scanned and then converted to plain text, this sometimes caused problems in that characters were occasionally misinterpreted. Another problem area was that maps, graphs, diagrams and tables were rendered meaningless when presented as extensible mark-up language (XML) since these appear as rows of figures and characters or other insignificant symbols. And other formatting features were also converted ineffectively. While this was not considered to be significant in terms of accuracy within the texts themselves, it often caused problems when the text was being tagged and lemmatised. In particular the Mathematics text caused problems, especially since solutions to questions were often listed in numerical and algebraic terms in whole sections at the end of each textbook. For this, abridged corpora were compiled as outlined in Appendix I.

TreeTagger:

As the name suggests, TreeTagger parses input texts and tags them with part-of-speech information using decision trees. The annotation system is that used by the Penn Treebank Project, and a list of tags with corresponding parts of speech is given in Appendix III. Closed classes are identified directly and others depend on the dictionary for their identification. The training was carried out on the Penn Treebank itself so it is considered to be quite comprehensive. “The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics at The University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese and old French texts and is adaptable to other languages if a lexicon and a manually tagged training corpus are available.” 4

4 Information provided by http://www.ims.uni-stuttgart.de.
The TreeTagger was used for various functions within this project. The tagged text however had to be manually modified to identify and eliminate whole sections that had been ignored since they were taken to be a single word by the tagger once enclosed by angle brackets. This decision was later reversed in case it caused misrepresentation, and the abridged corpora, as outlined in Appendix I, were manually modified to eliminate the occurrences of angle brackets, before being re-treated by the TreeTagger. Consideration was also given as to whether it might be necessary to compile a set of training data from the corpora, by manually tagging a selection from each of them. A very limited test corpus was used from each of these and processed by the TreeTagger. This preliminary test showed that the tagger didn't recognise certain words, notably the apostrophe used to indicate possessive endings and also some hyphenated words. However, an error rate of 2.44% for the English text, 1.49% for the Geography text and 1.42% for the Mathematics text did not seem to warrant such an action. This test corpus consisted of the course overview for each subject and samples of the type of text found in them, taken from documents on the website 'skool.ie'.

Custom Programme:

It was decided to write a custom programme in order to quantify the various tags identified by the TreeTagger. Firstly, the three column output consisting of the word (token), the part-of-speech code (tag) and the base form of the word (lemma) were imported into a spreadsheet. The column consisting of first the tag, then the lemma and later the token could then be isolated and exported to a text file. Each text file was further processed by the programme which performed various functions on them. The main function was basically to count the number of times that each tag occurred. This was done for each of the corpora in turn and a table comparing the results is shown in Table 2, in Chapter 4.2, in the section entitled 'Quantitative Analysis'. Then other features were added, for example punctuation marks could be quantified from the lemma text, and a further function was written which allowed any word to be searched for and quantified from the token text. This was extended still further to search for and quantify classes of words, such as time and place adverbials, etc. Another function to calculate word length was also included. Further extensions enabled the input file to be
entered dynamically to give greater flexibility. The output stream was further modified so that it could be sent to both the console and also in some cases to a 'comma separated values' (.csv) file for better compatibility when working with R, a statistical analysis programme that is discussed further on. Thus, together with the use of spreadsheets and R for statistics and graphical representation, the programme fulfilled a lot of the functions available in Hyperbase. While the data generated was limited by the size of the corpora, no problems would be envisaged if this was increased. That is to say that both the size of the corpora and the number of them could be increased as much as required. The programme is quite simple in terms of code, while executing effectively and efficiently. It is also highly extensible and flexible. Because it was extended in a piece-meal manner, there are several programs included. If this broader scope had been planned from the beginning, more generic code could have been used, with the main method calling methods from other classes and resulting in a more elegant programme. However, as it became apparent that Hyperbase would not be available for use in a reliable manner, or allowing like to be compared with like, an attempt was made to match a lot of the functionality normally available from it. It is felt that this has been achieved to a very large extent. In fact, when using Hyperbase for a previous project there were times when an ability to 'tweak' the code to enable other features to be searched for would have been useful, and this has now been achieved by the custom programme.

WordSmith:

WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. The WordList function generates word lists based on one or more plain text files. Word lists can be shown both in alphabetical and frequency order. The concordancer, Concord, shows any word or phrase in context -- to see what sort of company it keeps. It makes a concordance using DOS, Text Only, ASCII or ANSI text files. This allows the specification of a search word, which Concord will seek in all the text files chosen. It will then present a concordance display, and gives access to information about collocates of the search word and also of clustering patterns. The purpose of the KeyWords function is to locate and identify key words in a given text. To do so, it compares the words in
the text with a reference set of words usually taken from a large corpus of text. Any word which is found to be outstanding in its frequency in the text is considered "key". The key words are presented in order of frequency. This function needs access to two or more word lists, which must be created first, using the Word List programme. The distribution of the key words can be plotted. The tools have been used by Oxford University Press for their own lexicographic work in preparing dictionaries, by language teachers and students, and by researchers investigating language patterns in lots of different languages in many countries world-wide. Listings can be saved for later use, edited, printed, copied or saved as text files.  

Hyperbase:

The computer programme used to upload, tag, and carry out the factorial analysis of the texts is called 'Hyperbase'. Hyperbase originated in 1989 at the University of Nice, and was conceived by Étienne Brunet. In reality, the documentary and statistical functions were updated at this time, to develop an object-oriented programme from existing programs which were written in a procedural programming language. The present programme enables the realisation of 'hypertextual' databases of texts, with the ability to treat documents input in ASCII, XML, or HTML. It meets the classic needs of automatic text treatment providing the tools for processes such as selective or systematic indexing, dictionary of frequencies, concordance and compatibility tests, selection of wider context, co-occurrences, search for parts or groups of words; and it is statistically orientated. Hyperbase uses the TreeTagger for its tagging and parsing function.

Some problems were encountered when trying to upload corpora into the programme. While it uses TreeTagger to tag and lemmatise, apparently it's very sensitive to any anomalies within the data set. Although the abridged corpora, as described in the previous section, were successfully tagged and lemmatised by TreeTagger, before being successfully quantified by the custom programme, it was not possible to upload these into Hyperbase. The system continuously went out of bounds and failed to complete the creation of the database. The original corpora which had been sent to

Etienne Brunet, the programme designer, were however successfully uploaded by him and the completed database was returned for use. The only problem with this was that the corpora were not the same size, so it meant that like with like could not be compared. Basically, the English corpus is 51.85% of the whole, while the Geography corpus is 23.28% and the Mathematics corpus is 24.87%. Values were calculated for the English, Geography and Mathematics corpora in order to arrive at weights for each of them. These were found to be 2.07, 0.93 and 1.00, respectively. A spreadsheet showing the original values for the different syntactic structures, the calculations to arrive at the weights, and the revised values and percentages are included in Appendix VI. Relativised values were also calculated where the individual features in an individual corpus was compared with the total features in that same corpus, but since it had been decided not to focus on Hyperbase as the main data source, these were mainly used for comparison with the custom programme data. Other than that, one must remember when remarking on distributions displayed in graphs, that the English corpus is twice as big as the other two. Because of these problems, the use of Hyperbase was kept to a minimum.

R:

Some of the statistical analysis functions were carried out using 'R', which is free software. “R is a language and environment for statistical computing and graphics. It is a GNU Project, which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. R provides a wide variety of statistical functions (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and graphical techniques), and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.” 6

6 http://www.r-project.org/about.html
Appendix III: Tagset.

This is a list of the parts of speech encoded according to the Penn Treebank Project annotation system. The corresponding abbreviations, or tags, are listed alphabetically beside each part of speech. This is the same system used by the TreeTagger, although it should be noted that there are other annotation systems available. It should also be noted that there are linguistic issues that arise in connection with annotating texts by parts of speech. For further information the reader is directed to an article by Beatrice Santorini (1991).
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Part-of-Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>EX</td>
<td>Existential <em>there</em></td>
</tr>
<tr>
<td>FW</td>
<td>Foreign word</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition or coordinating conjunction</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>JJR</td>
<td>Adjective, comparative</td>
</tr>
<tr>
<td>JJS</td>
<td>Adjective, superlative</td>
</tr>
<tr>
<td>LS</td>
<td>List item marker</td>
</tr>
<tr>
<td>MD</td>
<td>Modal</td>
</tr>
<tr>
<td>NN</td>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>NNS</td>
<td>Noun Plural</td>
</tr>
<tr>
<td>NP</td>
<td>Proper noun, singular</td>
</tr>
<tr>
<td>NPS</td>
<td>Proper noun, plural</td>
</tr>
<tr>
<td>PDT</td>
<td>Predeterminer</td>
</tr>
<tr>
<td>POS</td>
<td>Possessive ending</td>
</tr>
<tr>
<td>PP</td>
<td>Personal Pronoun</td>
</tr>
<tr>
<td>PP$</td>
<td>Possessive Pronoun</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
</tr>
<tr>
<td>RBR</td>
<td>Adverb, comparative</td>
</tr>
<tr>
<td>RBS</td>
<td>Adverb, superlative</td>
</tr>
<tr>
<td>RP</td>
<td>Particle</td>
</tr>
<tr>
<td>SYM</td>
<td>Symbol</td>
</tr>
<tr>
<td>TO</td>
<td><em>to</em></td>
</tr>
<tr>
<td>UH</td>
<td>Interjection</td>
</tr>
<tr>
<td>VB, VH, VV</td>
<td>Verb, base form</td>
</tr>
<tr>
<td>VBD, VHD, VVD</td>
<td>Verb, past tense</td>
</tr>
<tr>
<td>VBG, VHG, VVG</td>
<td>Verb, gerund or present participle</td>
</tr>
<tr>
<td>VBN, VHN, VVN</td>
<td>Verb, past participle</td>
</tr>
<tr>
<td>VBP, VHP, VVP</td>
<td>Verb, non-3rd person singular, present</td>
</tr>
<tr>
<td>VBZ, VHZ, VVZ</td>
<td>Verb, 3rd person singular, present</td>
</tr>
<tr>
<td>WDT</td>
<td>Wh-determiner</td>
</tr>
<tr>
<td>WP</td>
<td>Wh-pronoun</td>
</tr>
<tr>
<td>WP$</td>
<td>Possessive wh-pronoun</td>
</tr>
<tr>
<td>WRB</td>
<td>Wh-adverb</td>
</tr>
</tbody>
</table>

This list has been updated to include the 'have' and 'normal' forms of the verbs, since the original list included only the codes for the 'be' form.

Updated information supplied by: http://trac.sketchengine.co.uk/wiki/tagsets/penn
import java.io.*;

/* This is a programme that reads a file containing
 * part-of-speech (tag) codes and quantifies them
 * for output to file and console */

class QuantifyTags {

    public static void main(String args[]){
        // create variable to store file name, entered dynamically
        String fileName = " ";
        // create a variable to hold tag from each line
        String tag = " ";
        // create variables to hold tag counts
        int countNN = 0;
        int countNNS = 0;
        int countNP = 0;
        int countNPS = 0;
        int countJJ = 0;
        int countJJR = 0;
        int countJJS = 0;
        int countCC = 0;
        int countDT = 0;
        int countIN = 0;
        int countCD = 0;
        int countEX = 0;
        int countFW = 0;
        int countLS = 0;
        int countMD = 0;
        int countPDT = 0;
        int countPOS = 0;
        int countPP = 0;
        int countPP$ = 0;
        int countRB = 0;
        int countRBR = 0;
        int countRBS = 0;
        int countRP = 0;
        int countSYM = 0;
        int countTO = 0;
        int countUH = 0;
        int countVB = 0;
        int countVBD = 0;
        int countVBG = 0;
        int countVBN = 0;
        int countVP = 0;
        int countVZ = 0;
        int countVH = 0;
        int countVHD = 0;
        int countVHG = 0;
        int countVHN = 0;
        int countVHP = 0;
        int countVHZ = 0;
        int countVV = 0;
        int countVVD = 0;
        int countVVG = 0;
int countVVN = 0;
int countVVP = 0;
int countVVZ = 0;
int countWDT = 0;
int countWP = 0;
int countWP$ = 0;
int countWRB = 0;

try {
    // get input file name from user
    System.out.println("enter file name: ");
    BufferedReader in = new BufferedReader(new InputStreamReader(System.in));
    fileName = in.readLine();

    // open file using file name entered
    FileInputStream fis = new FileInputStream(fileName);

    // create data input stream and pass in fis as argument
    DataInputStream dis = new DataInputStream(fis);

    // create buffered reader and pass in input
    // stream reader containing dis as argument
    BufferedReader br = new BufferedReader(new InputStreamReader(dis));

    // create output file
    String outputName = "csv_tags.csv";

    // wrap output file in file writer
    FileWriter fw = new FileWriter(outputName, true);

    // and send to buffered writer
    BufferedWriter bw = new BufferedWriter(fw);

    // read file one line at a time
    while ((tag = br.readLine()) != null) {
        // quantify relevant tags
        if (tag.equals("NN")) {
            countNN ++;
        }
        if (tag.equals("NNS")) {
            countNNS ++;
        }
        if (tag.equals("NP")) {
            countNP ++;
        }
        if (tag.equals("NPS")) {
            countNPS ++;
        }
        if (tag.equals("JJ")) {
            countJJ ++;
        }
        if (tag.equals("JJR")) {
            countJJR ++;
        }
        if (tag.equals("JJS")) {
            countJJS ++;
        }
        if (tag.equals("JJS")) {
            countJJS ++;
        }
    }
}
if(tag.equals("CC")){
    countCC ++;
}
if(tag.equals("DT")){
    countDT ++;
}
if(tag.equals("IN")){
    countIN ++;
}
if(tag.equals("CD")){
    countCD ++;
}
if(tag.equals("EX")){
    countEX ++;
}
if(tag.equals("FW")){
    countFW ++;
}
if(tag.equals("LS")){
    countLS ++;
}
if(tag.equals("MD")){
    countMD ++;
}
if(tag.equals("PDT")){
    countPDT ++;
}
if(tag.equals("POS")){
    countPOS ++;
}
if(tag.equals("PP")){
    countPP ++;
}
if(tag.equals("PP$")){
    countPP$ ++;
}
if(tag.equals("RB")){
    countRB ++;
}
if(tag.equals("RBR")){
    countRBR ++;
}
if(tag.equals("RBS")){
    countRBS ++;
}
if(tag.equals("RP")){
    countRP ++;
}
if(tag.equals("SYM")){
    countSYM ++;
}
if(tag.equals("TO")){
    countTO ++;
}
if(tag.equals("UH")){
    countUH ++;
}
if(tag.equals("VB")){
    countVB ++;
}
if(tag.equals("VBD")){
    countVBD ++;
}
if(tag.equals("VBG")){
    countVBG ++;
}
if(tag.equals("VBN")){
    countVBN ++;
}
if(tag.equals("VBP")){
    countVBP ++;
}
if(tag.equals("VBZ")){
    countVBZ ++;
}
if(tag.equals("VH")){
    countVH ++;
}
if(tag.equals("VHD")){
    countVHD ++;
}
if(tag.equals("VHG")){
    countVHG ++;
}
if(tag.equals("VHN")){
    countVHN ++;
}
if(tag.equals("VHP")){
    countVHP ++;
}
if(tag.equals("VHZ")){
    countVHZ ++;
}
if(tag.equals("VV")){
    countVV ++;
}
if(tag.equals("VVD")){
    countVVD ++;
}
if(tag.equals("VVG")){
    countVVG ++;
}
if(tag.equals("VVN")){
    countVVN ++;
}
if(tag.equals("VVP")){
    countVVP ++;
}
if(tag.equals("VVZ")){
    countVVZ ++;
}
if(tag.equals("WDT")){
    countWDT ++;
}
if(tag.equals("WP")){
  countWP ++;
}
if(tag.equals("WP$")){
  countWP$ ++;
}
if(tag.equals("WRB")){
  countWRB ++;
}

// close stream
dis.close();

// write results to file
bw.write(fileName + "," + countNN + "," + countNNS + "," + countNP + "," + countNPS + "," + countJJ + "," + countJJR + "," + countJJS + "," + countCC + "," + countDT + "," + countIN + "," + countCD + "," + countMD + "," + countEX + "," + countFW + "," + countLS + "," + countPDT + "," + countPOS + "," + countPP + "," + countPP$ + "," + countRB + "," + countRBR + "," + countRBS + "," + countRP + "," + countSYM + "," + countTO + "," + countUH + "," + countVB + "," + countVBD + "," + countVBG + "," + countVBN + "," + countVBZ + "," + countVH + "," + countVHD + "," + countVHN + "," + countVHP + "," + countVHZ + "," + countVVD + "," + countVVG + "," + countVVN + "," + countVVP + "," + countVVZ + "," + countWDT + "," + countWP + 
  + countWRB + 
  + countVHP + 
  + countVH + 
  + countVHD + 
  + countVHN + 
  + countVHP + 
  + countVHZ + 
  + countVVD + 
  + countVVG + 
  + countVVN + 
  + countVVP + 
  + countVVZ + 
  + countWDT + 
  + countWP +
  + countWRB + 
}

// close files
bw.close();
fw.close();
}

// catch exception if any
catch (Exception e){
  System.err.println("Error: "+ e.getMessage());
}

// print totals to console
System.out.println("Results: " + fileName);
System.out.println("Nouns: " + countNN);
System.out.println("Nouns, plural: " + countNNS);
System.out.println("Nouns, proper: " + countNP);
System.out.println("Nouns, proper, plural: " + countNPS);
System.out.println("Adjectives: " + countJJ);
System.out.println("Adjectives, comparative: " + countJJR);
System.out.println("Adjectives, superlative: " + countJJS);
System.out.println("Coordinators: " + countCC);
System.out.println("Determiners: " + countDT);
System.out.println("Prepositions: " + countIN);
System.out.println("Cardinals: " + countCD);
System.out.println("Existensial 'there's: " + countEX);
System.out.println("Foreign words: " + countFW);
System.out.println("List items: " + countLS);
System.out.println("Modals: " + countMD);
System.out.println("Predeterminers: " + countPDT);
System.out.println("Possessive endings: " + countPOS);
System.out.println("Personal pronouns: " + countPP);
System.out.println("Possessive pronouns: " + countPP$);
System.out.println("Adverbs: " + countRB);
System.out.println("Adverbs, comparative: " + countRBR);
System.out.println("Adverbs, superlative: " + countRBS);
System.out.println("Particles: " + countRP);
System.out.println("Symbols: " + countSYM);
System.out.println("'To's: " + countTO);
System.out.println("Interjections: " + countUH);
System.out.println("Verbs-be, base form: " + countVB);
System.out.println("Verbs-be, past tense: " + countVBD);
System.out.println("Verbs-be, gerund or present participle: " + countVBG);
System.out.println("Verbs-be, past participle: " + countVBN);
System.out.println("Verbs-be, 3rd person singular, present: " + countVBZ);
System.out.println("Verbs-have, base form: " + countVH);
System.out.println("Verbs-have, past tense: " + countVHD);
System.out.println("Verbs-have, gerund or present participle: " + countVHG);
System.out.println("Verbs-have, past participle: " + countVHN);
System.out.println("Verbs-have, non 3rd person singular, present: " + countVHP);
System.out.println("Verbs-have, 3rd person singular, present: " + countVHZ);
System.out.println("Verbs, base form: " + countVV);
System.out.println("Verbs, past tense: " + countVVD);
System.out.println("Verbs, gerund or present participle: " + countVVG);
System.out.println("Verbs, past participle: " + countVVN);
System.out.println("Verbs, non 3rd person singular, present: " + countVVP);
System.out.println("Verbs, 3rd person singular, present: " + countVVZ);
System.out.println("Wh-determiners: " + countWDT);
System.out.println("Wh-pronouns: " + countWP);
System.out.println("Possessive wh-pronouns: " + countWP$);
System.out.println("Wh-adverbs: " + countWRB);
import java.io.*;

/* This is a programme that reads a file containing tokens and quantifies relevant place adverbials */

class QuantifyPlaceAdv {
    public static void main(String args[])
    {
        // create variable to store file name, entered dynamically
        String fileName = " ";
        // create a variable to hold tag from each line
        String word = " ";
        // create variables to store tag counts
        int countABD = 0;
        int countABV = 0;
        int countABR = 0;
        int countACR = 0;
        int countAHD = 0;
        int countALS = 0;
        int countARN = 0;
        int countASH = 0;
        int countAST = 0;
        int countAW = 0;
        int countBD = 0;
        int countBW = 0;
        int countBT = 0;
        int countBS = 0;
        int countDL = 0;
        int countDS = 0;
        int countDM = 0;
        int countET = 0;
        int countFR = 0;
        int countHA = 0;
        int countIS = 0;
        int countIL = 0;
        int countISH = 0;
        int countINS = 0;
        int countLO = 0;
        int countNR = 0;
        int countNY = 0;
        int countNT = 0;
        int countNW = 0;
        int countOD = 0;
        int countOT = 0;
        int countOB = 0;
        int countOL = 0;
        int countOS = 0;
        int countSTH = 0;
        int countUF = 0;
        int countUG = 0;
        int countUN = 0;
        int countUH = 0;
        int countUS = 0;
        int countUM = 0;
        int countWT = 0;
        try
        {
            // get input file name from user
System.out.println("enter file name: ");
BufferedReader in = new BufferedReader(new InputStreamReader(System.in));
fileName = in.readLine();

// open file using file name entered
FileInputStream fis = new FileInputStream(fileName);

// create data input stream and pass in fis as argument
DataInputStream dis = new DataInputStream(fis);

// create buffered reader and pass in input
// stream reader containing dis as argument
BufferedReader br = new BufferedReader(new InputStreamReader(dis));

// create output file
String outputName = "csv_place.csv";

// wrap output file in file writer
FileWriter fw = new FileWriter(outputName, true);

// and send to buffered writer
BufferedWriter bw = new BufferedWriter(fw);

// read file one line at a time
while ((word = br.readLine()) != null) {
    // quantify relevant tags
    if (word.equals("aboard")) {
        countABD ++;
    }
    if (word.equals("above")) {
        countABV ++;
    }
    if (word.equals("abroad")) {
        countABR ++;
    }
    if (word.equals("across")) {
        countACR ++;
    }
    if (word.equals("ahead")) {
        countAHD ++;
    }
    if (word.equals("alongside")) {
        countALS ++;
    }
    if (word.equals("around")) {
        countARN ++;
    }
    if (word.equals("ashore")) {
        countASH ++;
    }
    if (word.equals("astern")) {
        countAST ++;
    }
    if (word.equals("away")) {
        countAW ++;
    }
}
if (word.equals("behind")) {
    countBD ++;
}
if (word.equals("below")) {
    countBW ++;
}
if (word.equals("beneath")) {
    countBT ++;
}
if (word.equals("beside")) {
    countBS ++;
}
if (word.equals("downhill")) {
    countDL ++;
}
if (word.equals("downstairs")) {
    countDS ++;
}
if (word.equals("downstream")) {
    countDM ++;
}
if (word.equals("east")) {
    countET ++;
}
if (word.equals("far")) {
    countFR ++;
}
if (word.equals("hereabouts")) {
    countHA ++;
}
if (word.equals("indoors")) {
    countIS ++;
}
if (word.equals("inland")) {
    countIL ++;
}
if (word.equals("inshore")) {
    countISH ++;
}
if (word.equals("inside")) {
    countINS ++;
}
if (word.equals("locally")) {
    countLO ++;
}
if (word.equals("near")) {
    countNR ++;
}
if (word.equals("nearby")) {
    countNY ++;
}
if (word.equals("north")) {
    countNT ++;
}
if (word.equals("nowhere")) {
    countNW ++;
}
if (word.equals("outdoors")){
    countOD ++;
}
if (word.equals("outside")){
    countOT ++;
}
if (word.equals("overboard")){
    countOB ++;
}
if (word.equals("overland")){
    countOL ++;
}
if (word.equals("overseas")){
    countOS ++;
}
if (word.equals("south")){
    countSTH ++;
}
if (word.equals("underfoot")){
    countUF ++;
}
if (word.equals("underground")){
    countUG ++;
}
if (word.equals("underneath")){
    countUN ++;
}
if (word.equals("uphill")){
    countUH ++;
}
if (word.equals("upstairs")){
    countUS ++;
}
if (word.equals("upstream")){
    countUM ++;
}
if (word.equals("west")){
    countWT ++;
}
// close stream
dis.close();

// write results to file
bw.write(fileName + "," + countABD + "," + countABV + "," + countABR + "," + countACR + "," + countAHD + "," + countALS + "," + countASH + "," + countAST + "," + countAW + "," + countBW + "," + countBT + "," + countBS + "," + countDS + "," + countDM + "," + countET + "," + countHA + "," + countIS + "," + countIL + ","
+ countNY + "," + countINS + "," + countLO + "," + countNR + ","
+ countOT + ","
+ countSTH + ","
+ countUH + ","
+ "\n");

    // close files
    bw.close();
    fw.close();
}

    // catch exception if any
    catch (Exception e){
        System.err.println("Error: " + e.getMessage());
    }

    // print totals to console
    System.out.println("Results: " + fileName);
    System.out.println("aboard: " + countABD);
    System.out.println("above: " + countABV);
    System.out.println("abroad: " + countABR);
    System.out.println("across: " + countACR);
    System.out.println("ahead: " + countAHD);
    System.out.println("alongside: " + countALS);
    System.out.println("around: " + countARN);
    System.out.println("ashore: " + countASH);
    System.out.println("astern: " + countAST);
    System.out.println("away: " + countAW);
    System.out.println("behind: " + countBD);
    System.out.println("below: " + countBW);
    System.out.println("beneath: " + countBT);
    System.out.println("beside: " + countBS);
    System.out.println("downhill: " + countDL);
    System.out.println("downstairs: " + countDS);
    System.out.println("downstream: " + countDM);
    System.out.println("east: " + countET);
    System.out.println("far: " + countFR);
    System.out.println("hereabouts: " + countHA);
    System.out.println("indoors: " + countIS);
    System.out.println("inland: " + countIL);
    System.out.println("inshore: " + countISH);
    System.out.println("inside: " + countINS);
    System.out.println("locally: " + countLO);
    System.out.println("near: " + countNR);
    System.out.println("nearby: " + countNY);
    System.out.println("north: " + countNT);
    System.out.println("nowhere: " + countNW);
    System.out.println("outdoors: " + countOD);
    System.out.println("outside: " + countOT);
    System.out.println("overboard: " + countOB);
    System.out.println("overland: " + countOL);
    System.out.println("overseas: " + countOS);
    System.out.println("south: " + countSTH);
System.out.println("underfoot: " + countUF);
System.out.println("underground: " + countUG);
System.out.println("underneath: " + countUN);
System.out.println("uphill: " + countUH);
System.out.println("upstairs: " + countUS);
System.out.println("upstream: " + countUM);
System.out.println("west: " + countWT);

import java.io.*;

/* This is a programme that reads a file containing tokens and quantifies relevant time adverbials */

class QuantifyTimeAdv {
    
    public static void main(String[] args) {
        // create variable to store file name, entered dynamically
        String fileName = " ";
        // create a variable to hold tag from each line
        String word = " ";
        // create variables to store tag counts
        int countAF = 0;
        int countAG = 0;
        int countER = 0;
        int countEY = 0;
        int countEV = 0;
        int countFY = 0;
        int countIM = 0;
        int countIN = 0;
        int countIY = 0;
        int countLT = 0;
        int countLY = 0;
        int countLR = 0;
        int countMY = 0;
        int countNW = 0;
        int countNY = 0;
        int countON = 0;
        int countOY = 0;
        int countPR = 0;
        int countPV = 0;
        int countRY = 0;
        int countSH = 0;
        int countSM = 0;
        int countSN = 0;
        int countSB = 0;
        int countTD = 0;
        int countTM = 0;
        int countTN = 0;
        int countYD = 0;

        try {
            // get input file name from user
            System.out.println("enter file name: ");
        }
    }
}
BufferedReader in = new BufferedReader(new InputStreamReader(System.in));
fileName = in.readLine();

// open file using file name entered
FileInputStream fis = new FileInputStream(fileName);

// create data input stream and pass in fis as argument
DataInputStream dis = new DataInputStream(fis);

// create buffered reader and pass in input
// stream reader containing dis as argument
BufferedReader br = new BufferedReader(new InputStreamReader(dis));

// create output file
String outputName = "csv_time.csv";

// wrap output file in file writer
FileWriter fw = new FileWriter(outputName, true);

// and send to buffered writer
BufferedWriter bw = new BufferedWriter(fw);

// read file one line at a time
while ((word = br.readLine()) != null) {
    // quantify relevant tags
    if (word.equals("afterwards")) {
        countAF ++;
    }
    if (word.equals("again")) {
        countAG ++;
    }
    if (word.equals("earlier")) {
        countER ++;
    }
    if (word.equals("early")) {
        countEY ++;
    }
    if (word.equals("eventually")) {
        countEV ++;
    }
    if (word.equals("formerly")) {
        countFY ++;
    }
    if (word.equals("immediately")) {
        countIM ++;
    }
    if (word.equals("initially")) {
        countIN ++;
    }
    if (word.equals("instantly")) {
        countIY ++;
    }
    if (word.equals("late")) {
        countLT ++;
    }
}
if (word.equals("lately")) {
    countLY ++;
}
if (word.equals("later")) {
    countLR ++;
}
if (word.equals("momentarily")) {
    countMY ++;
}
if (word.equals("now")) {
    countNW ++;
}
if (word.equals("nowadays")) {
    countNY ++;
}
if (word.equals("once")) {
    countON ++;
}
if (word.equals("originally")) {
    countOY ++;
}
if (word.equals("presently")) {
    countPR ++;
}
if (word.equals("previously")) {
    countPV ++;
}
if (word.equals("recently")) {
    countRY ++;
}
if (word.equals("shortly")) {
    countSH ++;
}
if (word.equals("simultaneously")) {
    countSM ++;
}
if (word.equals("soon")) {
    countSN ++;
}
if (word.equals("subsequently")) {
    countSB ++;
}
if (word.equals("today")) {
    countTD ++;
}
if (word.equals("tomorrow")) {
    countTM ++;
}
if (word.equals("tonight")) {
    countTN ++;
}
if (word.equals("yesterday")) {
    countYD ++;
}

// close stream
dis.close();

// write results to file
bw.write(fileName + "," + countAF + "," + countAG + "," +
        countER + "," + countIM + "," + countLY + "," +
        countNY + "," + countPV + "," + countSN + "," +
        countTN + ","
        + countFY + "," + countEV + "," + countMY + "," +
        countNW + "," +
        countIN + "," + countLY + "," + countNY + "," +
        countON + "," + countPR + "," +
        countSB + "," + countTD + "," + countTM + "," +
        countYD + "," + "\n");

// close files
bw.close();
fw.close();

} // catch exception if any
    catch (Exception e){
        System.err.println("Error: " + e.getMessage());
    }

// print totals to console
System.out.println("Results: " + fileName);
System.out.println("afterwards: " + countAF);
System.out.println("again: " + countAG);
System.out.println("earlier: " + countER);
System.out.println("early: " + countEY);
System.out.println("eventually: " + countEV);
System.out.println("formerly: " + countFY);
System.out.println("immediately: " + countIM);
System.out.println("initially: " + countIN);
System.out.println("instantly: " + countIY);
System.out.println("late: " + countLT);
System.out.println("lately: " + countLY);
System.out.println("later: " + countLR);
System.out.println("momentarily: " + countMY);
System.out.println("now: " + countNW);
System.out.println("nowadays: " + countNY);
System.out.println("once: " + countON);
System.out.println("originally: " + countOY);
System.out.println("presently: " + countPR);
System.out.println("previously: " + countPV);
System.out.println("recently: " + countFY);
System.out.println("shortly: " + countSH);
System.out.println("simultaneously: " + countSM);
System.out.println("soon: " + countSN);
System.out.println("subsequently: " + countSB);
System.out.println("today: " + countTD);
System.out.println("tomorrow: " + countTM);
System.out.println("today: " + countTN);
import java.io.*;

/* This is a programme that reads a file containing base forms (lemmas) * and quantifies punctuation marks, and other markers considered * relevant, for output to console and file */

class QuantifyLemma {

    public static void main(String args[]){
        // create variable to store file name, entered dynamically
        String fileName = " ";
        // create a variable to hold lemma from each line
        String lem = " ";
        // create variables to store lemma counts
        int countQM = 0;
        int countXC = 0;
        int countCN = 0;
        int countSC = 0;
        int countCA = 0;
        int countFS = 0;
        int countAP = 0;
        int countDQ = 0;
        int countOP = 0;
        int countCP = 0;
        int countUK = 0;

        try{
            // get input file name from user
            System.out.println("enter file name: ");
            BufferedReader in = new BufferedReader(new InputStreamReader(System.in));
            fileName = in.readLine();

            // open file using file name entered
            FileInputStream fis = new FileInputStream(fileName);

            // create data input stream and pass in fis as argument
            DataInputStream dis = new DataInputStream(fis);

            // create buffered reader and pass in input stream reader
            BufferedReader br = new BufferedReader(new InputStreamReader(dis));

            // create output file
            String outputName = "csv_lems.csv";

            // wrap output file in file writer
            FileWriter fw = new FileWriter(outputName, true);

            // and send to buffered writer
            System.out.println("yesterday: "+ countYD);
        }
    }
}
BufferedWriter bw = new BufferedWriter(fw);

// read file one line at a time
while ((lem = br.readLine()) != null)   {
    // quantify lemmas
    if(lem.equals(“?”)) {
        countQM ++;
    }
    if(lem.equals(“!”)) {
        countXC ++;
    }
    if(lem.equals(“:”)) {
        countCN ++;
    }
    if(lem.equals(“,”) ) {
        countSC ++;
    }
    if(lem.equals(“,”)) {
        countCA ++;
    }
    if(lem.equals(“.”)) {
        countFS ++;
    }
    if(lem.equals(“!”)) {
        countAP ++;
    }
    if(lem.equals(“\”)) {
        countDQ ++;
    }
    if(lem.equals(“(”) ) {
        countOP ++;
    }
    if(lem.equals(“<unknown>”)) {
        countUK ++;
    }
}

// close stream
dis.close();

// write results to file
bw.write(fileName + “,” + countQM + “,” + countXC + “,” +
          countCN + “,” + countSC + “,” + countCA + “,” + countFS + “,” +
          countAP + “,” + countDQ + “,” + countOP + “,” + countCP + “,” +
          countUK + “\n”);

// close files
bw.close();
fw.close();

// catch exception if any
catch (Exception e){
    System.err.println(“Error: “ + e.getMessage());
}
// write results to console
System.out.println("Question marks: " + countQM);
System.out.println("Exclamation marks: " + countXC);
System.out.println("Colons: " + countCN);
System.out.println("Semi-colons: " + countSC);
System.out.println("Full stops: " + countFS);
System.out.println("Apostrophes: " + countAP);
System.out.println("Double Quotes: " + countDQ);
System.out.println("Opened parenthesis: " + countOP);
System.out.println("Closed parenthesis: " + countCP);
System.out.println("Unkowns: " + countUK);
}

import java.io.*;

/* This is a programme that reads a tokenised file and * quantifies the occurrence of specific words as * required, and also calculates average word length */

class QuantifyTokens {
    public static void main(String args[]){
        //create a variable to store token required, entered dynamically
        String req = " ";
        // create variable to store file name, entered dynamically
        String fileName = " ";
        // create a variable to hold token from each line
        String tok = " ";
        // create variables to store token counts
        int countRQ = 0; // number of occurrences of token required
        // create variables for word length calculations
        int countNT = 0; // number of tokens
        int countNCH = 0; // number of characters
        int countTCH = 0; // total number of characters
        int countWL = 0; // average word length

        try{
            // get required token from user
            System.out.println("enter token required: ");
            BufferedReader tk = new BufferedReader(new InputStreamReader(System.in));
            req = tk.readLine();

            // get input file name from user
            System.out.println("enter file name: ");
            BufferedReader in = new BufferedReader(new InputStreamReader(System.in));
            fileName = in.readLine();

            // open file using file name entered
            FileInputStream fis = new FileInputStream(fileName);

            // create data input stream and pass in fis as argument
        }
    }
}
DataInputStream dis = new DataInputStream(fis);

// create buffered reader and pass in input stream reader containing dis as argument
BufferedReader br = new BufferedReader(new InputStreamReader(dis));

// read file one line at a time
while ((tok = br.readLine()) != null) {
    // method to count particular tokens
    if (tok.equalsIgnoreCase(req)) {
        countRQ++;
    }
    // method to find word length
    countNT++;
    countNCH = tok.length();
    countTCH += countNCH;
    countWL = (countTCH / countNT);
}

// close stream
dis.close();

// catch exception if any
try {
    System.err.println("Error: "+ e.getMessage());
}

// print totals to console
System.out.println("Results: "+ fileName);
System.out.println("number of tokens:"+ countNT);
System.out.println("number of characters:"+ countTCH);
System.out.println("average word length:"+ countWL);
System.out.println("number of occurrences of ":"+ req + ":"+ countRQ);
}
Appendix V: Linguistic Features in Detail.

The main group of linguistic features are those identified by the tagset in Appendix III, consisting of forty eigth parts-of-speech, when counting six different forms for each of the three verb types. These formed the basis of much of the analysis undertaken in this project. Biber (1988) suggested sixteen linguistic features for the English language and those available for investigation and considered relevant to this study are outlined below.

Verbs:

All verb forms were identified but those considered meaningful according to Biber (1988) are: (i) past tense forms which are indicative of narrative; (ii) perfect aspect which marks actions in past time with current relevance (Quirk et al. 1985:189) and are considered to be typical of a descriptive style; and (iii) present tense (in the base form or third person singular present) which are considered to bring the focus on immediate events and sometimes to remove the focus from temporal events in more academic styles.

Place and time adverbials:

These are thought to focus attention directly on the physical and temporal situation, and in this way they mark involvement and show situated content. They comprise the lists suggested by Quirk et al. (1985: 514 and 526 respectively,) as listed hereunder.

Place: aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west.

Time: afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday.
Pronouns:

While Biber differentiates between first, second and third person pronouns, these are simply grouped together as personal pronouns for the purposes of the present study. It would be possible to identify these individually, but it would be quite time-consuming and since transcripts of spoken texts are not being used, it is felt that there would be little to be gained from such an exercise. This is because first and second person pronouns are a marked feature of conversation and some other spoken genres since they bring an interpersonal focus to play. Third person pronouns however do place the focus outside the immediate domain and are therefore a common feature of reported styles. The impersonal pronoun it has been associated with lack of time for reflection and a non-informational focus. Demonstrative and indefinite pronouns have not been quantified either due to time constraints, but they generally have exophoric reference, to entities outside the text.

Questions:

While some of the 'Wh-' categories have been identified in the corpora, there is uncertainty as to their significance. Generally they relate to interpersonal focus and involvement levels, so they too would be expected to be more significant in spoken genres.

Nominal forms:

A huge distinction is made between verbal and nominal styles, since the use of nominal forms shows a focus on delivering information, as opposed to the use of verbal forms which are used more for an interpersonal focus in narratives and the like. Nominalisations occur when a verb is modified in order to use it as a noun. They are used to integrate information and expand idea units. Words for this category, as suggested by Biber (1988) are those ending in -tion#, -ment#, -ness#, or -ity# (plus plural forms). Figure 2. shows the distribution of nominalisations among the corpora, as found in Hyperbase, since the custom programme was not extended as far as looking for particular word endings.
Prepositions:

These are mainly used to pack information into academic nominal discourse. Following a list suggested by Quirk et al (1985: 665-667), these comprise: against, amid, amidst, among, amongst, at, besides, between, by, despite, during, except, for, from, in, into, minus, notwithstanding, of, off, on, onto, opposite, out, per, plus, pro, re, than, through, throughout, thru, to, toward, towards, upon, versus, via, with, within, without. See Figure 1. for the distribution pattern of prepositions among the corpora.

Adjectives and Adverbs:

These are also thought to “expand and elaborate the information presented in a text.” (Biber 1988:237). They do this through the use of detailed descriptions (by using adjectives) and by qualifying verbs in order to be more explicit (through the use of adverbs).

Lexical Specificity:

Lexical specificity is thought to correlate more with the differences in production circumstances between written and spoken discourse, so these may not be particularly relevant in this study. The type/token ratio basically shows whether the same words are repeated within a corpus, or whether different words are used instead. This is indicative of vocabulary richness which in turn might convey more specific information but also shows that there has been more time for reflection. This would possibly have been useful but due to time limitations it was not possible to calculate it, and this also applies to sentence length. However word length was calculated as being the average length of a word in terms of the number of characters. There is a tendency for longer words to be more specific and defining.

Other features were also considered, such as punctuation marks, but the relevance of these in the context of school-based genres could not be identified as being of significance. Their usage as markers of different genres might be better suited when this involves comparing both spoken and written discourse.
Appendix VI: Calculation of Weighted Values.

<table>
<thead>
<tr>
<th>pos</th>
<th>grand total</th>
<th>eng1</th>
<th>eng2</th>
<th>eng3</th>
<th>total eng</th>
<th>geo1</th>
<th>geo2</th>
<th>total geo</th>
<th>mat1</th>
<th>mat2</th>
<th>total mat</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbs</td>
<td>125881</td>
<td>20552</td>
<td>30159</td>
<td>27195</td>
<td>77906</td>
<td>12734</td>
<td>13554</td>
<td>26288</td>
<td>10455</td>
<td>11232</td>
<td>21687</td>
</tr>
<tr>
<td>adverbs</td>
<td>39031</td>
<td>6745</td>
<td>10636</td>
<td>8797</td>
<td>26178</td>
<td>3794</td>
<td>3916</td>
<td>7710</td>
<td>2285</td>
<td>2858</td>
<td>5143</td>
</tr>
<tr>
<td>nouns</td>
<td>278410</td>
<td>31340</td>
<td>45803</td>
<td>45720</td>
<td>122863</td>
<td>34209</td>
<td>33223</td>
<td>67432</td>
<td>44054</td>
<td>44061</td>
<td>88115</td>
</tr>
<tr>
<td>determiners</td>
<td>86806</td>
<td>11539</td>
<td>18039</td>
<td>15716</td>
<td>45294</td>
<td>9096</td>
<td>9119</td>
<td>18215</td>
<td>11423</td>
<td>11874</td>
<td>23297</td>
</tr>
<tr>
<td>adjectives</td>
<td>65564</td>
<td>8205</td>
<td>11363</td>
<td>10598</td>
<td>30166</td>
<td>8960</td>
<td>9398</td>
<td>18358</td>
<td>8382</td>
<td>8658</td>
<td>17040</td>
</tr>
<tr>
<td>prepositions</td>
<td>102188</td>
<td>13726</td>
<td>21201</td>
<td>19476</td>
<td>54403</td>
<td>12238</td>
<td>12469</td>
<td>24707</td>
<td>11196</td>
<td>11882</td>
<td>23078</td>
</tr>
<tr>
<td>numerals</td>
<td>55050</td>
<td>1830</td>
<td>2901</td>
<td>4389</td>
<td>9120</td>
<td>3890</td>
<td>2421</td>
<td>6311</td>
<td>19498</td>
<td>20121</td>
<td>39619</td>
</tr>
<tr>
<td>coordinators</td>
<td>25500</td>
<td>3496</td>
<td>5721</td>
<td>5431</td>
<td>14648</td>
<td>3129</td>
<td>2931</td>
<td>6060</td>
<td>2436</td>
<td>2356</td>
<td>4792</td>
</tr>
<tr>
<td>pronouns</td>
<td>61874</td>
<td>11382</td>
<td>18650</td>
<td>17784</td>
<td>47816</td>
<td>4010</td>
<td>3638</td>
<td>7648</td>
<td>3005</td>
<td>3405</td>
<td>6410</td>
</tr>
<tr>
<td>interjections</td>
<td>506</td>
<td>67</td>
<td>204</td>
<td>190</td>
<td>461</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>21</td>
<td>18</td>
<td>39</td>
</tr>
<tr>
<td>% / weight</td>
<td></td>
<td>13.35</td>
<td>19.75</td>
<td>18.75</td>
<td>2.07</td>
<td>11.63</td>
<td>11.65</td>
<td>0.93</td>
<td>12.00</td>
<td>12.87</td>
<td>1.00</td>
</tr>
<tr>
<td>weighted values :</td>
<td>verbs</td>
<td>%</td>
<td>adverbs</td>
<td>%</td>
<td>nouns</td>
<td>%</td>
<td>dets</td>
<td>%</td>
<td>adjectives</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>eng</td>
<td>37,568.17</td>
<td>0.43</td>
<td>12,623.67</td>
<td>0.48</td>
<td>59,247.53</td>
<td>0.27</td>
<td>21,841.87</td>
<td>0.34</td>
<td>14,546.78</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>geo</td>
<td>28,203.11</td>
<td>0.32</td>
<td>8,279.60</td>
<td>0.32</td>
<td>72,413.75</td>
<td>0.33</td>
<td>19,560.69</td>
<td>0.30</td>
<td>19,714.25</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>mat</td>
<td>21,794.40</td>
<td>0.25</td>
<td>5,168.47</td>
<td>0.20</td>
<td>88,551.39</td>
<td>0.40</td>
<td>23,412.38</td>
<td>0.36</td>
<td>17,124.39</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>87,592.68</td>
<td>1.00</td>
<td>26,071.74</td>
<td>1.00</td>
<td>220,212.66</td>
<td>1.00</td>
<td>64,814.94</td>
<td>1.00</td>
<td>51,385.42</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>weighted values :</td>
<td>preps</td>
<td>%</td>
<td>numerals</td>
<td>%</td>
<td>coords</td>
<td>%</td>
<td>pronouns</td>
<td>%</td>
<td>interjs</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>eng</td>
<td>26,234.45</td>
<td>0.35</td>
<td>4,397.89</td>
<td>0.09</td>
<td>7,063.62</td>
<td>0.38</td>
<td>23,058.04</td>
<td>0.61</td>
<td>222.31</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>geo</td>
<td>26,532.31</td>
<td>0.35</td>
<td>6,777.24</td>
<td>0.13</td>
<td>6,507.70</td>
<td>0.35</td>
<td>8,213.02</td>
<td>0.22</td>
<td>6.44</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>mat</td>
<td>23,192.29</td>
<td>0.31</td>
<td>39,815.21</td>
<td>0.78</td>
<td>4,815.73</td>
<td>0.26</td>
<td>6,441.75</td>
<td>0.17</td>
<td>39.19</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>75,959.05</td>
<td>1.00</td>
<td>50,990.34</td>
<td>1.00</td>
<td>18,387.06</td>
<td>1.00</td>
<td>37,712.80</td>
<td>1.00</td>
<td>267.94</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>