THE UNIVERSITY
of LIVERPOOL

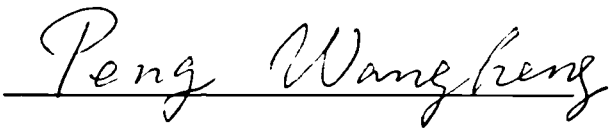
LEXICAL PATTERNING, KEY WORDS, AND THE THEME-RHEME SYSTEM

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by Peng Wangheng

January 1999

Declaration

This work is original and has not been submitted previously in support of any degree, qualification or course.

Signature 

To my parents

Acknowledgements

My heartfelt thanks firstly go to Professor Michael Hoey. I am really very lucky to have the most resourceful linguist and caring teacher as my supervisor. I have always enjoyed his supervision, from which I could be sure to get his encouragement and inspiration to carry on. Without his never-ceasing encouragement and inspiration, this thesis would not have been.

I'd like to thank Dr Mike Scott, who has showed his concern for me throughout my stay in Liverpool. Importantly, he kindly allowed me to use his computer program the "WordSmith Tools" and constantly offered technical support and tips to help me use the program more effectively. In addition, he has always been happy to discuss my problems and could always find a way to help me out.

I own my deepest gratitude to Margaret Berry, who has always been a source of inspiration. She has read some chapters in draft form and made invaluable comments especially on the notion of Theme. Martin Davies has showed his interest in my work and has tirelessly helped me to clarify the notion of "aboutness". Geoff Thompson has kindly helped me with insightful discussions and constructive comments on my previous drafts. Thanks also go to Dr Richard Kirkby and Jean Hill for their help to bring the thesis to a better shape.

The staff and students at the Applied English Language Studies Unit are always pleasant and helpful. I would especially express my thanks to Antoinette Reneuf, Sue Thompson, Nelia Scott, Tony Berber Sardinha, Mike Pacey, Steve Jones, Alfred Ndahiro, Celia Shalom, Maria Stella Martinez, Sarah Waite and Zargham Barganchi, for their interest and support in various forms. I would

also express my thanks to the clerical staff of AELSU, especially Maureen Molly, Karin Alecock, and Gill Lester, who have all been so kind and readily helpful.

Thanks to the Overseas Research Students Award Scheme (Reference number: ORS/95025003) which largely sponsored my tuition fees, and thanks to the English Department of the University of Liverpool which further supported me financially, I was able to pursue the degree of PhD. My funding also came from many individuals, both in China and in Britain. Here I would especially express my thanks to my parents and my wife's parents, Guo Lin, Lin Quanlin, Christopher Hampton, Antoinette Reneuf and Professor Michael Hoey.

Last but not least, I would express my thanks as well as my apology to my wife Yiping and my daughter Ruiting, for their trust, understanding and tolerance.

Abstract

This study explored the organisation of information in the text through the triple interface of lexical patterning, key words and the Theme-Rheme system in text. It was assumed that if a clause Theme is what the clause is about, all the clause Themes in a text may realise the prioritised meanings of the text. Thematic progression manifests the foundation of topic development, and patterns of thematic progression is basically realised by lexical repetition. On the other hand, computer generated key words of a text together were assumed to be able to reflect the 'aboutness' of the text. In an exploration of the distribution of computer generated key words in the Theme-Rheme system, this study found a very close relationship between the Theme-Rheme system and the distribution of key words. Key words were found to occur more frequently in Theme than in Rheme. In a further attempt to discover a clearer indicator of information distribution in the Theme and Rheme areas of the text, this study explored the distribution of key-word links in the Theme and Rheme areas of the sample texts. Key-word link density was shown to be closely related with the Theme-Rheme system and lexical patterning in text. There was a general trend for key-word link density to be higher in the Theme area than in the Rheme area of the texts.

This study has developed a statistical analytical methodology to explore text organisation. Further research may explore the relationship between different ratios of lexical patterning and key word distribution in the Theme and Rheme areas and the perception of the aboutness of the text.

Table of Contents

List of Tables

Note: The first digit in the serial number of the table indicates the chapter where the table is to be found.

List of Figures

Note: The first digit in the serial number of the figure indicates the chapter where the figure is to be found.

Chapter 1. Introduction

1.1 Background of this study

Firbas (1992b: 167 and 184) has repeatedly used Mathesius's metaphor that language is a fortress that must be attacked with all means from all sides. The present study is an attempt to join the attack on the fortress in terms of text organisation as manifested by the interplay of lexical patterning, key word distribution and the Theme-Rheme system.

With the rapid progress in computer technology, the study of linguistics has benefited from the increasing availability of electronically readable texts and easy-to-use packages of computer programs. This has made it possible to study language from new perspectives, and new insights are being gained, just as Clear predicted a decade ago,

> The power of the machine to store, search, classify, sort and otherwise manipulate language data in vast quantities liberates the study of language, much in the same way as the technology of the electron microscope opens new vistas for the researcher into physics and the natural sciences (1987: 61).

This thesis reports on an investigation into the organisation of text that makes extensive use of computer technology. A text may be either spoken or written,

but because the computer is especially capable of processing text in the written mode, and many of the characteristics are shared by both modes of the language, the present study is mainly concerned with written text. Therefore, as a short-hand form, the term 'text' will be used throughout this thesis mainly to mean 'written text'.

1.2 A working definition of text

The overall aim of this thesis is to investigate the organisation of text. Therefore, as a starting point, it is necessary to provide a working definition of the term *text* as it is used in this thesis.

A text is popularly defined as the main body of writing in a book, or any of the various forms in which a book, article, etc. exists (Procter et. al. 1988: 1146). This definition stresses the formal aspect of the text. There are also definitions which stress the semantic aspect of the text. For example, Halliday argued, 'A text is a semantic unit, not a grammatical one' (1985a: xvii). He went on, 'A text can be a highly complex phenomenon, the product of a highly complex ideational and interpersonal environment' (1985a: xvi). People use text to communicate, conveying ideational and interpersonal meanings. On the other hand, text also helps to shape people's ideology. Hoey (1997) even claimed that 'it is not we that generate the text, but the text, drawn from an immense network of previously encountered texts, that generates us' (1997: 264). What

is common between Halliday and Hoey is the fact that they both imply that a text is not a separate phenomenon from language or an epiphenomenon of language. It is an integral part of our language.

De Beaugrande and Dressler (1981) propose that a text should be defined as a communicative occurrence which meets seven standards of textuality: cohesion, coherence (which are text-centred notions), intentionality, acceptability, informativity, situationality, and intertextuality. They claim that 'if any of these standards is not considered to have been satisfied, the text will not be communicative. Hence, non-communicative texts are treated as non-texts' (1981: 3). However, Halliday is very cautious with the idea of 'non-text'. He says, 'We do not ordinarily meet with language that is not textured. ... People go to great lengths to interpret as text anything that is said or written, and are ready to assume any kind of displacement - some error in production, or in their own understanding - rather than admit that they are being faced with "non-text"' (1994: 334).

Halliday proposes that for linguistic analysis a text may be regarded as either process or product (1985a, 1994). But he emphasises that 'it is important to be able to think of text dynamically, as an ongoing process of meaning; and of textual cohesion as an aspect of this process, whereby the flow of meaning is channelled into a traceable current of discourse instead of spilling out formlessly in every possible direction' (Halliday 1994: 311).

For purposes of linguistic analysis, Hoey uses the term 'text' in two ways. 'Firstly, it refers to a piece of continuous language from a single source that is available for linguistic analysis... Secondly, it refers to the linguistic level between grammar and interaction' (1991: 269).

As far as size is concerned, a text is generally regarded as any stretch of language, from one or two words such as 'Stop!' or 'No Smoking', through to voluminous works such as the Encyclopaedia Britannia. Halliday observes, 'There are certain texts which the context of situation determines have to be short, like telegrams and newspaper head-lines' (1985a: 372). But on most occasions, a text can be defined as a unit larger than the sentence (cf. Pike 1967, Koch 1971, Heger 1976), or as a sequence of well-formed sentences (de Beaugrande and Dressler 1981: 28). In his analysis, Fries presumes a text to be larger than one sentence in length (1995: 54). Berber Sardinha's text is longer still, as he defines a segment of text as 'more than two sentences' (1997: 16).

By now I am in a position to provide a working definition of the term 'text' for use in the present thesis. A text is a semantic unit of language whose status is decided by the context of situation and the reader's perception of such context factors. If there is nothing in the context to challenge its being a complete message accomplishing a communicative task, then this piece of language should be regarded as a text, regardless of its length. The majority of texts we encounter everyday are longer than one sentence. Therefore, for the purpose of the present study, the term 'text' is used to mean a piece of writing which

accomplishes a communicative task and which is composed of more than one sentence.

1.3 Area of focus in this study

When a reader is confronted with a piece of written language, he/she will normally assume that it is a text. The text will sometimes be found to be fairly easy to read, because the reader can comfortably follow the flow of thoughts of the writer. Sometimes, however, a text may be felt to be rather difficult to read. The clues are hidden and the messages are opaque or disorganised. This suggests that there must be some properties in the text itself which enable the reader to judge whether the piece of language he/she is reading is a successful text or not.

Even though the reader will normally try not to reject a piece of language as 'non-text', he/she will certainly have an impression of what is a 'successful' text, at least to himself/herself (Berry 1989). It is assumed that one of the likely criteria that a reader has in mind when judging a text to be successful is coherence, or the way the text's messages 'hang together' (Halliday 1985a). A coherent text is normally easier to process than an incoherent one.

Halliday and Hasan (1976) claim that cohesion is an essential property of text, and cohesive devices connect sentences to make a unified text. However, this

claim is not accepted by all linguists. For example, Brown and Yule argue that 'formal cohesion will not guarantee identification as a text nor... will it guarantee textual coherence' (1983: 197). Hoey believes that 'coherence is a quality assigned to text by a reader or listener as a... unity' (1991: 265-266). In short, coherence is a property of the text in the reader's mind. Therefore different readers may judge the same text differently in terms of coherence.

Normally, a text may convey messages that are complex and thus could not be completely expressed by a single sentence. When a reader reads a text, he/she normally needs to process the information and reformulate it in his/her own mind. An explicit output of this processing may be in the form of a 'summary', which is useful for the reader himself/herself to remember the text and also useful for others who do not have the time to read the source text. A summary is a special kind of text which is derived from the source text and is thus closely related to the source text in terms of information content. A successful summary, therefore, should be a coherent text as well as one which retains the central messages of the source text.

Halliday holds that 'linguistic analysis may enable one to say why the text is, or is not, an effective text for its own purposes - in what respects it succeeds and in what respects it fails, or is less successful' (1985a: xv).

Naturally, two questions arise at this juncture.

1. If coherence is a property in the reader's mind, what is the property in the text itself which enables the reader to find it coherent? i.e. what makes a 'coherent text'?
2. If a successful summary is not only a coherent text but also one which retains the central messages of the source text, why is it so, and what features of the source text are retained?

This thesis is not in a position to give a comprehensive answer to the above questions, for a comprehensive answer is far too demanding for a thesis of this size. What this thesis can do, however, is to search for a discovery procedure for recognising patterns of information distribution. It seeks to reveal the relationship between text organisation and information distribution in text, which will be discussed in the succeeding sections.

1.4 Underlying principles of this study

Broadly speaking, the primary aim of this thesis is to investigate text organisation in terms of information distribution in the naturally written text. In determining which area of text organisation to investigate, certain general principles were kept in mind. The following paragraphs represent the underlying principles of this study.

Firstly, the field of investigation had to be general and fundamental. That is, the property of the text under investigation had to be representative of the language. For this reason, the sample texts would ideally need to be 'domain-free'. The results of the analysis ought to interest not only a restricted group of linguists but also a wide range of language researchers and ordinary language users.

Secondly, the features of text organisation under investigation had to be countable and quantitatively comparable. For example, in the analysis, it had to be possible to ask questions such as how occurrences of feature A in the text might be compared with occurrences of feature B in the same text in terms of their relative frequencies.

Thirdly, the categories of the analysis had to be computationally recognisable. That is, they needed to be clearly identifiable by the computer and ideally processed with minimum human intervention

Fourthly, hypotheses had to be testable. It had to be possible to set up hypotheses on the basis of existing knowledge plus intuition, and to test them against evidence obtained from the analysis of textual data drawn from my corpora.

Finally, the methods and the results had to be replicable. The procedures had to be clearly recorded, so that other researchers would be able to reproduce the same results given the same conditions of investigation.

With the above underlying principles in mind, the present study may be described as 'data-driven' (Johns 1991: 1ff, Stubbs 1993: 7ff) or 'corpus-driven' (Leech 1997: 3).

1.5 General research questions

As suggested by the title of this thesis, the present study focuses on the relationships amongst lexical patterning, key words, and the Theme-Rheme system, and as stated at the beginning of this thesis, this study makes extensive use of available computer technology. The notions of lexical patterning and the Theme-Rheme system are not unfamiliar to present day linguists. Nevertheless, the former will be discussed in detail in Chapter 3, and the latter will receive careful treatment in Chapter 2. However, it is the notion of key words that appears to be so commonplace that it might be misleading. Therefore, at this point, it is necessary to make a brief comment on this notion. This notion, as used in this thesis, comes from Scott (1996), who assumes that the comparative frequency of a certain word type in a text may indicate the textual status of this word. If the frequency of a word is unusually high or low in a text, as compared with its frequency in the language in general, this word is regarded as a key

word. If its frequency is unusually high in the text, it is a *positive key word.* If its frequency is unusually low in the text, it is a *negative key word.* In this thesis, however, focus is laid on the analysis of positive key words. It is assumed that all the positive key words in a text may indicate what the text is about, and the distribution of such key words in the text may reflect the patterns of information distribution in the text. Key words are defined in this thesis as lexical items which have a special status in the text organisation, and the degrees of textual significance of a key word may be quantitatively measured by *keyness*. This will be discussed in more detail in Chapter 5.

The decision to choose the aspects of lexical patterning, key words, and the Theme-Rheme system as my foci of investigation was made for three reasons. The reason for choosing lexical patterning as a focus of the thesis is that the approach to patterns of lexis is very suitable for computer aided textual analysis. There are three advantages to this approach. Firstly, it is feasible to program the computer to identify patterns of lexis in text by identifying repetition of lexical items. Indeed, there have been a number of attempts to construct computer programs on the basis of patterns of lexis to extract central messages from the text, as will be discussed in Chapter 3. Secondly, with the increasing availability of computer readable texts, it is possible to investigate text organisation on the basis of large corpora of textual data, which was not possible little more than a decade ago. As Halliday (1993: 24) predicted, 'with the potential for quantitative research opened up by corpus linguistics our

understanding of language, and hence of semiotic systems in general, seems likely to undergo a qualitative change.' Thirdly, patterns of lexis are able to reveal macro patterns of text organisation, and even inter-textual connections. Hoey (1997) observes, 'readers utilise a number of basic interpretative processes in order to identify common content. The most wide-ranging are the lexical processes' (1997: 258).

The reason for choosing key words as a focus of the thesis is that, since key words as defined in this thesis are a special kind of lexical item selected by a computer program, the analysis of key words has all the advantages noted about the analysis of lexical patterning. In addition, because key words are a special kind of lexical item, the distribution of key words in the text may be more sensitive to the property of text organisation than the distribution of average lexical items. Therefore, analysing key words may possibly reveal the property of text organisation more clearly than simply analysing lexical patterning. On the other hand, because key words are selected by the criterion of comparative frequency in the text and in the language in general, the distribution of key words in the text may shed some light on the relationship between text organisation and the use of language in general.

The reason for choosing the Theme-Rheme system as a focus of the thesis is that analysis of the Theme-Rheme system provides a discovery procedure for recognising patterns of information distribution at the clause level and, more important to the present study, at the text level. Daneš (1974, 1987), for

example, proposes patterns of thematic progression (TP) to reveal text organisation. Fries (1983, 1991) finds the choice of clause Themes closely related to the method of development of text. Berry (1989), working from a pedagogical perspective, recognises a close relationship between Thematic options and success in writing.

On the assumption that Thematic choices are characteristically realised by lexical choices, this thesis attempts to investigate the relationship between roles of the Theme-Rheme system and patterns of lexis in text organisation. Since Theme is an important area for identifying patterns of information distribution, and the Theme-Rheme system and patterns of lexis both play an important role in forming texture and carrying messages, the interrelationships between lexical patterning, key words and Theme-Rheme system may be of interest for investigation. It may shed some light on the understanding of the property of text organisation. Moreover, it is hoped that a better understanding of text organisation will contribute to the development of more efficient computer programs that will produce satisfactory summarisation of written texts, though this thesis is not itself concerned with the development of such programs.

With the above three foci in mind, this thesis attempts to answer the following questions:

1. Is there a relationship between lexical patterning and the Theme-Rheme system and, if there is, how is it manifested in the text?

2. Is there a relationship between key words and the Theme-Rheme system and, if there is, how is it manifested in the text?
3. Is there a three-way relationship amongst lexical patterning, key words and the Theme-Rheme system and, if there is, how is it manifested in the text?
4. What are the implications of such relationships for our understanding of text organisation?

1.6 Organisation of this thesis

Halliday (1985a: 313; 1994: 334) identified two major categories of features as those which combine to make up the 'textual' component of the grammar of English. One category is *structural,* which consists of the thematic structure and the information structure. The other category is *cohesive,* which is composed of reference, ellipsis and substitution, conjunction and lexical cohesion. His classification of the features which define a text may serve as a guideline of the organisation of the present thesis. More specifically, the thesis will focus on the Theme-Rheme system under the structural category, and lexis under the cohesive category.

The thesis will be divided into three parts. The first part runs from Chapter 1 to Chapter 3, and comprises a review of literature on the basis of Halliday's classification. Chapter 1, the present chapter, has provided a brief background

to the thesis, setting up principles and raising general research questions for the research work in this study. Chapter 2 will review work on structural/textual theories on text organisation, in particular work concerning the Theme-Rheme system. It will review the literature concerning approaches to the definition and identification of the Theme-Rheme system along two complementary lines. One is to review the work of linguists who treat Theme as a function of the clause and the work of linguists who treat Theme as a function of the text. The other is to review the work of linguists who combine the Theme-Rheme system and the information system and the work of linguists who separate the two systems. In the review of literature of the Theme-Rheme system, particular attention will be paid to the role Theme plays in the organisation of text, and approaches to the identification of Theme in actual analysis. Chapter 3 will review work on the role that lexis plays in text organisation. Firstly, it will review work which is concerned with the function of lexis in textual cohesion and in the macro-structures of text. Then it will review work which relates lexical patterning to the discourse topics in text and attempts to discover and formulate this relationship. Finally it will review attempts to create automatic summarisation by exploiting patterns of lexis in text.

The second part, which runs from Chapter 4 to Chapter 6, reports on experimental analyses of a corpus of thirty-two sample texts. Chapter 4 marks the beginning of the analysis in this study. It reports on a preliminary analysis of a sample text. Based on the literature as reviewed in the previous two

chapters, it assumes that a relationship exists between lexical patterning and the Theme-Rheme system which has impact on the text organisation. Therefore, it focuses on the analysis of distribution of lexical links in Theme and Rheme, and the relationship between such linking patterns and the perceived degrees of coherence in the text.

Chapter 5 carries on the analysis on expanded data with additional parameters of key words and keyness and their distribution in Theme and Rheme. As will be explained in detail in Chapter 5, the notion of key words is closely related to the notion of text and keyness is a measurement which indicates the degree of textual significance of the key word. Based on the results of analysis obtained in Chapter 4, it assumes that key words may concentrate in the Theme area rather than the Rheme area, and consequently the overall keyness may be higher in Theme rather than in Rheme of the text.

Chapter 6 continues on the basis of findings made in Chapters 4 and 5, and pulls the threads together. It assumes that if there are more lexical links in Theme than in Rheme, and if there are more key words in Theme than in Rheme, then it may also be more links formed by key words in Theme than in Rheme. Therefore it sets out to investigate key-word links in Theme and Rheme, in the hope that a discovery procedure may be worked out for recognising patterns of information distribution in the text.

In the third part of the thesis, conclusions will be provided in Chapter 7, which will summarise the study and discuss implications of the results in regard to the theory of text organisation and practice of text summarisation. Finally, after briefly discussing constraints and limitations, Chapter 7 will suggest directions for further research.

Chapter 2. The Theme-Rheme System and Text Organisation

2.1 Introduction to this chapter

In Chapter 1, a working definition of the term text as it is used in this thesis was provided. This chapter will move on to reviewing some of the properties of text. As noted in Chapter 1, Halliday categorises the features which make up the 'textual' component of English as structural (Theme-Rheme and Given-New) and cohesive (reference, ellipsis and substitution, conjunction, and lexical cohesion) (1985a: 313; 1994: 334). Following his categorisation, this chapter will review the structural features of text, with a focus on the Theme-Rheme system, and Chapter 3 will review the cohesive features, with a focus on lexical cohesion.

However, before the review of the structural and cohesive features of text is carried out, the term 'structural' needs to be reconsidered more carefully. Text is viewed as structured by Thorndyke (1977), van Dijk and Kintsch (1983) and van Dijk (1980, 1983), who emphasise the central role of text structure in the

understanding of written text. But, Halliday does not accept this view. He argues, 'Below the sentence, the typical relationship is a constructional one, of parts into wholes... One manifestation of this structural relationship is the sequence in which the elements occur; but this is only one variable among others... Above the sentence, the position is reversed. Here the non-constructional forms of organisation take over and become the norm' (1985a: xxi). Hoey represents this position more clearly. He proposes that text is best viewed as organised instead of structured; structural descriptions of the text have to be reinterpreted as descriptions of culturally popular patterns of organisation. Hoey explains, 'The difference between a structural and organisational view of text will not necessarily show in the detail of presentation, but it is of some consequence for the way we view language'(1991: 194). 'Structural statements claim to say what is possible; organisational statements claim to describe what is done'(1991: 193).

In fact, despite the apparent similarity in the use of the word 'structural', the above linguists are talking about very different things. Halliday's notion of 'structural' is basically that of structural features of the clause or clause complex that contribute to texture, while Meyer, Thorndyke, van Dijk, Kinsch, and Hoey see the text itself as having organisational properties, though they differ as to the appropriateness of the label 'structural'.

In this thesis, Hoey's stand will be adopted; that is, text will be viewed as organised rather than structured. However, because Halliday's categorisation

clearly marks the different roles played by grammatical and semantic devices in text organisation, and since the difference between a structural and organisational view of text will not necessarily influence the detail of presentation, the term 'structural' as used by Halliday will be borrowed in this chapter to describe one aspect of text organisation, namely that of the Theme-Rheme system.

2.2 The notion of Theme

It is commonly acknowledged that if a text is to be appreciated as a text, there should be some kind of 'continuity' to what it is about. As Beaugrande and Dressler claim,

> ... the vital importance of continuity has all too frequently been overlooked in linguists' preoccupation with analysis into units and constituents. All the standards of textuality are closely related to continuity (1981: 46).

One of the factors contributing to the realisation of text continuity is the clause Theme, which is part of the binary system of Theme and Rheme.

However, in spite of the attraction it has for linguists, the notion of Theme is still far from being agreed. The root of the disagreement is typically indicated by the following two questions: What exactly is Theme? And where is it best to place the borderline between Theme and Rheme? After many years of debates over these questions, Theme remains an area of dispute concerning both its

definition and its identification, or in other words, its functions and scope (see especially Hasan and Fries, 1995, and Berry, 1996, for a summary of the different arguments).

According to Halliday (1985a: 38), Theme is what the clause is about. On the other hand, many linguists also hold that Theme is indicative of what the text is about. The patterns of thematic development between the clauses are regarded as being able to reveal some aspect of the continuity of the text (e.g. Daneš 1974, Fries 1983, Berry 1996). In this thesis it is assumed that in order to reveal text organisation it is helpful to study the patterns of thematic development in the text.

The remainder of this chapter will review some major approaches to the interpretation of the Theme-Rheme system.

Initially, perhaps influenced by the general trend of interest in clause or sentence structure, the notion of Theme attracted attention from linguists as a clause phenomenon. But with the shift of interest from separate clauses to continuous text, the notion of Theme became more and more attractive as a text phenomenon, because it provides evidence for the organisation of text. Therefore, to trace the shift and draw insight from previous studies on the notion of Theme, this review of the notion of Theme will move from Theme as a function of the clause to Theme as a function of the text.

In a summary of work exploring the notion of Theme, Fries (1981: 1-2) distinguishes two major approaches: the *combining approach*, which uses Mathesius's two criteria of 'that which is known or at least obvious in the given situation, and from which the speaker proceeds' to define Theme (see Section 2.3.1 below), and the *separating approach*, which takes only the second criterion as definitive of Theme. Fries argues that the first criterion, 'that which is known or at least obvious in the given situation', should belong to a different category, that of *Given-New*, since Theme may also contain new information. He concludes,

> The difference... between the combining and the separating approach to the definition of Theme is that while the combiners either ignore the contribution of word order... or treat it as contributing to the same concept as the given-new distinction... separators tease out and separate the contributions of word-order and of the distinction between given and new information, and they use the term Theme to indicate the meaning of initial position in the clause. (1981: 3)

In this chapter, Fries's distinction will be used as a way of organising a more detailed discussion of the two major approaches in the framework of Theme as a function of the clause and as a function of the text.

2.3 Theme as a function of the clause

2.3.1 'The combiners'

The notion of *Theme* is not a new concept. In a review of the initial development of research into Theme, de Beaugrande and Dressler (1981) recall,

> Comparing word order in ancient and modern languages, Henri Weil (1844, 1887) detected another principle besides grammar: the relationship of 'thoughts' to each other evidently affects the arrangement of words in sentences. His investigations were renewed by Czech linguists (many of them belonging to the 'Prague School') under the designation of Functional Sentence Perspective (1981: 20).

Among the Prague School linguists, Mathesius was the first to propose the use of Theme for the description of text organisation; he defined Theme as 'that which is known or at least obvious in the given situation, and from which the speaker proceeds' (1939: 234, quoted in Firbas 1964: 268). As for the part of the utterance that is not Theme, Mathesius termed it Rheme (from an ancient Greek word 'rhēma', meaning 'a saying'), defining it as the element 'which the speaker states about, or in regard to, the Theme of utterance' (quoted in Firbas 1964: 277).

Firbas, one of the leading Prague School linguists, proposed the notion of *functional sentence perspective (FSP).* Theme was regarded as the part of a sentence constituted by the foundation-laying elements and functioning under

the broader category of FSP. This notion soon became the focus of the Prague School linguists. Skalička (1960) claimed that the whole linguistic theory of grammar of texts should be reduced to FSP. Trost (1962) Sgall (1969), Hausenblas (1964), and Isenberg (1970) studied the role FSP contributed to the inner connexity of texts. Daneš (1970, 1974) studied contextual and Thematic aspects of ESP. Beneš (1968) distinguished the 'point of departure' and what he called the 'foundation' of the utterance. So, in order to understand what Firbas meant by Theme, it is necessary to briefly review what is meant by FSP.

2.3.1.1 Theme and FSP: Firbas (1957, 1992a, 1992b)

Firbas (1957) first used the term FSP to describe the organisation of text through degrees of *communicative dynamism (CD)* and information flow in text. By degrees of CD he meant the relative extent to which the element contributes to the development of the communication within the sentence. But he warned that the notion of 'development' was not to be understood as linear positioning. For example, in the sentence 'Last night, I was reading a fascinating book while I was waiting for you', 'I' carries the lowest degree of CD, whereas 'a fascinating book' carries the highest degree of CD. Thus this sentence is orientated, or perspectived, to 'a fascinating book', which is the element carrying the highest degree of CD in the sentence.

The degrees of CD are the result of the interplay of three factors which are essential for the FSP of written text. The three factors are the contextual factor,

linear modification and the semantic factor, which will be reviewed in more detail below.

According to Firbas, the first factor is the *contextual factor.* In written text, if a piece of information is retrievable from the immediately relevant context, it will be regarded as *context-dependent* or *old* information. On the other hand, if a piece of information is irretrievable from the immediately relevant context, it will be regarded as *context-independent* or *New* information, regardless of its position in the sentence.

Firbas admits that 'context is a complex phenomenon' (1992b: 171). This is largely because of the difficulty in determining the scope of the 'immediately relevant context'. As a way of limiting 'immediately relevant context', Firbas (1992b: 170) proposes a span not exceeding three sentences (simple or complex), though Svoboda (1981) proposes a span of seven clauses.

This criterion seems fairly straightforward at first glance, but in actual analysis it is blurred by Firbas's own modifications. Firbas notes that the immediately relevant context is embedded within the entire preceding context, which in its turn is embedded within a still wider context of common knowledge and experience shared by the producer and receiver of the message, and eventually all the above context complex is embedded within the context of human knowledge and experience. While he proposes focusing on 'immediately relevant preceding context', he admits that sometimes the wider context might

become predominant in determining context dependency. For example, in the sentence 'Give me "The Three Musketeers"' used in reply to 'Which of the two books would you like to have, "The Three Musketeers" or "An Introduction to Meteorology"?' and the sentence 'Well, they've elected me', in reply to 'Whom did they elect in the end?' both '"The Three Musketeers"' and 'me' contain context-dependent information, but at the same time they also convey information that is context-independent. The reason for this , as Firbas explains, is that 'they express the result of a choice, an irretrievable piece of information' (1992b: 171).

Following his approach, simply referring to the immediate context will not enable one to determine convincingly whether a piece of information is retrievable or not, or context-dependent or not, because information outside the immediate context might be more influential on the determination of context-dependency, or retrievability, thus leaving room for more than one interpretation of the function of an element in the sentence. It is easy to envisage a situation in which one and the same element of a sentence may be regarded by different analysts, or even by the same analyst at different times, as containing either context-dependent or context-independent information. Therefore, impressive though his argument may be, the complicated machinery of determining context dependency causes difficulties in actual operation.

Firbas's second factor is *linear modification*. Under this category, the linear position gradually raises the degrees of CD from the beginning towards the end

of the sentence. Linear modification undoubtedly involves sentence linearity, but it is again a rather unreliable criterion in the FSP approach, as noted by Firbas: 'it cannot be claimed that the actual linear arrangement of sentence elements is always in perfect agreement with a gradual rise in CD' (1992a: 8).

As will be shown later in Section 2.3.2.1, this factor is regarded as essential in Halliday's functional grammar concerning information distribution. However, it is not regarded by Firbas as equally powerful with the other two factors: it is indeed the weakest of the three FSP factors in Firbas's framework, easily modified by the other two factors. This point will be returned to later with the discussion of the definition of Theme.

The third factor proposed by Firbas is the *semantic factor*, which involves not only the semantic content of the linguistic element, but also the semantic relations between the elements. The semantic factor performs two communicative roles, or dynamic semantic functions: the *Presentation* function and the *Quality* function. The two semantic functions consist of two scales respectively. The first is the Presentation Scale, which contains *Setting (Set) - Presentation (Pr) - Phenomenon (Ph).* The second scale is the Quality Scale, which contains *Setting (Set) - Bearer (B) - Quality (Q) - Specification (Sp) - Further Specification (FSp).* The above functions are determined on the basis of context dependency. For example, in the sentence 'Then Peter came into the room', if only 'into the room' is context-dependent, the verb performs the Pr-function, the Subject performs the Ph-function, and the adverbial performs the

Set-function. However, in the same sentence, if only 'Peter' is context-dependent, the verb performs the Q-function, the Subject performs the B-function, and the adverbial performs the Sp-function. The two scales can be fused into a Combined Scale (Set - Pr - Ph - B - Q - Sp - FSp). In principle, the scales reflect a gradual rise in degrees of CD as they are carried by context-independent elements. But they do not reflect the actual linear arrangement. Firbas claims that 'the Presentation Scale and the Quality Scale can be regarded as established and looked upon as belonging to the centre of the system of language. A central feature of primary importance indeed are the two communicative perspectives: the Ph-perspective and the Q/Sp-perspective' (1992a: 69).

As with the contextual factor and linear modification, the semantic factor also interplays with the other factors in determining the degrees of CD over the written sentence. The semantic content is also likely to be influenced by context dependency. For example, in the sentence 'Peter flew to Edinburgh yesterday', if only 'Peter' is context dependent, then the adverbial 'to Edinburgh' performs the dynamic semantic function of Specification. It completes the development of the communication and therefore carries the highest degree of CD. On the other hand, 'yesterday', though also context independent, expresses mere background information and therefore performs the dynamic semantic function of Setting. But if 'Peter', 'flew' and 'to Edinburgh' are all context-dependent, 'yesterday' no longer acts as Setting, but as a Specification, and carries the highest CD in the sentence. Even the notional

component of the verb can be rendered context-dependent. For example, in the sentence 'Peter did fly to Edinburgh yesterday', if the notional component 'fly' and the non-verbal elements both convey context-dependent information, the positive polarity conveyed by 'did' becomes the piece of information towards which the communication is perspectived.

Firbas claims that context-dependency overpowers the semantic factor:

> If rendered context-dependent by the contextual factor, an element has the communicative value of its semantic content weakened to such an extent that irrespective of sentence position it does not exceed in CD any context-independent element. (1992b: 173)

On the other hand, if an element is context-independent, in accordance with the character of its semantic content and the character of the semantic relations entered into, it is either capable of working counter to linear modification or incapable of doing so.

> If capable of doing so, the communicative value of its semantic content is unaffected by linear modification. If it is incapable of doing so, linear modification will affect its communicative value. (1992b: 173)

However, Firbas does not state clearly under what conditions an element is regarded as being capable of working counter to linear modification.

In conclusion, Firbas claims,

> An interplay of the three factors determines the distribution of degrees of CD over the written sentence. It determines the perspective in which a semantic and grammatical sentence structure is to function in the act of communication; that is, it determines its functional sentence perspective (1992a: 11).

However, as Firbas admits, the interplay of FSP factors is rather complicated; thus 'it cannot... be expected that the outcome of the interplay of factors will always be invariably unequivocal' (1992a: 11). For instance, there may be different interpretations of the same language phenomenon, depending on whether an element is regarded as context-independent or context-dependent. Since the same element in a sentence has the potential of being either context-independent or context-dependent, it is necessary to rely on the individual reader's judgement, thus giving rise to the potentiality of more than one interpretation of the function of an element in the sentence. Although Firbas hopes that the number of types of potentiality can be reduced by further research, at present, he admits, FSP is still an inadequate explanation of how text is organised.

In spite of its limitations, FSP undoubtedly sheds light on the study of language. Firstly, language is regarded as functional in human communication, which is meaningful and purposeful. Secondly, language is viewed not as a static product, but as a dynamic process. Further, evidence is provided that different elements in the sentence contribute to communication to a varied extent, which can be explained by the degrees of CD in the sentence.

Having reviewed the notion of FSP, we now turn to the notion of Theme within that perspective. Firbas (1992b: 175) defines Theme as being constituted by three kinds of elements: all context-dependent elements, all elements (either context-dependent or context-independent) fulfilling the Setting-function and

all elements (either context-dependent or context-independent) fulfilling the Bearer-function. The rest of the elements (all context-independent) constitute the non-Theme.

Firbas further divides Theme into four sub-categories: *Theme proper (ThP), Theme oriented element (ThPO), diatheme oriented element (DthO)* and *diatheme (Dth)*, arranged in accordance with a gradual rise in degrees of CD. Theme proper is an element that performs a thematic function which is always linked with the lowest degree of CD within the Theme, and in consequence also within the entire sentence. At the other end of the scale is Diatheme, which is a thematic element that carries irretrievable information. Diatheme always carries the highest degree of CD within the Theme. Between Theme proper and Diatheme stand Theme oriented elements and Diatheme oriented elements, which have moderate degrees of CD close to ThP or Dth, respectively, as shown in Figure 2.1 below.

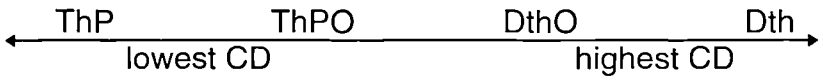


Figure 2.1. Four sub-categories of Theme proposed by Firbas

Note that in Figure 2.1 the sub-categories of Theme are not defined by their linear position in the sentence, but by the degree of CD they carry; that is, it is quite possible for the Diatheme to occur before Theme proper in a sentence. Firbas's subcategorisation of Theme is insightful in that the key to understanding of the nature of Theme lies in its function in language

communication, and in that the subcategorisation of Theme is sensitive to the interplay of the three FSP factors.

In conclusion, FSP was the first attempt to observe sentences from a functional perspective. The sentence is viewed as dynamic rather than static. Theme is defined by the degrees of CD, from the interplay of the three FSP factors, which show that language analysis can be carried out from a broader view than within the language itself, and also show an attempt to reflect the complexity of communication. However, the complicated mechanics needed to decide context-dependency in order to delimit Theme appear to be over-delicate. It is thus hardly replicable with objective criteria and is not easy to apply in computational corpus analysis.

2.3.2 'The separators'

2.3.2.1 The systemic-functional definition: Halliday (1985a)

Following the terminology of the Prague school of linguists, Halliday integrated the notion of Theme into the systemic-functional model, using the term Theme as the label for the dual function of 'point of departure' and 'what the clause is concerned with' in the clause as message (1985a: 38).

The Theme is defined by Halliday as follows:

- 'The Theme is what is being talked about, the point of departure for the clause as a message...' (1967b: 212)
- 'The Theme is the element which serves as the point of departure of the message; it is that with which the clause is concerned.'(1985a: 38)
- '...the Theme is the starting-point for the message; it is what the clause is going to be about.'(1985a: 39)

Although Halliday rightly separates Theme-Rheme from Given-New, he fails to further separate two distinct aspects, namely the 'point of departure'(from the logical perspective) and 'what the clause is about'(from the semantic perspective).

Unlike the 'Prague School' linguists who regard linearity as a rather weak factor in determining Theme, Halliday focuses the definition of Theme largely on the linear position in the sentence. Although he has tried repeatedly to clarify that Theme is not defined by its position, his practice nevertheless shows that Theme can be invariably identified as that element which comes in first position in the English clause (cf. 1985a: 39).

As for the delimitation of Theme, Halliday sets the borderline between Theme and Rheme on the first 'ideational element': 'The Theme of any clause, therefore, extends up to (and includes) the topical Theme. The topical Theme is the first element in the clause that has some function in the ideational structure (i.e. transitivity...)' (1985a: 56).

This way of delimiting Theme has its advantages. At least it seems to be simple to operate, preventing the confusion caused by the rather complicated operation

in determining context dependence and degrees of CD. However, the apparent simplicity has its disadvantages. For example, to use Halliday's own words, it has caused 'one of the most fundamental confusions in linguistics.' Concerning this confusion Halliday had to explain that when he wrote 'The Theme of an English clause is the element that is put in first position', what he actually meant was (although ungrammatically) 'The Theme of an English clause is been [sic] by the element that is put in first position' (1988: 33-34). In other words, Theme does not 'represent' the first position, but is 'represented by' the first position.

However, the confusion is not so easily smoothed out. It involves the correlation of 'aboutness' and the scope of the thematical element. For example, there are complaints that following Halliday's delimitation of Theme, it is often found that 'not every initial sentence element need have anything to do with the discourse topic' (Wikberg 1990: 238). Huddleston has commented, 'I can't make any sense of the idea that *Nothing will satisfy you, You could buy a bar of chocolate like this for 6d before the War,... There's a fallacy in your argument*, are receptively about "nothing", "you" and "there"'(1988: 158). Berry has pointed out that under the assumptions of the 'first ideational element' approach, the grammatical Subject is often outside the Theme, in spite of the fact that Subject very clearly relates to the concerns of the writer and has been recognised as such by the reader (1996: 37).

Halliday himself has not been able to avoid the same problem in his analysis of the clause complex 'For all his integrity and high principles, Robert pulled a slightly fast one over his father and business partners' where he had to resort to the rather unsatisfactory notion of 'displaced Theme' to account for what the clause is about, i.e. the topical element 'Robert', which is not the first ideational element in the clause (1985a: 64).

2.3.2.2 A dynamic perspective: Ravelli (1991, 1995)

Based on Halliday's definition of Theme as 'the point of departure', but attempting to provide a more satisfactory delimitation from a dynamic perspective so as to account for 'what the clause is about', Ravelli argues that the main verb should be the definite limit for the Theme; that is, Theme stops at the main verb of the clause.

From a dynamic perspective, language is viewed as process rather than product. Moreover, it is a process of making choices. At every moment in the production of a text, for example, there are numerous choices to be made. Therefore, Ravelli proposes a *path* analysis, which provides a mechanism for keeping track of choices as they unfold in a text. Such a path analysis is more insightful when carried out simultaneously for the Theme, Mood and Transitivity systems in the systemic functional linguistics model.

In Thematic analysis, any initial element of the clause will be taken to open the *Theme path.* Once a candidate for a topical element is reached, steps into

further elements will be taken to close the Theme path and open the *Rheme path* (1995: 222).

In Mood analysis, if a Finite or wh-element is encountered first, then the Mood will most probably be interrogative; if a nominal element is encountered first, then it is highly likely that a declarative is being formed. For declarative clauses, it is necessary to identify elements which are potential Subjects, and then to wait for a Finite element to confirm this. Once a Finite element is reached, the Mood analysis ceases to be of interest, and further steps must pertain to the Residue(1995: 223).

For Transitivity, the presence of the Process element marks a critical dividing line in the ideational path between 'little' and 'significant' information: the sense of what the clause is 'about' in ideational terms grows as the clause grows, with the Process providing the essential pivot (1995: 226).

In deciding the extent of Theme, then, it is necessary to take account of Mood and Transitivity systems, especially the Mood system. The expectations for Mood and Theme are inter-related. Because of the unmarked association between Subject and Theme, the simultaneous expectations in the Mood analysis affect the Theme analysis. If an element has the potential to function as Subject, the same element will be interpreted as being Thematic. Once the Process is reached, then the Subject is confirmed, and Theme is at the same time completed.

Ravelli explains her proposal of the integrated path analysis from the perspective of the three meta-functions of language as advocated by systemic linguists in this way:

> Ideationally, there is a sense that the departure point of the clause is not fully elaborated until the process is reached, and it is the interpersonal structure which gives rise to the expectation that the message is off the ground and ready to be elaborated. Textually, everything up to that critical dividing line can be seen to be thematic; once there is an element which is not only thematic but also likely to be functioning as Subject, the ideational information is expected to be increased imminently, and the departure point of the message is therefore fully elaborated. (1995: 368)

In her dynamic analysis of the Theme-Rheme structure, Ravelli claims that each of the ideational elements continues to contribute to the 'departure point' of the message, but once the process is reached, the clause is unequivocally 'under way'(1991: 364).

Ravelli's proposal to analyse the text from a dynamic perspective broadens our understanding of text organisation. Her description of how Theme is delimited is particularly relevant to the present thesis. While Ravelli provides a theoretical motive for taking Theme up to the Process, Davies provides criteria for doing so practically. The next section will review Davies's proposal to include Grammatical Subject as the obligatory element of Theme.

2.3.2.3 Extension of the scope of Theme: Davies (1988)

Attempting to overcome the difficulties Halliday has encountered in the delimitation of Theme, Davies (1988) reconsiders how Theme is identified and what function it serves both within the sentence and within the discourse. She assigns to Theme two potential functions: identification of Topic and provision of Contextual Frame. Davies proposes that Grammatical Subject is an obligatory element in Theme, which serves an 'equally obligatory semantic function' of identifying 'topic' in the clause. 'Thus Subject is equated with the intuitive notion of "what the clause is about"'(1988: 177).

On the other hand, the optional element, which includes Circumstantial Adjunct, and/or modal or conjunctive adjuncts, and/or conjunctions, preceding the Grammatical Subject in a clause, serves the distinct function of providing 'different frameworks or contexts for the development of topic' as the discourse proceeds. Unlike the topical elements which are the recurrent elements in coherent text, the 'framing' elements are typically non-recurrent and signal changes or stages in the progression of the discourse (1988: 178).

An outcome of including Subject as an obligatory element in Theme is that it allows Theme to extend up to and including Subject and, hence, when Subject is preceded by an Adjunct or Complement, extends the domain of Theme beyond Halliday's delineation (1988: 178).

Another point where Davies deviates from Halliday is in her treatment of existential 'there' and anticipatory 'it', which Halliday terms 'empty Theme' and 'predicated Theme' respectively. Davies regarded them both as empty Subjects. So, for clauses beginning with existential 'there' or anticipatory 'it', she extended the domain of Theme to include the 'first ideational element' after the verb. For example, in '*It is clear* that...' and '*There is no reason* to suppose that...', the italicised parts are both regarded as thematic elements. Although this inconsistency in Theme delimitation remains to be justified, her method is nevertheless helpful in bringing the 'topic' or 'what the clause is about' under focus in such clauses.

However, Davies's method fails to account for inverted sentence structures in English. For example, she is unable to explain how Theme might be identified in such a sentence as 'Between where I stood by the rail and the lobby was but a few yards', taken from my data, where the Grammatical Subject as the obligatory element of Theme occurs as the last element in the sentence. In spite of this limitation, Davies's proposal is very useful in revealing the semantic function of the clause Theme. Moreover, as Davies claims, the semantic function of Theme is not restricted to the clause. As the topic element is typically recurrent, the repeated occurrence or re-occurrence of the same topical element as Subject also specifies the topic in the discourse, indicating what the text or a particular stretch of text is about.

The notion of Theme is important in describing language at the clause level. But as Fries comments, 'Halliday generally discussed isolated sentences as examples... his case could be made stronger by considering the thematic contribution to texts considered as wholes' (1995: 318-319). In fact, the notion of Theme will be more revealing of the function of language if it is approached from the text level. These approaches will be reviewed in the next section.

2.4 Theme as a function of the text

2.4.1 'The combiners'

2.4.1.1 Thematic Progression: Daneš (1974)

Unlike Firbas, who focused on the role Theme plays in FSP at the sentence level, Daneš (1974) focused on the function of clause Themes in organising a text. In the text, Thematic content may pass from one clause to the next, continuously, thus bringing about continuity of the content of the text. Daneš held that the choice of the clause Themes is not unmotivated. It is connected with the relative position of the clause in the text. He argued that 'the progression of the presentation of subject-matter must necessarily be governed by some regularities, must be patterned' (1974: 109). In support of his argument, he proposed three main types of *Thematic Progression (TP).*

The first type of TP is *simple linear Thematic Progression*, or TP with linear thematisation of rhemes, in which the Theme of a sentence develops from the Rheme of a preceding sentence, as shown in Figure 2.2 below.

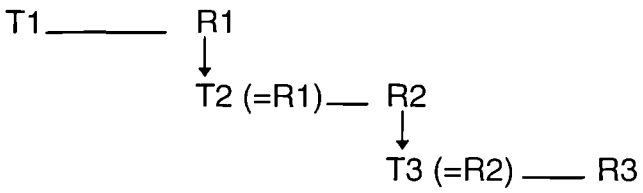


Figure 2.2. Simple linear Thematic Progression

An example of simple linear Thematic Progression from the corpus of sample texts in the present study is as follows.

Example 2.1.

> *Many synthetic drugs have* molecules that come in left- and right-handed forms; *sometimes* one *is beneficial while* its mirror image *is harmful.*

The second type is *Thematic Progression with a continuous Theme*, or TP with a constant Theme, in which the Theme of a sentence develops from the Theme of a preceding sentence, as shown in Figure 2.3 below.

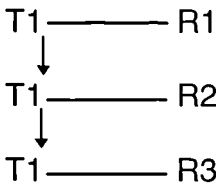


Figure 2.3. Thematic Progression with a continuous Theme

An example of Thematic Progression with a continuous Theme from the corpus of sample texts in the present study is shown in Example 2.2 below.

Example 2.2.

> *Chimps* hunt co-operatively, taking up positions of scout and beater like human tribal huntsmen. *They* signal food availability to one another.

Daneš's third type of TP is *Thematic Progression with derived Themes*, in which the Theme of a sentence develops into Themes of several subsequent sentences. In the derived Theme type of TP, the first Theme is also called the 'hypertheme', which may or may not actually appear in the text. This is shown in Figure 2.4 below.

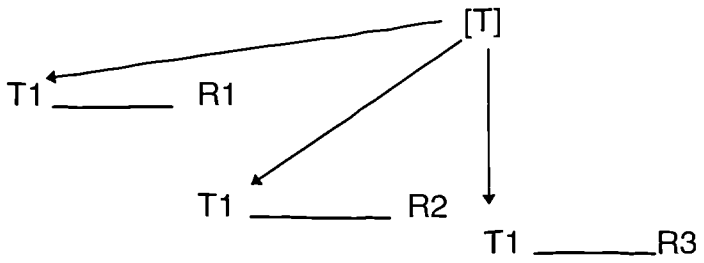


Figure 2.4. Thematic Progression with derived Themes

An example of Thematic Progression with derived Themes, taken from the corpus of sample texts in the present study, is the following:

Example 2.3.

> *The Royal Botanic Gardens at Kew* has built a time machine, designed to transport visitors back 4,000 million years.... *The entrance* is dominated by a pool of bubbling primeval mud, representing the environment in which primitive life first evolved, some 4,000 million

> years ago.... *Under the rock bridge*, you come to the Devonian period, when the first seed plants, the seed ferns, make their initial appearance... *Across the swamp* lurks an extinct woodlouse the size of a corgi, which once dined within the decaying stems of the tree-like lycopsids.

According to Daneš, Thematic Progression can be viewed as the skeleton of the plot. The first type apparently, in Daneš's view, presents the most elementary and basic TP (1974: 118). However, there seems to be a general agreement that the first two of Daneš's TP types are equally basic, but the third type seems to be less than satisfactory, as will be shown later in Section 2.4.1.2. The main problem is in the criteria for the identification of 'hypertheme'. The hypertheme may or may not actually occur in the co-text or immediate context of situation, and even if it does occur, there is no direct link between the hypertheme and the derived themes. The determination of the hypertheme requires extra-linguistic knowledge and thus relies heavily on the intuition of the reader. This is especially a problem for computer processing of textual data.

Nevertheless, Daneš's framework of TP opens a promising area in the exploration of text organisation. Daneš draws attention to the mutual influence between individual clause and the text as a whole. In response to Halliday's statement that 'thematisation is independent of what has gone before' (1967a: 17), Daneš claims that 'the choice of the themes of particular utterances can hardly be fortuitous, unmotivated, and without any structural connexion to the text' (1974: 109). The clause and the text are mutually dependent. The clause realises the text, but is at the same time constrained by the linguistic environment of the text.

Daneš's insight has significant impact on later studies of Theme both at the clause level and the text level. His three main types of TP will serve as one of the starting points of the present study. However, because of the limitation of his framework, especially the third type of TP, attempts have been made to refine his framework. One such attempt was made by Dubois (1987), which will be reviewed in the next sub-section.

2.4.1.2 A reformulation of TP typology: Dubois (1987)

Following Daneš in shifting the focus of studies of Theme from individual clauses or sentences to the text as a whole, Dubois (1987) argues that it is more helpful to study Thematic progression (TP) across authentic discourse than to study Theme in isolated sentences which are often made up by the analysts themselves. Although she takes Halliday and Daneš's works as her starting point, she nevertheless criticises both of them for failing to systematise the bases for their conclusions (1987: 90).

Through analysis of a corpus of authentic discourse, of a comparatively large size at the time, Dubois was able to make interesting findings. Firstly, she discovered that isolated instances of Daneš's three basic types of Thematic Progression, namely constant TP, simple linear TP, and hypertheme, constitute only a minority in her data. Instead, they were often found to occur in combination; or in her words, multiple development is 'rather the rule than the exception' (1987: 95). Secondly, she found in her data that the marked Theme,

which is a clause initial element that is not the grammatical Subject, is usually followed by an unmarked Theme, which is grammatical Subject in the declarative clause. The unmarked Themes following marked Themes develop by the same kinds of progression as those found without marked Themes. Therefore, it may be inferred that for Dubois the scope of Theme should be wider than the Hallidayan convention. Moreover, Dubois found that Thematic Progression may happen between sentences at a distance. She called this kind of Thematic Progression 'gapped Thematic Progression', as opposed to Daneš's examples of Thematic Progression between adjacent sentences. In gapped Thematic Progression, Theme does not develop from the immediately preceding sentence, but the development may be interfered by other sentences.

In her attempt to systematise and simplify Daneš's (1974) framework, Dubois (1987: 109) proposes that recoverable or given Thematic Progression takes only two forms:

1. *themic*: from a previous Theme or Themes;
2. *rhemic*: from a previous Rheme or Rhemes.

Each of the two types of TP can be *simple* (from a single source) and, if so, will be *contiguous* or *gapped*, so that repetitions of Theme can be called simple contiguous TP or simple gapped TP. Alternatively, thematic progression can be *multiple* (from more than one source), with at least two subtypes: *(re)integration* (two or more items from the previous text are integrated into one Theme) and *separation* (an item from the previous text is separated into

two or more Themes). Dubois's classification of Thematic Progression is represented diagrammatically in Figure 2.5 below.

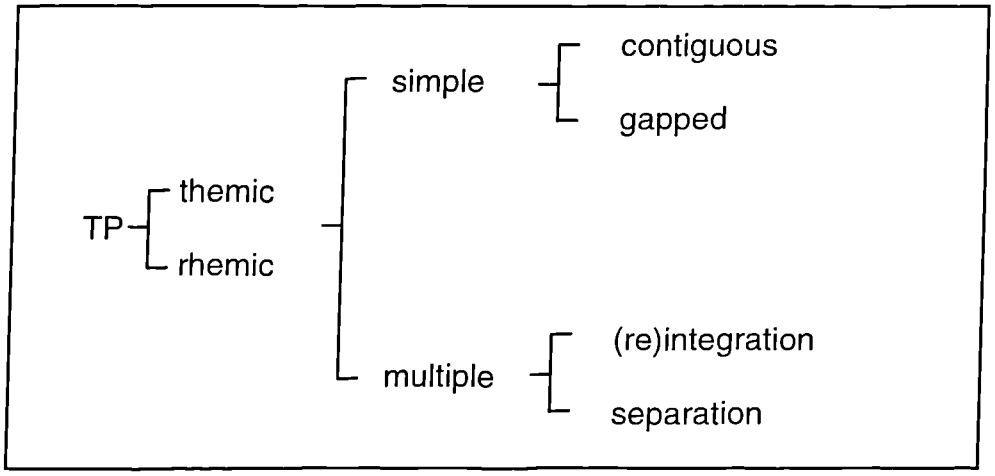


Figure 2.5. Dubois's categorisation of Themes

Dubois notes that the above categorisation does not exhaust the possibilities. For example, it is possible to have *hyperthemic* Thematic Progression and *hyperrhemic* Thematic Progression, although the asymmetric term *hypertheme* as used in Daneš's framework was avoided in her model.

2.4.1.3 TP Modified: Daneš (1989)

In a later paper, Daneš (1989) has modified the model of Thematic Progression (TP) which he proposed in 1974, and which was discussed in Section 2.4.1.1 above. In this later paper, he admits that Theme need not be a known piece of information, and agrees that 'thematisation is independent of what has gone before' (1989: 25). He further admits that TP does not contribute directly to the connectedness of text. The choice of Theme and the starting point of the utterance may be motivated by context, but they are not necessarily dependent

on context. Rather, the connectedness of text is dependent on the conditions of an underlying network of isotopic relations (NIR). By isotopic relations, Daneš means a qualified class of semantic relations among members of the set of particular *discourse subjects (DS)* of a given text. DS stands for anything that the speaker or writer has in mind when applying a nominating or deictic unit in the process of text production in order to introduce or re-introduce something. DS may be divided into three subclasses: total identity of DSs, partial identity of DSs, and objectively statable semantic affinity obtaining between DSs (1989: 24).

Daneš argues that although Halliday (1970: 357) may be right in claiming that 'the speaker is free to select whatever Theme he likes', in reality 'the speaker very often chooses as the Theme of subsequent utterances a DS known from the context, i.e. an isotopic one' (1989: 25). He adds,

> The operation of thematic choice on NIR, taking place in those cases when the speaker selects for thematisation an isotopic (known) DS, may be described and presented as a kind of Thematic Progression (TP), which represents a component of the whole development of thematic relations in a text (thematic route) (1989: 25).

TPs belong to different types and may be classified according to the three kinds of criteria shown in Figure 2.6:

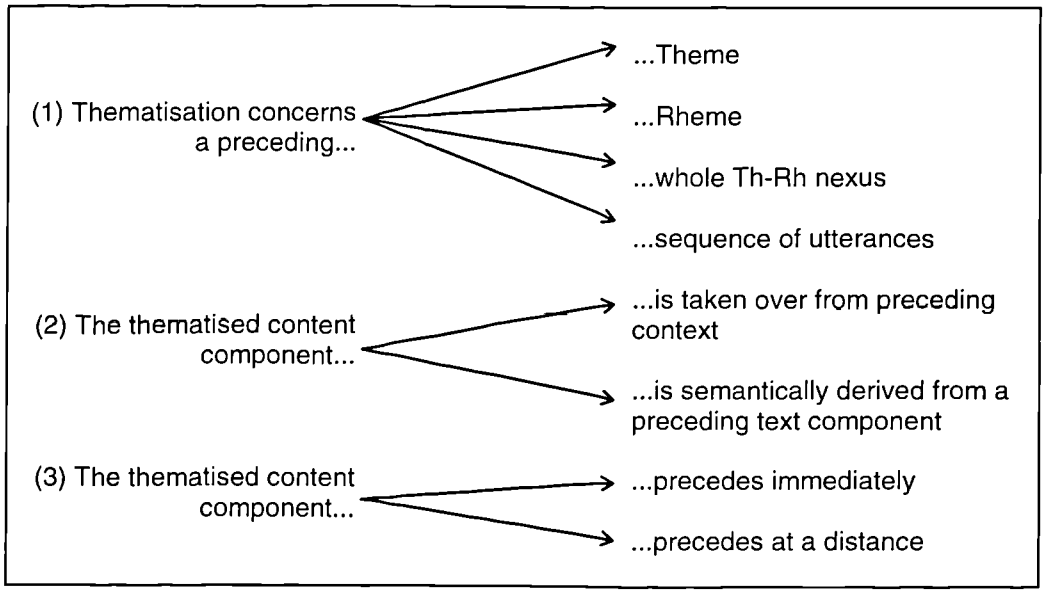


Figure 2.6. Criteria for TP classification (taken from Daneš 1989: 25-26)

Here Daneš is attempting to make TP more comprehensive in its coverage. In addition to the TPs from previous Theme or Rheme, as shown in his 1974 paper, Theme may also develop from the whole sentence or group of sentences of the previous text. The semantic content of Theme may be taken directly or derived indirectly from the previous context. There may be or may not be a gap between the Theme and the immediately preceding context; in other words, Theme may progress not only from the immediate sentence but also from a sentence at some distance.

Daneš's (1989) modified model of TP is undoubtedly an important improvement on his 1974 model. It takes more account of the semantic aspect of the thematic element in the organisation of text. This shift to the semantic aspect naturally leads to the exploration of the lexical items in Theme, which

are the major carrier of semantic content. Some studies in this area will be reviewed in the succeeding subsections.

2.4.2 'The separators'

2.4.2.1 Theme and 'method of development' of text: Fries (1983, 1995)

Exploring the status of Theme in text, Fries (1983: 144) makes six observations:

1. Thematic choice is meaningful;
2. Thematic choice is a choice of what information to use as the point of departure of the message expressed by the sentence;
3. The consistent choice of terms from a single lexical system for the Themes of the component sentences of a passage correlates with the perception that the passage has a single method of development;
4. The meaning of Theme explains the strong tendency for given information to occur at the beginning of the sentences;
5. Thematic choice is independent of the choice of what is given or new information; and
6. The Theme-Rheme organisation of the sentence forms part of a larger pattern which governs the flow of information in any English discourse.

Of the above observations, it is the third point that is of particular interest for the purpose of the present study. Therefore, it is worthy of some more detailed discussion.

Fries carried out some experiments with the lexical patterning of Themes of a text. In order to demonstrate the text-organising function of clause Themes, he

destroyed the patterns of Thematic Progression of the normal text by simply changing the order of some words in the sentences, so that no underlying organisation governed the choice of the information which was thematic within each clause. As a result, he observed, 'Each sentence when read in isolation is a perfectly possible sentence. However, when read as a paragraph they do not flow' (1983: 133).

Fries noticed that the lexical system within a text may interact with the thematic organisation of the text. For example, one of his sample texts contains three lexical systems: the first concerns living, growing and changing, the second concerns wisdom versus chance, and the third concerns concepts relating to government. He observed that the terms concerning wisdom and chance typically occur in the Themes of the component sentences, the terms concerning living, growing and changing typically occur in the Rhemes, and the terms concerning concepts relating to government occur more or less equally within the Themes and Rhemes of the component sentences of the sample text.

In his sample text, while the English constitution is the topic of the text, the semantic field of wisdom and chance is perceived to be the method of development of the text. Fries concluded that although in other paragraphs the paragraph topic may appear consistently within the Themes of the component sentences, 'there is no necessary correlation between the topic of the paragraph and the Themes of the component sentences of the paragraph' (1983: 135).

Fries makes a distinction between the method of development of the text and the topic of the text, which is signalled by a high frequency of mention in the component sentences. In addition, he distinguishes the method of development of the text from the point of the text, which indicates that lexical items revealing the point of the text occur more frequently in the Rheme than in the Theme of the component sentences. In other words, there is a negative correlation between the degree to which a lexical set occurs within the Themes of the component sentences and the degree to which the lexical set is perceived to indicate part of the point of the text.

These findings are repeated in a later paper (1995), in which Fries points out that the notion of the 'method of development' is not a structural idea but a semantic one. In this paper, he investigates the chain relation in Thematic Progression, dividing each clause in his data into three sections: Theme, N-Rheme, and Other. The Theme is the initial constituent of a clause and provides a framework for the interpretation of the message expressed by the clause. The N-Rheme is the final constituent of a clause and is the location of the unmarked placement of New information. The category of Other is for material which is neither Theme nor N-Rheme (1995: 349).

As a means of exploring the chain relation in Thematic Progression, Fries examines nominal groups which are found as part of the Themes of the clauses of the texts in his data, identifying the location of the most recent previous

member of the chain either in the Theme, N-Rheme, or Other of the earlier clause.

Fries notes that 'Thematic progressions and the experiential content of the Themes do not occur randomly in these texts' (1995: 354). The difference of TP patterns in different texts results from their being members of different genres and the different purposes of the authors. 'As the author changes task and responds to different pressures, the Themes of the clauses of the text will reflect and encode those changes.' (1995: 355),

Fries's contribution to the research into the status and function of Theme may be summarised as follows. Firstly, he has explored the study of Theme beyond the clause level, thus obtaining more useful results with regard to the text organising function of Theme than his predecessors who had concentrated on studying Theme in isolated sentences. Secondly, he uses the nominal group as the means of his investigation, and from the textual perspective, studies the chains of nominal groups in Theme. Thirdly, he distinguishes the topic of the text and the point of the text from the method of development of the text. The three categories may conflate, but they do not necessarily do so. Fourthly, Fries observes the difference between TP patterns in texts of different genres. Fifthly, he not only observes the function of Theme in text organisation, but also notices the function of Rheme as well. Especially he draws attention to the final constituent of the clause, which he names N-Rheme, i.e. Rheme as unmarked

carrier of new information. Finally, his method of analysis is objective and replicable.

Fries rightly focuses his study of Theme at the text level rather than at the clause level and used the nominal group as a means of investigation. This enables him to provide new insights into the function and status of Theme in the text. However, his data are of very limited size, mostly composed of one-paragraph passages or excerpts. This is a major limitation, which has prevented him from drawing very convincing conclusions about the text types he has described.

2.4.2.2 Thematisation and staging: Brown and Yule (1983)

Brown and Yule (1983) are more interested in the general phenomenon of 'thematisation' or 'staging' at the level of discourse than in the function of Theme within the sentence. They claim that 'what the speaker or writer puts first will influence the interpretation of everything that follows' (1983: 133). This refers not only to what is fronted within the sentence, but also to the thematisation or staging at the discourse level. For example, the text's title, headings and subheadings are all regarded as thematisation devices used in the organisation of discourse structure. They note that 'thematised elements provide not only a staging point around which what follows in the discourse is structured, but also a starting point which constrains our interpretation of what

follows' (1983: 139). They use the term *Theme* to refer to a formal category, the left-most constituent of the sentence. Following Daneš's (1974) typology of Thematic Progression, they assign two main functions to the Theme:

1. connecting back and linking up with the previous discourse, maintaining a coherent point of view;
2. serving as a point of departure for the further development of discourse (1983: 133).

They focus on the linking function and discourage the use of Theme to mean 'the main character/object/idea', which they suggested should be better termed *topic entity* instead.

They demonstrate in an experiment they report that some thematic sequences are preferred in real life communication. When subjects are required to choose from a set of possible continuation sentences, as in Example 2.4. below.

Example 2.4.

(a) The Prime Minister stepped off the plane.
(b) Journalists immediately surrounded her.
(c) She was immediately surrounded by journalists.

They noticed that there is a preference for (c) to (b) as the continuation sentence. This is because, they assume, that readers prefer to maintain the same subject or topic entity (1983: 130). They claim that a thematised referent occurring as syntactic subject is a better prompt for sentence recall; that is, sets of sentences with the main referent in the Theme position may be easier to remember. When a referent is thematised in most of the clauses in the text, it

may be claimed to be the Theme of the text. They claim that not all Themes are grammatical Subject, but rather,

> In general it seems reasonable to suggest that the constituent which is thematised in a sentence is, in some sense, 'what the sentence is about', regardless of whether or not the constituent is the grammatical subject (1983: 132).

Brown and Yule's discussions about thematisation and staging, particularly that about the preference of maintaining the same subject as Theme in continuation sentences, serves as an important basis for the present study, especially in the relationship between Theme and lexical patterns in text. This point will be carried on to the next subsection.

2.4.2.3 TP and patterns of lexis: Wikberg (1990)

Wikberg (1990) illustrates different patterns of lexis management in relation to Thematic Progression and topic development in an analysis of four short expository and procedural texts on the topic of natural light, excerpted from photography manuals. He chooses to focus on Thematic Progression, which he regards as mainly belonging to the semantic category, 'since it serves both to link up initial sentence constituents with the previous discourse and as a starting point for their further semantic development' (1990: 231). By 'previous discourse' he means the immediately preceding sentence as well as more remote parts of the text.

Wikberg notes, 'The most concrete surface manifestation of Thematic Progression in the text is the lexical patterns' (1990: 231). Therefore, he decides to analyse the patterns of Thematic Progression in his corpus using a computer program to extract key words in thematic positions from the four texts, and then compares their frequencies across texts.

Wikberg notes that although there is a great deal of agreement on the functions of Theme, there are always disputes regarding its identification. He argues that for a sentence element to be 'Theme' it should be initial, given, topical and make up the mood Subject as defined in systemic grammar. However, Theme is not equal to topic. Although topic entity can formally overlap with Theme, it does not need to do so. Further, he notes that 'even if we exclude textual (e.g. sentence connectors and conjuncts) and interpersonal Themes (e.g. disjuncts), not every initial sentence element need have anything to do with the discourse topic' (1990: 238).

Like Dubois (1987), Wikberg also distinguishes themic Theme from rhemic Theme (cf. Section 2.4.1.2 of this chapter), which are categories related to Daneš's (1974) typology of Thematic Progression. Themic Theme is related to constant Thematic Progression, repeating the entity from the Theme of the previous sentence. Rhemic Theme refers to simple linear progression, continuing the entity from the Rheme of the previous sentence. According to Wikberg, Themes may be identified as themic, rhemic, or both.

His analysis shows that the Themes in his corpus are dependent on several different factors, including the way in which the authors describe their subjects to achieve special effects and the extent to which the authors explicitly manifest the interpersonal function of language. Wikberg claims, 'Particularly expository discourse is built up by chunks of recurring lexical items which attract thematically related members as the text develops and which contribute to cohesion and thematic continuity' (1990: 233). However, contrary to the common belief that Theme correlates with 'given' information, he concludes from his analysis that 'the thematic elements contain a fair amount of new information, although the Theme as a whole is rarely brand new'(1990: 251). Wikberg's argument is supported by Ndahiro, who finds that Theme as defined as the element in sentence initial position is not associated with given information. In his sample texts, given information and new information are scattered in various sentential positions (1998: 338-339). Nevertheless, Ndahiro agrees that in the absence of intonational marking of new information in written text, the expectation would be that in the vast majority of cases given information will map onto Theme and new information onto Rheme.

2.4.2.4 Theme and nominalisation: Taylor (1983)

Like Wikberg, Taylor (1983) also tries to tackle the property of Theme from a lexical perspective, narrowing his focus specifically on to the nominal group. He regards Theme as 'necessarily a nominalised concept' (1983: 216), and notes that 'Themes must typically be nominalisations' (1983: 210). In his view,

the notion of Theme is a semantic notion, and first position is also interpretable semantically (1983: 217). 'The question of "what am I talking about?" is best answered with an independent noun phrase of some kind...' (1983: 210). Based on an analysis of a 700,000-word corpus of high school textbooks, Taylor concludes that 'a good deal of the connections between sentences is done by lexical linkage rather than by discourse markers like the conjunctions' (1983: 214). Lexical linkage, he suggests, mostly lies in the Themes of the sentences.

In his analysis of 1,232 sentences from a corpus of science and history textbooks, he found that the overwhelming majority of sentence Themes are nominalisations (1,002; 81%). Moreover, and this is important to the present thesis, these Themes are closely related to the topics of the sentences.

In addition, Taylor also noticed different types of thematic element in relation to different text types. In history textbooks, for example, he found a strong preference for adjunct Themes, introducing each proposition with a circumstantial setting. On the other hand, the science textbooks prefer clausal Themes, which present a noticeable trend towards setting up some condition as a Theme.

2.4.3 Discourse Theme

2.4.3.1 Theme and discourse topic: McCarthy and Carter (1994)

McCarthy and Carter (1994: 70) believe that the relative coherence of a text can be established across sentence boundaries by means of Theme development. They state, 'It is clear that whatever we decide to bring to the front of a clause is a signal of what is to be taken as a framework within which what we want to say is to be understood. This includes the particular attitudes or point of view which we wish to communicate... It also includes words and phrases which provide explicit organisational signals of how the text is to be read' (1994: 72).

Distinguishing interactive Themes and topic-based Themes, McCarthy and Carter claim that a Theme is interactive if it contains words or phrases which refer to the sender or receivers of the message. On the other hand, a Theme is topic-based if it contains words or phrases which refer to some aspect of the topic (including pronouns or deictics which allow a continuation of the topic) (1994: 71). Drawing on the results of a mini-experiment, McCarthy and Carter found that the reader is sometimes able to say what the whole text is about after only an initial reading of topic-based Themes isolated from the text. This last point is of particular interest for our purpose in the present thesis. If their claim is well grounded, there must be some properties of the Theme which enable the

reader to make inferences about the topic of the text. These properties will be one of the foci of investigation in the present studies.

2.4.3.2 Interactive and informational Themes: Berry (1995)

A more comprehensive framework of different types of Themes is proposed by Berry (1995), who hypothesises that thematic options are closely relevant to success or failure in writing. On the basis of a thematic analysis of the writings of four school children, she presents a network of different thematic options in a text, as shown below in Figure 2.7.

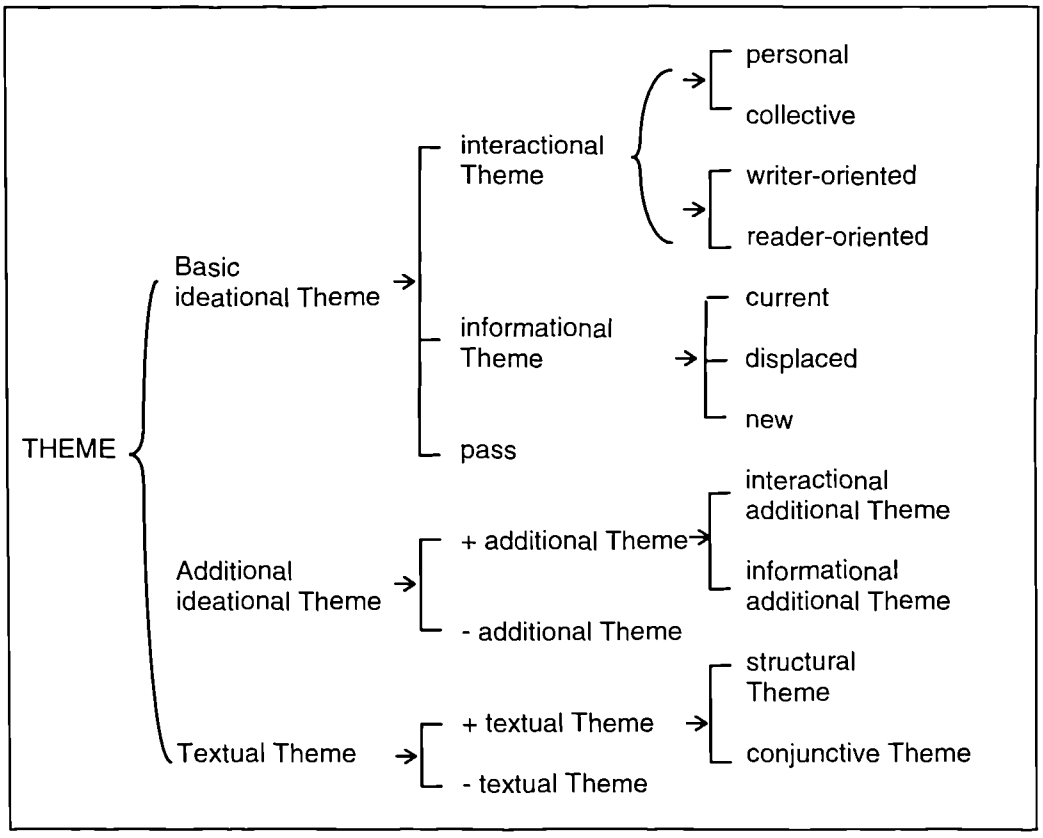


Figure 2.7. A system network of thematic options (from Berry 1995: 25)

From the start, Themes can be divided into *basic ideational Themes, additional ideational Themes,* and *textual Themes*. Basic ideational Themes, which are realised by the headword of the grammatical subject, can be further divided into interactive or informational Themes. *Interactive Themes* are mainly realised by words referring to the speaker/writer or listener/reader, especially first and second person pronouns. If the interactive Themes refer to the writer, they are writer-orientated. If they refer to the reader, they are reader-orientated. *Informational Themes*, which are also called topic-based Themes, are mainly realised by reference to the topic or aspects of the topic of the text. If they continue the topic from the Theme of the immediately preceding sentence, they are called *current Themes*. If the Themes contain retrievable information, or return to the topic mentioned earlier, but not to the Theme of the immediately preceding sentence, they are called *displaced Themes*. If the Themes introduce the topic for the first time, they are called *new informational Themes*.

Additional Themes, as the name suggests, are optional. These Themes are realised by adjuncts. They may be *interactional*, realised by for-adjuncts, such as in 'For those who like long walks...', and *informational*, realised by other adjuncts, such as in 'A few miles out from Grantham...' and 'In the centre of Grantham...'.

In conclusion, Berry claims that the two kinds of Themes, namely interactive Themes and informational Themes, which are within the category of basic ideational Themes, play different roles in writings for different purposes.

Interactive Themes may, for example, be important for distinguishing successful from unsuccessful promotional writing, whereas new informational Themes may be the distinguishing features of successful informational writing.

Berry's system networks for Theme clearly demonstrate the role of different kinds of Themes in allowing one to evaluate the success of different kinds of writing. Of these different kinds of Themes, informational Themes are of central importance to the research work of the present thesis, because this thesis is interested in finding the relationship between thematic options and patterns of lexis in terms of patterns of information distribution.

2.4.3.3 Discourse Theme$_M$: Berry (1996)

As was pointed out in Section 2.3.2.1, Halliday's (1985a) definition of Theme assigned it the dual function of being 'the point of departure in the clause as message' and 'what the clause is about'. This aroused heated debate, especially over the second function. Attempting to give a more sensible definition of Theme, Berry (1996) separates the two functions by distinguishing *Theme$_F$* from *Theme$_M$*. The former correlates with the function of 'point of departure'; the latter is more concerned with 'what the clause is about', or the priority meaning of the clause.

Berry claims that her inquiry is meaning-centred rather than form-centred. She says, 'Form is only of interest in so far as it is a means to the end of making/discovering meaning' (1996: 8). Following Hallidayan convention,

meaning can be further distinguished as ideational meaning, interpersonal meaning and textual meaning. In her analysis of Theme, Berry is 'generally more concerned with interpersonal meaning and textual meaning than with ideational meaning' (1996: 9).

Moreover, Berry focuses her analysis more on meaning at the discourse level than at the clause level. She claims that her main concern in analysing Theme is to find out the 'prioritised meaning' of the discourse as a whole rather than that of individual clauses.

Berry regards the following two points as particularly important: the points 'that Theme has to do with the concerns of the speaker or writer; and that it is the cumulative force of the Theme of the clauses of a text which indicates these concerns, rather than the Theme of any one clause individually' (1996: 18).

By analogy with the distinction between *sentential topic* and *discourse topic* in Brown and Yule (1983: 71), Berry distinguishes between discourse $Theme_M$ and clause $Theme_M$ (1996: 18). Discourse $Theme_M$ is a priority set of types of meaning that reflects the writer's underlying concerns for the duration of the text or large section of text, whereas clause $Theme_M$s may reflect the writer's priority for a particular clause. Here Berry states rather cautiously, 'If the overall priorities are to be communicated, it would seem likely that at least some of the clause priorities will reflect the overall priorities' (1996: 18).

As for $Theme_F$, however, Berry seems to treat it as an asymmetrical term, since she only talks about clause $Theme_F$ and avoids using the term 'discourse $Theme_F$'. Indeed, because $Theme_F$ is basically a structural category, and clauses are structured, it is easy to understand that clause $Theme_M$ is realised by clause $Theme_F$. Conversely, as stated at the beginning of this chapter, texts are organised rather than structured. Therefore, it is easy to infer that for Berry, discourse $Theme_F$ should not exist; discourse $Theme_M$ is realised by the accumulation of clause $Theme_F$s.

Berry assumes that '$Theme_F$, if it exists, is a grammatical means of prioritising the meanings of discourse $Theme_M$s' (1996: 31). Further, '$Theme_F$ is the location, not for prioritising meanings in general, but for priority meanings of the types that are communicatively important... i.e. interpersonal meanings...; and... logical meanings' (1996: 60). On the other hand, she suggests, '$Rheme_F$ is assumed to be the part of the clause associated with the weight of the ideational meaning, the main matter to be communicated...' (1996: 46).

Berry adopts a 'top-down' approach to the identification of $Theme_F$; that is, before $Theme_F$ is determined, she would first of all try to determine the discourse $Theme_M$, which is the priority meaning of the discourse, in a method independent from the grammatical categorisation. Her analysis can be divided into three stages. First she obtained information concerning the priority of meaning of her sample passages by interviews with the writers and readers. What the writer claimed, and the reader perceived, as prioritised meaning in a

passage was regarded as being discourse $Theme_M$. Then she counted the instances of prioritised meanings in each of her sample passages. If significantly more instances were found in the predicted passage, it was regarded as an indication that the meaning had indeed been prioritised in the passage. Finally she considered whether the meanings prioritised by frequency of mention were consistently associated with some grammatical feature such as position in clause. The part of the clause that contained the instances of prioritised meanings was regarded as the clause $Theme_F$. The clause $Theme_F$ realises the clause $Theme_M$.

In a review of major approaches to the identification of $Theme_F$, Berry (1996: 29-31) tested a list of ten hypotheses:

1. first position hypothesis (Halliday 1985a: 38);
2. first ideational element hypothesis (Halliday 1985a: 54);
3. the subject hypothesis (Enkvist 1973);
4. the preverb hypothesis (Berry 1989, 1995);
5. auxiliary verb hypothesis (Stainton 1993);
6. the lexical verb hypothesis (Berry 1996);
7. the continuum hypothesis (Matthiessen 1992: 51);
8. the overlapping hypothesis (Ravelli forthcoming);
9. unusual position hypothesis (Halliday 1985a: 45); and
10. the main clause hypothesis (n.a.).

Berry concludes that 'the sixth, the lexical verb hypothesis, looks most promising' (1996: 46). Like Ravelli, Berry regards the Process as an important pivot for identifying Theme. But unlike Ravelli, she argues for including Process in the Theme domain. Berry claims that 'the process constituent acts as

a cut off point for interpersonal meaning,... very few interpersonal meanings will occur after the process constituent, and such that do will be perceived to have a low degree of prominence' (1996: 58). Similarly, 'the process constituent acts as a cut off point with regard to the mobility of "logical meanings" ...very few "logical meanings" will occur after the process constituent and such that do will be perceived to have a low degree of prominence' (1996: 58). Logical meaning, as Berry uses it, is a subtype of ideational meaning, the other subtype of ideational meaning being experiential meaning. Berry suggests that 'logical' meaning might be broadened to include 'topical' meaning, so that it would bring together all types of meaning that have to do with making connections between the meanings of different parts of the text.

Berry's major contribution to the study of Theme is her distinction of $Theme_F$ from $Theme_M$, the latter being divided into clause $Theme_M$ and discourse $Theme_M$. This distinction rightly draws a line between the form and meaning of the notion Theme, so that it helps greatly to eliminate the confusion caused by Halliday's definition, which fuses two very distinct functions into one broad category. Furthermore, her method of analysis sets a standard for the present study. Her hypotheses were established on the basis of researcher's intuitions, and evidence was drawn from independent sources to test the hypotheses. In her analysis, the evidence was distributive, and could be tested quantitatively; thus the results could be repeated by other analysts if the same procedure was followed. In her study, she tried to find whether the prioritised meanings of

discourse $Theme_M$s were regularly associated with a particular section of a clause or other grammatical feature. This is exactly the method the present study adopts: this thesis will try to find whether there is regular association between Theme-Rheme system and the distribution of lexical repetition in the text.

2.5 Conclusion

From the above review of literature on the notion of Theme, there seems to be no disagreement that Theme can be defined as the 'point of departure' of the clause, as is manifested in McGregor's comment that 'what comes first is necessarily the point of departure of the message, whatever the language. It is hard to imagine a Theme final language in which the starting point of a clause is at the end' (1990: 7).

However, the other definition offered by Halliday (1985a) that Theme is 'what the clause is about' remains an issue of much debate. As shown in Section 2.3.2.1, Huddleston (1988) questioned the 'aboutness' of certain kinds of Themes delimited by Halliday's criterion. Downing, likewise, also argued that there are often initial elements in the clause which 'are not even remotely concerned with what the clause is about' (1991: 124). This issue may be settled, not by modifying the definition, but by clarifying criteria for the identification of Theme, i.e. the determination of the scope of Theme in the

clause. In other words, what exactly is included in Theme determines what Theme exactly is. Davies's partition of Theme into obligatory and optional elements seems to suit well for this purpose. The grammatical Subject of the clause, which is assigned as obligatory element of Theme, normally reflects the main concern of the speaker, whereas the Context Frame, made up of optional elements of Theme, which precedes the Grammatical Subject, often serves as the basis on which the discourse develops. However, as pointed out in Section 2.3.2.3, Davies's definition has its difficulties, especially in its failure to account for inverted sentences. Ravelli's (1991, 1995) path analysis of Theme from the dynamic perspective is insightful, but Berry's (1989) definition seems to have the merit of both being clear and easy to operate. However, her proposal in 1996 to include Process as Theme does not seem to fit the purpose of the present thesis, for two reasons. Firstly, more evidence is needed to show convincingly that the Process could still be suitably regarded as the 'starting point', as it is positioned so far away from the initial elements of the clause. Intuitively, it is suspected that following Berry's (1996) proposal might result in treating many sentences in my data as having only Themes without Rhemes. Secondly, it is not easy to see that the Process represents 'what the clause is about' in the usual sense more than the Subject and its preceding elements can do. Following Berry's suggestion that the prioritised meaning, or 'what the clause is about' as I understand it, may be realised by frequency of mention, a small experiment on my data was carried out, which showed that lexical verbs in a text did not repeat as frequently as the nominals in the sentence initial

position. In the event that a verb was repeated, it was more likely to refer to actions performed by different actors rather than by the same actor. Therefore, in this thesis, Berry's (1996) proposal was not followed. Instead, her (1989) method in identifying clause Theme was adopted. This will be described in more detail in Chapter 4.

If Theme functioned only on the clause level, it would not have attracted so much attention amongst linguists and language teachers. More revealing is the study of the function of Theme at the discourse level. Discourse Themes, or discourse $Theme_M$s as Berry calls them, reflect the method of development of the discourse and especially prioritise the textual and interpersonal meanings the speaker intends to convey. Therefore, in this thesis, Berry's (1996) distinction of $Theme_F$ from $Theme_M$ will be observed. Berry's claim that discourse $Theme_M$s are realised by the accumulative force of clause $Theme_M$s which in turn are realised by $Theme_F$s will serve as one of the theoretical bases for exploring the distribution of patterns of lexis in the Theme and Rheme areas of the text.

This chapter has looked at one aspect of text organisation, the Theme-Rheme system, which functions both at the clause level and the text level. It has also noted that the patterns of development of clause Themes in the text are basically realised by the choice of lexical items in Theme and Rheme. Therefore, we must now turn to another aspect of text organisation identified in

our research question, namely that of the text-organising function of lexis. This leads us to the next chapter of the present thesis.

Chapter 3. Lexis and Text Organisation

3.1. Introduction to this chapter

Chapter 2 reviewed the role the Theme-Rheme system plays both at the clause level and in the organisation of whole texts. It was noted there that the networks of clause Themes in the text are basically lexical, that is, they are realised by the choice of lexical items in Theme and Rheme. Therefore, this chapter moves on to review the literature on studies of lexis as a factor in text organisation.

As early as 1966, Sinclair called for 'beginning the study of lexis'. However, in the later 1980's Carter and McCarthy (1988) were still claiming that vocabulary study was neglected by linguists, applied linguists and language teachers. They noted that 'although interest has grown quite rapidly during the 1980s, there is certainly not much evidence of interest in vocabulary in the last twenty-five years taken as a whole, and relative to investigation at other linguistic levels' (1988: 1).

Earlier studies on the properties of text generally focused on grammar, and lexis was only mentioned as a peripheral factor. For example, in Quirk et al (1972), an influential book on grammar, progress was made in the study of text by the authors's drawing attention to phenomena on the text level in addition to those on the sentence level, with an elaborate discussion of syntactic devices that help connect sentences into a coherent text, but little was said with regard to the role lexis plays in text organisation. However, in their revised version more than a decade later, the same authors (Quirk et. al. 1985) were beginning to pay more attention to lexical cohesive devices.

Since the later 1980's, more attention has been drawn to the study of lexis. Hoey (1995) even predicts that the first decade of the new millennium will be the decade of lexis studies. His prediction is not groundless. Given the rapid development of computer tools and electronically readable texts, language already can be viewed and examined on a scale that previous studies in language could not even imagine. However, there were a few forerunners in this field. One of them is Ure, whose work published in 1971 is particularly relevant to the present thesis. So, as a starting point, this chapter will first review her work.

3.2. Lexical items and lexical density

3.2.1. Lexical density: Ure (1971)

Ure (1971) made one of the earliest attempts to study lexis in a corpus which consisted not of single texts but of sets of texts. Her corpus consisted of a number of spoken and written texts, amounting to over 42,000 words.

Ure assumed that language form is closely related to the function it plays. For example, language used in different situations might have different patterns. In the 1971 paper, her purpose was to investigate the relation between a situational classification of texts according to language use and a classification of texts according to language patterning. Her focus was on a functional variety of language, that of language-in-action.

One important notion proposed in her analysis is that of *lexical density*, which measures the proportional occurrence of lexis in the text. Lexical density is obtained by computing the percentage of word tokens with lexical properties to all the orthographic word tokens in the text. It should be noted that this notion is not as straightforward as it appears at first sight. In the first place, the 'word tokens with lexical properties' are not easy to define exactly. This point will be returned to later in this subsection. Nevertheless, Ure discovered that lexical density varies with the way the text was produced. For example, most of the spoken texts in her corpus had a lexical density of under 40%, while all but two

of the written texts had a lexical density of 40% and over. Thus she initially assumed that lexical density might largely be a matter of choice of spoken or written medium. However, she found later that the interaction between the speaker/writer and the hearer/reader, which she called 'feed-back', was an even more powerful factor in determining lexical density than the spoken/written choice. For example, if there was feed-back from the hearer/reader while the text was being produced, there would be a low level of lexical density in the text. The results of her analysis indicate that in her data all texts with feedback had a density of 36% or under, while all monologue texts (with one exceptional case) had a density of 37% and over. Other less powerful factors affecting lexical density included preparedness, personal and social relations between participants, personal-impersonal contrast, and subject matter.

Ure's attention to the overall patterns of the distribution of lexical items is insightful. As Stubbs comments, 'One important reason for studying L and G words [i.e. lexical and grammatical words] is that their relative frequency in a text differs considerably according to various features of the context in which the text is produced' (1986: 33). Ure's analysis certainly had a great influence on later work in lexis studies. The notion of lexical density, as measured by the ratio of lexical items to all running words in a text, provides a quantitative measurement to describe the characteristics of texts. For example, Halliday (1985b), when comparing written and spoken language, observes that 'written language displays a much higher ratio of lexical items to total running words'

(1985b: 61). Written language is indeed summarised as 'a language with a high lexical density' (1985b: 75).

However, there remains a number of practical problems with Ure's approach, especially in the identification of lexical items. As pointed out by Ure herself, it is not easy to accurately count the number of lexical items in a text. For example, 'turn up' constitutes one unified semantic unit, and separation of the two words may change the meaning of the unit. But in her analysis only 'turn' was counted as a lexical item and 'up' was not counted, on the grounds that it is a grammatical item. On the other hand, the word 'bookshop' seems to be undoubtedly one lexical item, but if it is written as 'book shop', should it still be counted as one, or should it be counted as two lexical items? Consequent to the above problem, there is a further problem, which is that some words are ambiguous. For example, the word form 'can' might be used as a lexical item, as in 'a can of worms', or a grammatical item, as in 'she can do it'. These problems remain a big obstacle in computer-aided analysis of text.

Although, as already noted, Ure's influence on later studies on lexis is significant, it is her methodology in obtaining objective data to support the researcher's intuitions about text properties, using text corpus and obtaining computed statistics, that is of the greatest relevance to the present thesis

3.3. Lexical items and cohesion

3.3.1. Coherence and Cohesion: Halliday and Hasan (1976)

Halliday and Hasan (1976) made the first systematic and comprehensive exploration into the system which ties up the components of a text and gives it texture. In their work, cohesive devices are categorised into reference, substitution, ellipsis, conjunction, and lexical cohesion, of which four categories belong to the grammatical cohesive devices, while lexical cohesion is only briefly mentioned in one short chapter. Most relevant to the present thesis is the notion of lexical cohesion, which is defined by Halliday and Hasan as 'selecting the same lexical item twice, or selecting two that are closely related' (1976: 12).

Lexical cohesive devices are further divided into two major types. One type is *reiteration*; the other type is *collocation.* Reiteration is realised by (a) same word (repetition), (b) synonym (or near synonym), (c) superordinate, and (d) general word. For illustration of these categories, see Example 3.1, which contain sentences taken from a sample text in the present thesis.

Example 3.1

(1) The textbooks say there are nine planets in our solar system.

(2) The most distant is Pluto, discovered on 18 February 1930 by astronomers at the Lowell Observatory, Arizona.

(7) For a number of years before Mr Tombaugh's discovery, the existence of a ninth planet was suspected, because something large was affecting the orbital path of Uranus around the Sun.

(8) The locating of Pluto was thought to explain the Uranus effect.

In Example 3.1, there are several cases of reiteration by the same word, such as 'Pluto-Pluto', 'planets-planet' and 'Uranus-Uranus'. Reiteration of a synonym or near synonym may be exemplified by 'discovery-locating' and 'affecting-effect'. Reiteration by superordinate may be exemplified by 'planet-Pluto' and 'planet-Uranus'. Reiteration by general word may be exemplified by 'a ninth planet-something large'. Reiteration means either restating an item in a later part of the discourse by direct repetition or else reasserting its meaning by exploiting lexical relations, which are the stable semantic relationships that exist between words and which are the bases of descriptions given in dictionaries and thesauri. Reiteration is not a chance event; writers and speakers make conscious choices whether to repeat, or find a synonym or a superordinate.

Collocation is defined as the association of lexical items that regularly co-occur' (1976: 284). It covers everything that shares 'the same lexical environment' (1976: 286-8). This includes items which are members from the same ordered series, like 'January-February' 'Sunday-Monday', or items from unordered lexical sets, like 'book-page', 'sky-fly' or 'flower-red', etc. Halliday and Hasan seem to be a little vague about the concept of collocation. In their description, collocation means 'the mutual expectancy between words that

arises from the one occurring frequently in the environment of the other, or of the two occurring in a range of environments common to both' (1976: 320). This virtually allows every word to be included in the domain of lexical cohesion, since whenever words co-occur in a text they come into the same environment.

The contribution of Halliday and Hasan (1976) is that they made the first systematic analysis of the devices that help make the text a coherent whole. Their definition of lexical cohesion as 'selecting the same lexical item twice, or selecting two that are closely related' links cohesion to repetition, thus making it possible (albeit inaccurately) to measure coherence quantitatively. This is particularly useful for computer-aided text analysis, as repetition can be fairly easily recognised by computer programs. Nevertheless, their work is not without limitations. One limitation is the difficulty in identifying collocation. Because of the difficulties they have in providing a clear definition of lexical cohesion, Halliday and Hasan are forced to admit that 'the effect of lexical, especially collocational, cohesion is subtle and difficult to estimate' (1976: 288). Indeed, Hasan (1984) abandoned the category of collocation.

3.3.2. Re-entry systems: Jordan (1984)

Jordan (1984) attempts to show the complexity of lexical cohesion in text organisation. He argues that although it is often educationally helpful to use very brief examples of text which are relatively simple in their relative lack of

lexical cohesion, larger and more complex texts which exhibit complexities of lexical cohesion cannot be ignored in linguistic analysis. Jordan divides lexical cohesion into systems of nominal cohesion and verbal semantics, but the former are the focus of his work. With examples from naturally occurring texts, he presents a detailed description of the systems of nominal cohesion in English use.

In his analysis of lexical cohesion both within and between sentences of the text, Jordan uses the term *re-entry* to signify the means by which the speaker/writer re-includes a previous topic into the text to say something more about it. He divides re-entry systems into two sub-categories of co-referential re-entry systems and identified re-entry systems. The former include lexical repetition, substitution, ellipsis, listing, synonymy, naming, and generic nouns; the latter include relative, non-finite and verbless clauses, and appositives between clauses of a sentence.

He further distinguishes basic re-entry systems from associated re-entry systems. The latter include nominal groups which, in context, are semantically associated with a previously introduced nominal group, 'often in such a way that the association is, or can be, overtly "triggered" within the associated nominal'(1984: 224). He holds that 'the associates can have any combination of association with one or more triggers'(1984: 224). The association can be either direct or separated in stages. If it is separated, there may be some linking stages between the distantly associated nominal groups. However, the linking

stages may or may not actually occur in the text. This is a potentially difficult point in his argument, which makes it possible to have any lexical item related with any other item, so long as they occur in the same text. In Example 3.2, taken from Jordan (1984: 234), which Halliday and Hasan (1976: 341) use in their analysis of 'ties', Jordan finds several re-entries that are either direct or triggered in stages.

Example 3.2

> Well, **I** met a thief in **my** *house*. **I** had one of those nice old *houses* - **I** was very lucky. *It* was about thirty years old, on stone pillars, with a long staircase up and folding *doors* back on to a *veranda*. And **I** came through the *door* from the *kitchen*, and a thief carrying **my** handbag emerged through **my** *bedroom door* into the *living room* at the same moment.
>
> Key: **Bold** = main item with re-entries;
> *italicised without underline* = associates;
> *italicised with single underline* = 2 stage associates;
> *italicised with double underline* = 3 stage associates.

In Example 3.2, the main participant is the speaker 'I', with many re-entries. The speaker's house is an associate with the speaker. The house is then seen to become a trigger for several other re-entries (2 stage associates) such as 'veranda, kitchen, bedroom, living room', which in turn trigger further re-entries (3 stage associates) such as the 'doors'.

The notion of association is basically semantic. It is useful in determining the relationships of nominals between clauses and sentences. However, his system

is rather complicated, and Jordan seems to use the term 'association' rather loosely. Actually, one may extend the associates in an indefinite sequence of stages, and the associates may be combined and interlocked. So, in extreme cases, every item may be seen to have some relationship with every other item. After all, the fact that they occur in the same text provides a common ground for their semantic interpretation.

Theoretically, one may set a limit to the stages, so that only prominent items are included for semantic analysis. However, this will not entirely solve the problem, since division into stages is not always clear-cut, even when the stages are all present in the text. For instance, in Example 3.2, it is not easy to accurately decide how many stages there are between the 'folding doors' and the house, and in turn the main participant.

Jordan notes that 'further work is needed to relate the concepts of re-entry and association with Winter's analysis of the matching relation and particularly comparative affirmation and denial within and between sentences' (1984: 234).

Jordan rightly focuses his study of cohesion on the nominal cohesion. This is in agreement with Halliday's observation that 'the overwhelming proportion of "content", in the sense of lexicalised meaning, is carried in the nominal groups - by nouns and their premodifying nouns and adjectives.... All the meat of the message is in the nominals' (1985b: 72). Jordan also rightly points out the complexity of lexical cohesion . His proposal to use the 're-entry' system as a

means to explore the complex phenomenon of cohesion is very useful for the description of text organisation.

However, Jordan's work has its own limitations. Firstly, he claims that grammatical cohesion is basic, and lexical cohesion is only additional to grammatical cohesion. This greatly impairs the significance of his discovery. As text is a semantic unit rather than a grammatical one, the unity of text should logically be realised mostly by means of semantic devices rather than grammatical ones. Nevertheless, in reality, it is apparent that his own focus is on the relations between lexical items in the sentences, and most of the examples he provided are examples of lexical cohesion (1984: 225). Secondly, although the identification of 'stages' between the associates is helpful in revealing the closeness of relations between the nominal items, it is not possible to achieve convincing accuracy, since it depends so much on the intuition as to the nature of relations between entities in the real world. Finally, Jordan shows the complexity of the lexical relations without also showing how they clarify the relationship between the text and the meaning it carries. Indeed, we should say, the major objective of the study of relations amongst lexical items should be to reveal the meaning of the text.

3.3.3. Large-scale patterns of lexical organisation: Philips (1985, 1988)

Philips (1988) aims to reveal the relationship between the text and the world of reality. Claiming that the meaning of text lies in its relation with 'reality', he observes that 'perception of textual meaning is a high order process not directly dependent on the organisation of text at the local level' (1988: 106). Therefore, he sets out to explore the 'large-scale patterning of text' in the hope that the relationship between text and reality may be better understood. His basic assumption is that text is composed of elements of linguistic substance juxtaposed in linear sequence whereas reality is full of complex and non-linear phenomena. Starting from this assumption, the question which naturally follows is how the complex non-linear conceptual structures of 'reality' are realised through the ultimately linear organisation of language substance. Philips suggests that lexical patterning may be relevant to the question, for 'a study which seeks evidence of such patterning [of textual substance] might be expected to throw light on how texts relate to reality and hence on how they mean'(1988: 106).

Based on the evidence from the lexical analysis of five scientific texts and two non-scientific texts, Philips claims that a text as a whole constitutes a vast network of lexical items. The vast network is composed of smaller networks, some of which are more tightly interwoven than others. Since every word in the network over the span of the whole text co-occurs with every other word, it is

possible to isolate individual networks by focusing on significantly high frequencies of co-occurrence of the lexical items (1988: 107). Using this method of analysis, Philips shows that 'the text is organised through the association of certain lexical items and thereby a particular conceptualisation of real-world phenomena is presented to the reader' (1988: 108). The identification of the association of lexical items is particularly useful because, as he argues, 'groupings of lexical items can be discerned which clearly articulate the principal cognitive content and functional purposes of their texts' (1988: 108).

Philips concludes, 'The recognition that similar networks thus exist in widely separated locations in the text indicates that semantic relationships extend over long stretches of text and create large-scale patterning... [So] large-scale patterns of lexical organisation are responsible for the structure of the subject matter as projected by the text' (1988: 109).

Philip's work is very important for the study in the present thesis, as the purpose of this thesis is also to reveal text organisation in terms of information distribution. Text organisation is essentially realised through choice and association of lexical items, but there are other factors which contribute to the text organisation, such as the Theme-Rheme system. So, to explore text organisation, it will be fruitful to further explore the interaction between large-scale lexical patterns and the patterns of Thematic Progression in the text.

3.3.4. Cohesive harmony: Hasan (1984), Halliday and Hasan (1985)

As pointed out in Section 3.3.1, Halliday and Hasan (1976) appear to regard lexical cohesion as peripheral. However, they do redress this imbalance by assigning a more even distribution to the text-forming role of lexical items and grammatical items in their later work. In Hasan (1984) and Halliday and Hasan (1985), Hasan turns her attention to the role lexical items play in the creation of cohesion in text, her aim being to measure coherence by means of cohesion. She uses the term 'coherence' to refer to the property of 'unity' or 'hanging together' of text, which relies on the semantic property of text, because, she argues, coherence involves the idea of unity and the idea 'that the patterns of language manifest or realise the existence of semantic bonds' (Halliday and Hasan 1985: 94).

According to Hasan, coherence is not an absolute property of the text but is gradable. The coherence of a text is measurable in terms of the meaning relations between component parts of the text. 'The many differing kinds of semantic relations operate at one and the same time through sizeable portions of a text' (Halliday and Hasan 1985: 83).

In discussing the notion of coherence, Hasan (1984: 184) makes two important points. Firstly, coherence is not only an essential property of texts, but it is also a relative property of the text, so it is possible to rank a group of texts on a

cline from most coherent to least coherent. Secondly, the variation in coherence in a text does not correlate with structural facts, and therefore an examination of coherence necessarily involves an examination of non-structural relations generally grouped under cohesion. Hasan emphasises that

> Non-structural relations are crucial to the creation of coherence not because structure is entirely irrelevant to it, but rather because structure is a uniformly integrative device; and as an integrative device, it does not go far enough in the explication of the notion (1984: 183)

Consequently, 'cohesion is the foundation upon which the edifice of coherence is built' (Halliday and Hasan 1985:94), and 'the cohesive devices create texture because they establish relations of meaning' (Halliday and Hasan 1985: 96). However, Hasan admits that in previous studies of cohesion, especially in Halliday and Hasan (1976), 'lexis is a neglected area' (1984: 194). Therefore, she pays greater attention to the role lexical cohesion has in the creation of coherence in text. For this purpose, she makes some revisions to the categories of lexical cohesion she had previously proposed with Halliday.

In the first place, collocation is excluded from the categories of lexical cohesion, because it is very difficult to identify collocates objectively and consistently (1984: 202). Secondly, Hasan argues, lexical cohesion belongs to two primary types: that mediated through 'general' lexical relations and that mediated through 'instantial' ones. General lexical cohesive devices are based upon the language system in general, whereas instantial lexical relations are

text-bound (1984: 201). Moreover, the categories of lexical cohesion are semantically motivated. They are shown in Table 3.1 below.

Cohesive devices		*Examples*
A. General	i. repetition	leave, leaving, left
	ii. synonymy	leave, depart
	iii. antonymy	leave, arrive
	iv. hyponymy	travel, leave
	v. meronymy	hand, finger
B. instantial	i. equivalence	the *sailor* was their *daddy*
	ii. naming	the *dog* was called *Toto*
	iii. semblance	all my *pleasure* was like *yesterdays*

Table 3.1. Categories of lexical cohesion (taken from Hasan 1984: 202)

Central to the notion of cohesion is the idea of 'two-ness'(1984: 185). This manifests in the notion *cohesive tie*, which is the relationship between two items in the text (see Figure 3.1, taken from Halliday and Hasan 1985: 73).

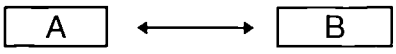


Figure 3.1. The cohesive tie

Hasan explains, 'The nature of this link is semantic: the two terms of any tie are tied together through some meaning relation. Such semantic relations form the basis for cohesion between the messages of a text' (Halliday and Hasan 1985: 73). However, she found in her initial analysis that variation in the number or quality of the cohesive ties had no effect on the reader's perception

of coherence. Therefore, she developed the notion of *cohesive chain*, which is defined as 'a set of items each of which is related to the others by the semantic relation of co-reference, co-classification, and/or co-extension' (Halliday and Hasan 1985: 84) (see Figure 3.2, adapted from Halliday and Hasan 1985: 83).

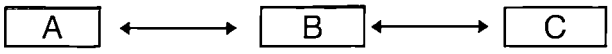


Figure 3.2. The cohesive chain

Cohesive chains are threads of continuity running through the text. These chains are composed of a succession of message components. 'The linking of components creates cohesion between messages' (Halliday and Hasan 1985: 83). Cohesive chains are further divided into two types: *identity chains* and *similarity chains*. Identity chains are text bound. They are formed by items in a relationship of co-reference; that is, each item has the same referent within the context of a specific text. The semantic bond between them may be realised through pronominal cohesion, simple equivalence, simple lexical repetition, or a combined operation of grammatical and lexical cohesion. Similarity chains, on the other hand, are not text bound. They are formed by items related by co-classification and/or co-extension (Halliday and Hasan 1985: 84-85). Co-classification may be realised by substitutive or elliptical cohesion, or under certain conditions by simple lexical repetition. Co-extension is realised through lexical repetition, synonymy, antonymy, hyponymy or meronymy (1984: 205-206).

However, Hasan observes, it is possible for a text to have all or most of its lexical items entering into chains and still not be coherent. Essential in the creation of coherence is what Hasan terms *chain interaction,* which requires that at least two members of one chain should stand in the same relation to two members of another chain (Halliday and Hasan 1985: 91). The relationships can be one of the five types as shown in Table 3.2.

	chain interaction types	*examples*
1.	actor - action	girl - went
2.	action - acted-upon	took - teddy bear
3.	action and/or actor location	girl - home
4.	saying - text	speak - English
5.	attribute - attribution	lovely - teddy bear

Table 3.2. Types of chain interaction (based on Halliday and Hasan 1985: 72)

All the chains are composed of lexical items, or what Hasan calls 'tokens'. The tokens that enter into identity or similarity chains are relevant tokens, which may be either *central tokens* or *non-central tokens*. Central tokens are those relevant tokens that interact, whereas non-central tokens are those relevant tokens that do not interact. The tokens that do not enter into any kind of chain are called *peripheral tokens*. With regard to the relation of the tokens to the coherence of text, a text is likely to be coherent if:

1. the proportion of peripheral tokens to relevant ones is low,
2. the proportion of central tokens to non-central ones is high, and
3. there are few breaks in the chain interaction (Halliday and Hasan 1985: 93).

The sum of the above three phenomena is described as *cohesive harmony.* Variation in coherence, according to Hasan, is the function of variation in the cohesive harmony of a text.

Hasan (1984) and Halliday and Hasan (1985) are important contributions to the study of cohesion. Hasan rightly deviates from her (1976) position (with Halliday) to give more recognition to the role lexis plays in text organisation. Particularly, the notion of cohesive harmony opens up the possibility of quantitatively measuring the coherence of text. However, her samples are too limited for her claims to be fully convincing. Following her work, some attempts have been made to test her claims on larger corpora of texts. Parsons (1996) is one such attempts, one that is reviewed in the next subsection.

3.3.5. Significant chains: Parsons (1996)

Parsons (1996) makes an attempt to examine Hasan's theory of cohesive harmony using statistical analysis. Parsons's data consisted of sixteen texts written by graduate students and ranking scores of coherence of the texts by a group of informants. Building on Hasan's hypothesis that the proportion of central tokens to non-central ones correlates with one's perception of coherence of the text, Parsons sought to test the correlation between the percentage of central tokens (%CT) in a text and the rank score given the text by his informants. He found that the proportion of peripheral tokens to relevant tokens was low in texts perceived to be coherent, so that Hasan's first hypothesis was

supported by his statistical tests. However, Hasan's other hypotheses were not supported. So Parsons proposed to modify Hasan's concept of cohesive harmony.

Feeling that more coherent texts might have a larger number of longer interacting chains than less coherent texts, Parsons decided to investigate whether there was a correlation between the number of long chains and the coherence of the texts. For this purpose, Parsons developed the notion of *significant chains*. A significant chain is defined as one containing more tokens than average for chains in the text. In his data a significant chain was composed of four or more tokens each. Further, he developed the notion of *significant tokens,* which are tokens in significant chains. Significant tokens are automatically *central tokens*. Those central tokens that do not enter significant chains are called *non-significant central tokens*. For his data, which had an average chain length of 3.13 tokens, he set the minimum number of tokens for significant chains as 4, slightly above the average. The significant tokens were called S4, and the percentage of significant tokens to all lexical tokens in the texts was expressed as %S4. Parsons found that the correlation between %S4 and a high rank score of coherence was statistically significant, as was the correlation between the ratio of %S4 to peripheral tokens (PT) and a high rank score of coherence.

Following this, Parsons raised the minimum number of tokens required for significant chains, setting it at the level of 5 tokens; thus the significant tokens

were referred to as S5. He found that %S5 correlated with the high rank score of coherence even better. Finally, he set the minimum number at 3. Now the tokens were referred to as 'central tokens' but not 'significant tokens' because in his data 3 was below the average number of tokens in a chain. The results with %C3 showed worse correlation than %S4.

Parsons's work is significant in that quantitative evidence was provided to support Hasan's notion of cohesive harmony, thus making a valuable contribution to the study of text organisation. However, Parsons failed to adequately explain the results obtained from his analysis. For example, to the question 'Why does the CT analysis provide a better explanation of informants' perceptions than the C3 analysis when the results arising from the examination of significant tokens would suggest the reverse?' he could only conjecture that perhaps 'there is a balance between a text being too cohesive and being not cohesive enough and that a writer needs to steer a course between not making a text insufficiently cohesive and making it too cohesive' (Parsons 1996: 598). But as to why this should be so, he could provide no explanation.

Nevertheless, Parsons proposes that we should study 'the extent to which cohesion is contributing to coherence compared with other possible factors such as thematic progression' (1996: 598). That is exactly what is intended in the present thesis, as will be reported from Chapter 4 to Chapter 6.

3.4. Lexical items and discourse topics

3.4.1. Computing lexical cohesion: Morris and Hirst (1991)

Morris and Hirst (1991) argue that lexical chains are a direct result of units of text 'being about the same thing' and that chains of related words contribute to the continuity of lexical meaning in text (1991: 21). For them, lexical chains are defined as chains of semantically related words. Therefore, 'when a unit of text is about the same thing there is a strong tendency for semantically related words to be used within that unit.' (1991: 35).

According to Morris and Hirst, lexical chains vary in strength. Reiteration, density, and length are three factors that contribute to the strength of lexical chains: the more repetitions, the stronger the chain; the denser the chain, the stronger it is; and the longer the chain, again the stronger it is (1991: 32). The three factors can combine to indicate the strength of a lexical chain. Stronger chains may be more closely related to the central topic of the text, while weaker chains may indicate supplementary information, and thus not be central to the general topic of the text.

Morris and Hirst (1991) argue that lexical cohesion is a computationally feasible clue to identifying a coherent stretch of text. Specifically, they attempt to use lexical chains to identify units in a text on specific topics. They assume

that lexical chains tend to indicate the topicality of text segments of more than one sentence. 'If a new lexical chain begins, this is an indication or clue that a new segment has begun,' whereas 'when a lexical chain ends, there is a tendency for a linguistic segment to end' (1991: 24).

Although their assumptions seem to be promising, their methods of analysis are not equally reliable. For example, serious problems occur in their identification of lexical chains. For them, apart from reiteration, any two items in the relationship of collocation may also enter into a lexical chain. As noted in Section 3.3.4, it is very difficult to identify this relationship, because the notion of collocation itself is rather fuzzy: in a broad sense, any two items that co-occur in the same text may be regarded as collocational.

Their analysis consists of two steps, the first step being to select candidate words as the basis of lexical chains, and the second being the process of building the chains. Unfortunately, they do not have objective criteria for determining lexical chains. So, they have to rely heavily on their own intuition in both steps. In selecting candidate words, they exclude all 'closed-class' words such as prepositions, verbal auxiliaries and pronouns. As I will show in my own analysis in Chapter 6 of the present thesis, the exclusion of pronouns is a serious omission, because pronouns by definition form one of the closest semantic links. Pronouns very frequently stand for a nominal group whose meaning is often easily recoverable in the same context.

Morris and Hirst also exclude high frequency words such as 'good' 'do' and 'taking' from their analysis. Although this practice may be intuitively plausible, their explanation does not appear to be persuasive: high frequency alone seems not to be the right criterion for excluding such words.

In building the lexical chains in their sample texts, Morris and Hirst have to rely on an existing thesaurus for determining the semantic relations of the items in lexical chains. This produces serious limitations. As they acknowledge, 'A thesaurus is as good as the work that went into creating it, and also depends on the perceptions, experience, and knowledge of its creators' (1991: 41). No one thesaurus exists that meets all needs. So, for example, while they identify the lexical items 'environment, setting, surrounding' as belonging to one chain, these items are not related in the thesaurus they used.

Nonetheless, Morris and Hirst's analysis is of importance to the present thesis. Their observation that lexical cohesion is the result of chains of related words which in turn are the result of units of text being about the same thing is most useful. Following their observation, identifying lexical chains may enable one to identify topics of text or sub-topics of text units. This is also a goal of the present thesis.

3.4.2. Identifying scientific terms: Yang (1986)

An attempt to extract the main content of the text by automatic means was made by Yang (1986), who proposed a method to identify automatically what was described as scientific/technical terms, i.e. lexical items reflecting the subject-matter of the scientific text.

Working on the data of a corpus of 10 scientific texts in different fields including electronics, chemistry, biology and humanity, plus a corpus of two non-scientific texts for comparative purposes, Yang noticed that some words have far greater frequencies than others in individual texts in his corpora. His intuition was that 'it is probably the scientific/technical terms or the nominals in a science text that carry most of the subject matter information' (1986: 93). Further, he hypothesised that 'since scientific/technical terms are sensitive to subject matter, they should have fairly high frequencies of occurrence in texts where they occur' (1986: 94). So 'it is possible to identify single-word terms on the basis of their frequencies of occurrence and distribution' (1986: 93).

In order to identify the terms, Yang obtained statistical information relating to the Distribution (D) of words in different fields, Average Frequency (F) of words in all the fields, and Relative Standard Deviation (SD) of a word across all specialised fields. On this basis, he further obtained information of Peak-ratio (P), which is the maximum frequency of occurrence divided by the

average frequency, and Range-ratio (R) which is the maximum frequency of occurrence divided by the minimum frequency.

Yang observed that words with very high (D) value and fairly high (F) value were mostly function words, whereas words with very high (D) value but relatively low (F) value were not function words. They were what some linguists call 'sub-technical words', similar to Winter's 'Vocabulary 3' words. So it was possible for him to filter out those words. Further, he found that while both function words and sub-technical words had fairly low SD, P and R, scientific/technical terms showed very low D but very high P and R values. The respective fields for the high P and R values characterise the field specificity of scientific/technical terms. For instance, the term 'hexadecimal' would occur with very high frequency in computer science literature, but may never occur in dentistry literature. Therefore, he concluded, it seemed possible to identify single-word scientific/technical terms by setting appropriate P and R ratio filters.

However, Yang did not put forward any objective criteria for setting the P and R ratios. His ratios were chosen empirically, with the P value set to greater than 5 and R greater than 10. In spite of this limitation, Yang's results are promising. The lists of terms abstracted from his corpus seem fairly satisfactory.

Yang's discovery is significant in that the terms were abstracted automatically mainly on the basis of frequency of occurrence and distribution. If his method is reliable, it may apply to the automatic summarisation of scientific texts. Since the terms so abstracted were closely related to the specific subject matter of the text, they might indicate the content of the text. Yang's methodology in selecting scientific terms is similar to the methodology whereby key words as defined later in this thesis are selected. Both the selection of scientific terms and the selection of key words are based on the frequency of occurrence of a certain word in the text as compared with the frequency of occurrence of the same word in a larger body of text(s) or even in the language in general. For detailed description of how key words are selected, see Sections 1.5 and 5.1.1.

3.4.3. Patterns of lexis in text: Hoey (1991a)

As noted in Chapter 2, Hoey (1991a) proposes that text is best viewed as organised instead of structured. Moreover, he strongly suggests that texts are not organised linearly and that 'motivated selections from a text may make sense' (1991a: 187). Aiming to shed fresh insights into text organisation, Hoey (1991a) proposes a new system of analysis based on the study of cohesion, particularly lexical cohesion. With a focus on the repetition patterns of lexical items, Hoey (1991a) notes that if cohesion is to be interpreted correctly, it must be interpreted in the context of the sentences where it occurs.

In a separate paper, published in the same year, Hoey (1991b) proposed to view cohesive harmony from another perspective. He notes that Hasan's (1984) notion of cohesive harmony depends on the interaction of cohesive chains. Hasan's perspective is one that focuses on global relationships between chains and uses local relationships between elements from different chains to establish such larger relationships, as presented in Figure 3.3.

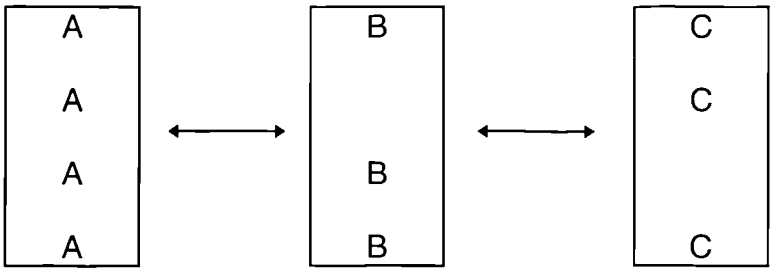


Figure 3.3. Abstract representation of interacting chains (taken from Hoey 1991b: 389)

While lexical items A, B and C respectively form cohesive chains running through sentences, what is important is the interaction between the three chains. Hasan's chain interaction is doubtless insightful in that it provides a measurement for the coherence of text. However, Hoey comments on Hasan's model of chain interaction that

> just as the existence of two or more chains may be grounds for looking at the relationship that may hold between the chains, so also the presence of a cluster of cohesive ties linking two sentences may be grounds for looking more closely at the way that the pair of sentences relates (1991b: 390).

He views the phenomenon of text cohesion from a new angle. For him, the relationships between the cohesive ties that form cohesive chains may be

viewed both vertically and horizontally. Therefore, he claims, 'A cluster of cohesive ties connecting two sentences in a non-narrative text has the effect of relating the messages in a manner not exhaustively accounted for by the ties alone' (1991b: 390). Not only do chains run through sentences as single threads, but the sentences which the chains run through may also be regarded as related 'whole messages'. This may be seen more clearly from Figure 3.4 below.

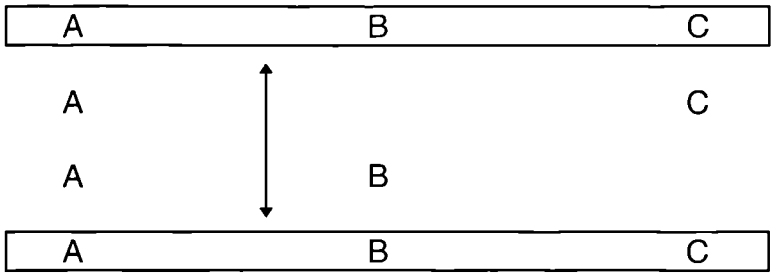


Figure 3.4. A complementary abstract representation of interacting messages (taken from Hoey 1991b: 390)

Sentences sharing the cluster of lexical items A, B, and C will, Hoey argues, be more closely related than sentences that do not share all the three lexical items. Hoey explains, 'In some cases the relationship will be shown to be so close that the pair may be found as coherent as normal prose;... In many cases ... the two sentences are seen to be mutually relevant.' In any case, 'the pairs - through which cohesive chains pass - are related as "whole messages"...' (1991b: 390).

It is from this perspective that Hoey develops his own system to reveal the organisation of text. In this system, Hoey proposes that lexical items which are repeated are linked with each other, and any pair of sentences which have an

above average number of links (in any case no less than three) may be regarded as 'bonded'. When the bonded sentences are placed together, they make sense, and very often produce coherent text. Moreover, 'bonded pairs make sense *in the same way as they did in the original text*' (1991a: 192, original emphasis).

Hoey divides lexical cohesive devices into nine categories, which are listed below in decreasing order of importance (1991a: 83):

1. simple lexical repetition
2. complex lexical repetition
3. simple mutual paraphrase
4. simple partial paraphrase
5. antonymous complex paraphrase
6. other complex paraphrase
7. substitution
8. co-reference
9. ellipsis

Simple lexical repetition occurs whenever a lexical item is repeated with no variation other than that allowed by the item's grammatical paradigms; e.g. *ape – apes, woman – women*. Complex lexical repetition occurs whenever two items share a lexical morpheme but differ with respect to other morpheme(s) or with regard to their grammatical function; e.g. *argue – argument, work (verb) – work (noun)*. Simple paraphrase occurs whenever a lexical item may substitute for another in context without loss or gain in specificity and with no discernible change in meaning (1991a: 62). The paraphrase is mutual when the two items

involved can substitute for each other. See Example 3.3, taken from a sample text in the present thesis.

Example 3.3

1. Climate has again been indicted, but if this were the cause we would expect small *species* to suffer more than large, for they are more sensitive.
2. Yet in the late Pleistocene, the extinctions are of larger *animals.*

In Example 3.3, '*species*' (in Sentence 1) and '*animals*' (in Sentence 2) are cases of mutual paraphrase. The paraphrase is partial if the substitution works in one direction only. See Example 3.4.

Example 3.4

1. Professor Fouts taught Booee sign language at the Institute of Primate Studies at Oklahoma, but resigned when the Institute sold the *chimp* and other apes for medical research.
2. Years later, when Professor Fouts visited Booee, the *animal* remembered him and signed his name.

In Example 3.4, it is possible to replace the word *chimp* by the word *animal,* but not vice versa. Therefore it is partial paraphrase. This is because, Hoey notes, although the distinction is trivial semantically, it is more acceptable to replace a particular term with a general term, but less acceptable to replace a general term with a particular term.

Complex paraphrase may be said to occur when two lexical items, which share no lexical morpheme, are semantically related, mediated by a third lexical item, which may or may not actually occur in the context. For example, *writer* and *writings* are complex repetition, *writer* and *author* are simple paraphrase, thus *writings* and *author* are complex paraphrase, whether the word *writer* actually occurs in the context or not. Antonymous complex paraphrase covers cases where two items that do not share a lexical morpheme are nevertheless understood to be the antonyms, or opposites, of each other; e.g. *hot* and *cold*, *dry* and *wet*. This relationship may be best illustrated by a 'link triangle' as shown in Figure 3.5, which is based on Hoey (1991a: 65).

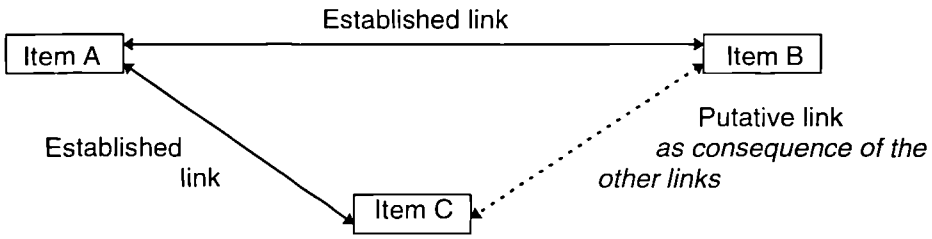


Figure 3.5. Link triangle

In the link triangle, if item B is in a complex repetition relation with item A, and item A is in a simple paraphrase relation with item C, then item B and item C are in a complex paraphrase relation, as a consequence of their relations with item A.

Substitution is used as in Quirk et al (1972) to mean any item which stands in for one or more earlier lexical items; e.g. personal pronouns. Co-reference occurs if, and only if, two items are interpreted as having identical referents in

the same context. It often requires encyclopaedic knowledge; e.g. *Mrs Thatcher* and *the Prime Minister* in a text written in the 1980s in Britain are likely to be co-referential, but this is much less likely in the 1990s. Ellipsis is considered to exist where a sentence is grammatically incomplete unless something is supplied from earlier in the text (1991a:74).

In the analysis, links internal to the sentence are not recorded and their connection with earlier sentences not noted (1991a: 84). That is to say, if more than one link is formed by two occurrences of the same lexical item in a sentence with an item in another sentence, it is counted as only one link.

Any pairs of sentences with three links or more are regarded as bonded. The complete set of all the bonds in a text can be presented as a net to show the interconnections of the bonded sentences (1991a: 92). It is interesting to note that 'bonding accurately identifies related pairs of sentences in a text, and the net they combine to create accurately reflects the organisation of the text' (1991a: 193). Non-adjacent sentences connected by multiple repetition may make sense when placed together, and the bonded sentences can be used to produce coherent sub-texts from the main text (1991a: 48).

Hoey's study in patterns of lexis in text has three outstanding features. In the first place, Hoey places much more emphasis on the lexical items than grammatical items as the cohesive devices to give the text texture. He argues that lexical cohesion is the dominant mode of creating texture. 'In other words,

the study of the greater part of cohesion is the study of lexis, and the study of cohesion in text is to a considerable degree the study of patterns of lexis in text' (1991a: 10).

Secondly, Hoey (1991a) regards the relationship of cohesive chains as much more complex than that of a one-to-one linear relationship, as implied by Hasan. He points out, 'Lexical cohesion is the only type of cohesion that regularly forms multiple relationships' (1991a: 10). The lexical item that links with its immediate predecessor also has links with previous items at distance. Thus Figure 3.2 in Section 3.3.4 should be redrawn as Figure 3.6, where item C is not only related to item B, but also related with item A.

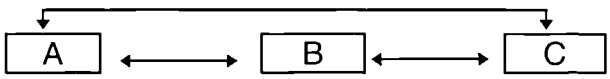


Figure 3.6. Multiple relations of the lexical chain (adapted from Hoey 1991a: 72)

Thirdly, Hoey (1991a) regards text as composed of a succession of whole messages as opposed to a succession of only message components. He argues that, when we are reading or listening, for example, 'in addition to perceiving ties between words in the sentences we encounter, we also see relationships between the sentences as whole units' (1991a: 12). Moreover, Hoey regards sentence pairs which are related by means of repetition of lexical items also as whole messages. This not only happens between adjacent pairs of sentences, but also happens between sentences bonded by repetitions of lexical items, however far apart the sentences may be.

Hoey's system employing lexical repetition has a direct consequence on our understanding of the nature of language. This can be clearly shown from a comparison of his model of language with a popularly recognised model proposed by Halliday (1961), who placed grammar at the heart of language. On the basis of Halliday's model of language, Hoey (1991a) proposes a model which gives much more prominence to lexis. See Figure 3.7 below, which is based on Hoey (1991a: 208 and 218).

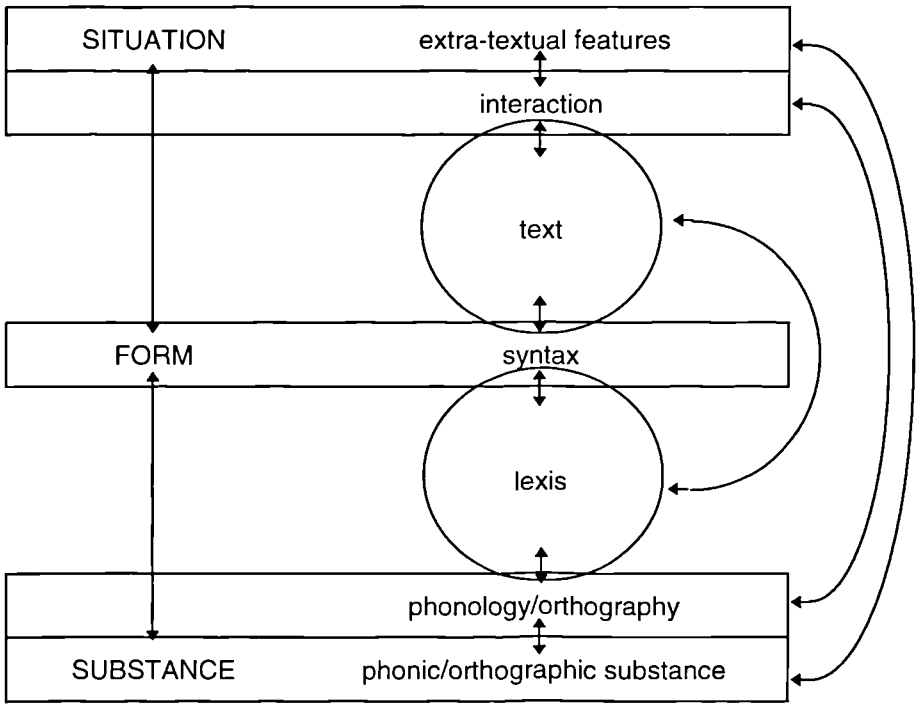


Figure 3.7: Hoey's model of language

In Hoey's model, circles represent levels which are viewed as organised rather than structured. These are the levels of text and lexis. As Hoey reiterates throughout his book, the difference between a structural view and an organisational view lies in predictability. The structured levels are predictable, whereas predictions cannot be made about text and lexis.

Hoey's model of language has three major implications. Firstly, lexical study must track the use of lexis in text. Secondly, lexical study must treat collocation as a textual phenomenon. Thirdly, any textual theory and/or description will be expected to make close reference to, and find at least some of its evidence in, the lexical realisation of the text (1991a: 218-219).

The central notion of his system, as shown by the title of his book (1991a), is that of patterns of lexis, which are realised by the repetition of lexical items in a text. Lexical repetition occurs both within and between sentences, and in the latter case, both between adjacent sentences and sentences at long distances. The central claim of his book is that

> ...all non-narrative texts produced by competent adult writers are made up of a network of relations between non-adjacent as well as adjacent sentences, and that some of these will take the form of bonded pairs (1991a: 190).

A practical output of Hoey's system is 'a methodology for the production of readable abridgements of text that is capable of some degree of automation' (1991a: 3). For the application of his system to text summarisation, Hoey proposes three different procedures: omitting *marginal sentences*, collecting *central sentences*, and collecting *topic controlling sentences*.

Marginal sentences are sentences that have no bonds with other sentences in the text. These sentences may be regarded as not making any direct contribution to the main theme of the text (1991: 105). Hoey shows convincingly that omitting the marginal sentences from a source text does not

affect the main concerns of the text. Although only one sixth of the sentences in his example are marginal, he assumes that it is not unusual for more than fifty percent of sentences to be marginal, so that it is possible to produce a shorter version of the original text that is both representative and readable.

Central sentences are sentences that have an unusually high level of bonding with other sentences in the text. Centrality is a comparative concept. It may be adjusted according to the nature of the text and the required length of the abridgement. Hoey presented two abridgements of different lengths of the same source text by adjusting the centrality threshold. Although the style of the resulting abridgement may be arguable, there seems no doubt that the central content of the source text is well preserved in either case.

Topic controlling sentences are themselves central sentences. They may be further divided into *topic opening sentences* and *topic closing sentences*. The former refers to sentences which have more bonds with their succeeding sentences than with their preceding sentences, thus are regarded as setting up topics in the text. The latter are sentences which have more bonds with their preceding sentences than with their succeeding sentences, thus are regarded as closing down topics in the text. Representing three versions of abridgement using the topic controlling sentences, Hoey observes that abridgements produced in this way may be more focused topically than abridgement produced by simply selecting central sentences. The reason is that, he assumes, 'the chosen sentence controls the subject-matter for all others' (1991: 143).

Hoey's work opens an exciting area in the study of text organisation. But it is not without limitations. On the one hand, his theory is claimed to be only applicable to non-narrative texts, which forces the analyst to be very careful in choosing texts for analysis. On the other hand, since all his analysis was carried out manually, he had to use a text of limited size. Following his work, attempts have been made to apply his system to automatic summarisation of longer texts, one of which will be reviewed in the next sub-section.

3.4.4. Lexical items and text summarisation: Benbrahim and Ahmad (1994)

Benbrahim and Ahmad (1994) set out to examine Hoey's theory related to the existence of patterns of lexis in text. Applying Hoey's theory to text analysis, Benbrahim and Ahmad designed a computer program called TELEPATTAN, an acronym standing for '*te*xts and *le*xical *patt*erns *an*alyser'. This program is an interactive program, in which the user can specify the source text, the thesaurus or dictionary of the language of the text, a patterns file and output text file where the results of the analysis can be stored. The user can also adjust the bond threshold, namely the number of links required to form a bond between sentences with both upper and lower bond values.

Their analysis is carried out in three stages. The first stage is text analysis, which consists of parsing, sentence indexing, and lexical patterns extraction. The second stage is to represent analysed text, either in graphical patterns or in

textual patterns. The third stage produces output, in the forms of text summaries and candidate terms specific to the text.

In the text analysis stage, they first use a parser to extract tokens from a stream of characters of a text. A text may contain non-lexical tokens, such as numerals, punctuation marks and mathematical expressions, as well as lexical tokens. In their analysis, they emphasise lexical cohesion based on lexical tokens only, so all the non-lexical tokens are excluded from their analysis. Then they index the sentences, adding sentence numbers, mostly for the purpose of reference. Finally, they extract lexical repetition patterns. This last step distinguishes their work from others. Whereas most other attempts either identify only simple repetition or identify limited forms of complex repetition by means of inflection, they identify complex repetition and paraphrase repetition by exploring the use of thesaural information. They use a modern English thesaurus called the Macquarie Thesaurus, which contains over 180,000 entries. This procedure is important but it is not as straightforward as it appears. For example, they need to conduct a simple morphological analysis prior to any comparison, so as to bring together word forms such as 'phenomena's-phenomena-phenomenon', which may be related to 'fact', 'event', or 'circumstance' in the thesaurus.

An output of the analysis is text summary, which is produced by selecting sentences above a certain bond threshold. Following Hoey closely, Benbrahim and Ahmad experiment on various procedures, such as omitting marginal

sentences, selecting central sentences and selecting topic controlling sentences, to produce different versions of text summaries. They find that all of Hoey's claims are valid over much longer texts than experimented on by Hoey himself. However, they hypothesise that topic opening sentences may be more useful to produce a brief account of the theme of the source text than other categories of sentences.

Trying to extend Hoey's theory, Benbrahim and Ahmad propose that it is possible to select candidate terms specific to a text by computing the relative frequencies of the terms in the text and recalling corresponding frequencies of these terms in the general language. They claim that these terms may be indicative of the main topic of the text. Unfortunately, because they are over-brief in their description of how the terms could be computed, it is difficult to assess the validity of their proposal, although their notion of 'terms' may be related to Scott's (1996) notion of 'key words', which is described in Sections 1.5 and 5.1.1 of the present thesis.

Nevertheless, Benbrahim and Ahmad's work is significant, especially in three aspects. Firstly, following Hoey (1991), they have made one of the first attempts to apply lexical analysis to practical text summarisation. Secondly, they try to apply Hoey's theory to more than one language, using not only an English text, but also a Welsh text. Thirdly, they are among the first to use computer technology to analyse patterns of lexis in text. Fourthly, benefiting

from the use of computer technology, they are able to analyse a larger quantity of textual data than is possibly handled manually.

3.5. Conclusion

This chapter has reviewed the literature of studies on the function of lexis in text organisation. Halliday and Hasan (1976) suggested that texture, which is the essential property of text, was created mostly by grammatical cohesive devices. However, Hasan (1984) admits the neglect of lexical cohesive devices in her work with Halliday in 1976 and pays more attention to the role lexis plays in creating texture. She notes that cohesive devices create texture because they establish relationships of meaning in the text and meaning in ideational terms is basically carried by lexical items. Therefore she proposes the notion of 'cohesive chain', which is composed of items that are semantically related to each other in the text. But essential to the creation of coherence in the text is 'chain interaction', which reflects the interrelationships of meaning between individual cohesive chains. There are two important points in Hasan's proposals. The first is that the cohesive chains are basically formed by lexical items in the text, thus logically raising the status of lexis in text organisation. The second is that the notions of cohesive chain and chain interaction are closely related to the notion of repetition, which may be regarded as a fundamental device for creating texture and text organisation. It is by repetition that the chains are realised and consequently are identified. The movement of

Hasan's position from 1976 to 1984 and 1985 reflects a general trend towards the recognition of the status of lexis in text organisation. However, it is in Hoey (1991) that the status of lexis in text organisation, and consequently in the study of language system, is fully recognised. Hoey's position is well grounded, because if text is a semantic unit, as claimed by Halliday (1985a) and others, then it follows logically that it should be organised mostly by semantic means rather than grammatical means. Therefore, in order to understand how text is organised, and how language works for human communication, it is necessary to study the relationship between lexical items in the text. In other words, the study of text organisation should be largely the study of patterns of lexis in text.

The next chapter will begin the analysis of the present study. It is based on the assumption that Thematic Progression patterns as proposed by Daneš (1974) are basically realised by the development of lexical chains, or patterns of lexis in the Theme and Rheme areas. Therefore, it will investigate Thematic Progression patterns in relation to lexical patterns in the text.

Chapter 4. Lexical Links in Theme and Rheme

4.1 Introduction to this chapter

We have seen in previous chapters that the Theme-Rheme system and patterns of lexis both play an important role in text organisation. In this and next two chapters we attempt to answer the following research questions raised in Chapter 1:

1. Is there a relationship between lexical patterning and the Theme-Rheme system and, if there is, how is it manifested in the text?
2. Is there a relationship between key words and the Theme-Rheme system and, if there is, how is it manifested in the text?
3. Is there a three-way relationship amongst lexical patterning, key words and the Theme-Rheme system and, if there is, how is it manifested in the text?
4. What are the implications of such relationships for our understanding of text organisation?

We begin in this chapter by attempting to answer the first question, that is, attempting to investigate the nature of lexical patterning in the Theme and Rheme areas of clauses in the written English text, and in so doing to test a discovery procedure for identifying the patterning of information distribution in text, and ultimately to shed some light on the property of text organisation. It is hypothesised that the framework of thematic progression (Daneš 1974) may indicate the distribution of information in the text in a way correlated to lexical 'bonding' (Hoey 1991a), and that it should be possible to make use of lexical repetition in the Theme area to help improve our understanding of written English texts.

It is widely acknowledged that throughout a text information is distributed in recognisable patterns. We saw in Chapter 3 that Hoey proposes a theory of using patterns of lexis in text to find the 'central sentences' which represent the central messages of the text. He claims that repetition of lexical items helps form 'bonded' sentences and bonding may accurately reflect the organisation of the text (1991a: 193). As was reviewed in Chapter 2 of the present thesis, patterns of thematic structures also reflect the distribution of information in the text. For example, Fries (1990) confirms that 'it was fairly easy to show that the information placed in the Themes of the component clause complexes of short texts relates strongly to the perceived flow of information of that text' (1990: 3). Likewise, Hasan and Fries (1995) suggest that Theme choice construes method of development and indicates topic continuity in a text. 'There are some fairly obvious grounds for suggesting that the patterns of

thematic progression concern only topical theme, and that these may be primarily relevant to some aspect of the field of discourse' (1995: xxxiii).

4.2 Basic assumptions of lexical links in Theme and Rheme

As we saw in Chapter 2 of the present thesis, Daneš (1974) proposes three major types of thematic progression (TP), which are thought of as the most common methods of text development. From his own examples it may be assumed that the criteria for identifying TP patterns are basically lexical.

Hoey (1991a) observes, 'relations between sentences established by repetition need not be adjacent and may be multiple' (1991a: 20). Indeed, he claims elsewhere that 'in a discourse the relationship between adjacent sentences may be considerably less strong than that between non-adjacent ones' (1983: 32). Hoey provides examples to show that sentences at a distance of around 4,000 intervening sentences apart can still be bonded by repetition of lexical items, and there seems to be no upper limit for the distance of bonding (1991a: 149 ff.). From his analysis, it can be seen that lexical links are characteristically extended to sentences at a distance. Likewise, Daneš (1988) modifies his TP patterns by considering the possibility that TP may exist between sentences at a distance in addition to TP between adjacent sentences in the text.

In this chapter, we shall base our investigation of lexical relationships on the TP patterns proposed by Daneš (1974), and we shall also follow Daneš (1988) to include TP relationships between sentences at a distance, no matter how far apart the sentences are in the text.

Two of the basic types of lexical link in the TP patterns proposed by Daneš (1974), which we call Type A and Type B links, are represented in Figure 4.1 below.

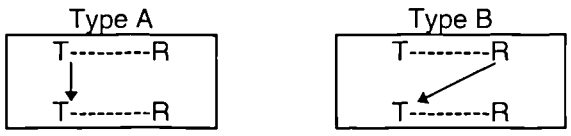


Figure 4.1. Two basic types of lexical link between Theme and Rheme based on Daneš's (1974) TP patterns

In Figure 4.1, each arrow represents a link formed by the repetition of a lexical item between two sentences. A vertical arrow between T and T indicates a lexical link between the Themes of two sentences, and a slanted arrow between R and T indicates a lexical link between the Rheme of a preceding sentence and the Theme of its succeeding sentence.

Daneš's third TP pattern, that of 'super-Theme', is not considered here. One of the reasons for the exclusion of the super-Theme is that this kind of TP involves semantic relations which depend so heavily on the context and individual readers' intuition that consensus is not always reached. In addition, Daneš's third TP pattern is realised by lexical items which are morphologically

different, so that it poses a great difficulty for the available computer programs to identify. Further, according to Hoey (1991a), the most important means by which lexical links are created is by simple repetition of lexical items. This implies that lexical links created by paraphrase or collocation are inessential. Therefore, at the present stage of text-linguistic theory and computer software development, we would content ourselves with an investigation of lexical links realised by simple repetition, though some forms of complex repetition were also considered whenever possible.

For the purpose of this study, however, Daneš's two basic TP patterns will be extended to include lexical links between Rhemes (Type C) and those from Theme to Rheme (Type D) (cf. Lau 1992), as shown in Figure 4.2 below.

Figure 4.2. Two more basic types of lexical link between Theme and Rheme

As in Figure 4.1, the arrows in Figure 4.2 represent links formed by the repetition of lexical items between sentences. Here a vertical arrow between R and R indicates a lexical link between the Rhemes of two sentences, and a slanted arrow between T and R indicates a lexical link between the Theme of a preceding sentence and Rheme of its succeeding sentence.

We could easily combine the basic Types of A to D to get more types according to the repetition patterns of lexis between the sentences. Figure 4.3 below presents six more types of lexical link:

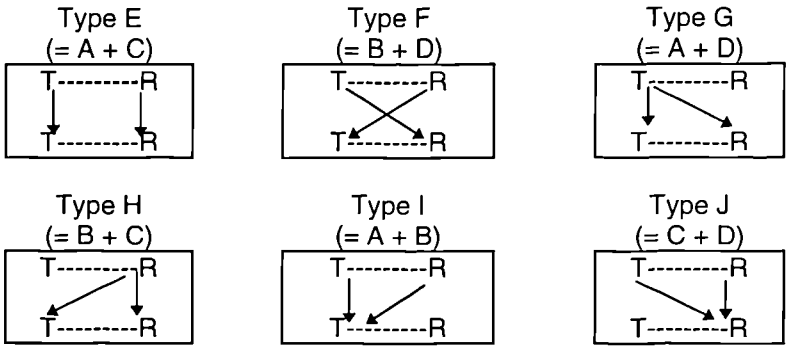


Figure 4.3. Combination types of lexical link

In their discussion of the meaning of cohesion, Halliday and Hasan (1976) imply that Theme is more capable of creating cohesion than Rheme. They assume that '*other things being equal*, it seems that the most probable target of a cohesive reference item is the Theme of the preceding sentence'(1976: 312, original emphasis).

In Hasan (1984) and later in Halliday and Hasan (1989), as we saw in the previous chapter, Hasan explores the notion of cohesive chains. The chains consist of repeated items that create the continuity of the text. However, what is significant is not the occurrence of isolated chains, but the interaction of chains, when two or more members of a chain stand in an identical functional relation to two or more members of another chain. In her example of a story about a little girl and a teddy bear, Hasan notes that the lexical item 'girl' (she/her) occur in a chain of 17 members, the lexical item 'bear' (it) occur in a chain of

14 members, and the lexical item 'home' occur in a chain of 2 members (1989: 90-92). In her presentation of the chain interaction, however, the chains interact largely around the little girl, instead of the teddy bear, in spite of the fact that the two items 'girl' and 'bear' occur with similar frequencies in the text. Why do the cohesive chains tend to interact around the item 'girl'? One possible answer may be that the entity that functions as the node of interaction occurs more often in the Theme than in the Rheme. The item 'girl' almost invariably occur in the Theme area of the text, with only one occurrence in the Rheme area, whereas the 'bear' occurs mostly in the Rheme area with only three occurrences in the Theme area. Intuitively, the passage is felt to be more about the little girl than the teddy bear. Because the writer has her focus on the girl and wishes to draw the audience's attention to the same entity, she can achieve her goal by exploiting the two devices of lexical repetition and thematisation simultaneously.

In summary, Theme is well recognised as having a scene-setting or frame-setting function. Viewing text organisation from a different angle, Hoey (1991a) observes when he discusses pattern of lexis in text that 'there is informational value to repetition, in that it provides a framework for interpreting what is changed'(1991a: 20). If this framework formed by lexical repetition is set in the Theme area, it will presumably play an important role in the building of topical continuity in the text. There is a further logical connection to be made here, which is that Rheme's association (albeit flawed)

with 'New' information would support the idea that this is what will be changed.

4.3 Hypotheses

Since Theme is an important area for identifying patterns of information distribution, and Theme-Rheme structure and patterns of lexis both play an important role in forming texture and bearing messages, it is natural to assume that some correlation between Theme-Rheme structure and patterns of lexis may exist, which will be of interest for investigation. On the basis of this general assumption, we shall set up five hypotheses concerning the relations between the Theme-Rheme structure and patterns of lexis in text.

4.3.1 Hypothesis 4.1

As reviewed in Chapter 2, Brown and Yule (1983) observe that in their use of language people prefer to maintain the same subject or discourse topic entity. They also note that 'pieces of discourse about a "main character" are frequently organised into sets of sentences in which the character is referred to by the noun phrase acting as syntactic subject' (1983: 135). This is in line with Halliday's idea of grammatical subject as unmarked Theme. Likewise, Katz (1980: 26) comments, 'The notion of discourse topic is that of the common theme of the previous sentences in the discourse, the topic carried from

sentence to sentence as the subject of their predication.' Theme is more closely related to 'given' information, which implies repetition of information from the preceding text, whilst Rheme is likely to carry 'new' information, which may or may not be a repetition of the information in the preceding part of the text. Therefore, it may be assumed that in a naturally occurring text, most of its constituting sentences will have a common topic, which is likely to be realised by the repetition of lexical items in the Theme position.

In order to test this claim, we need to set up testable hypotheses. My first hypothesis is therefore that proportionally the set of lexical items in Theme will tend to form more links than lexical items in Rheme, or in other words

Hypothesis 4.1. Proportionally, there are more T-T links than R-R links between sentences in a text.

T-T links and R-R links, as well as other links in the following hypotheses, refer to the lexical links represented in the figures in Section 4.2 of this chapter.

4.3.2 Hypothesis 4.2

If the links run across the Theme-Rheme boundary between sentences in a text, they may either run from the Rheme of the preceding sentence to the Theme of its succeeding sentence (R-T links) or from the Theme of the preceding sentence to the Rheme of its succeeding sentence (T-R links). Information in a text tends to flow from the 'new' in a preceding sentence to the 'given' in a

succeeding sentence, and generally Theme is more associated with the 'given' and Rheme is more associated with the 'new', so, if the links run across the Theme-Rheme boundary between sentences, it is more likely for them to run from the Rheme of the preceding sentence to the Theme of its succeeding sentence than from the Theme of the preceding sentence to the Rheme of its succeeding sentence. My second hypothesis is therefore that,

Hypothesis 4.2. Proportionally, there are more R-T links than T-R links between sentences in a text.

This line of reasoning may be supported by Giora (1983: 155), who regards the presentation of a theme in a previous Rheme constituent as the most elementary of Daneš's (1974) TP patterns. It seems logical to assume that the 'given' information, which is associated with the Theme of a sentence, is likely to continue from the 'new' information associated with the Rheme of a preceding sentence, whereas it is unlikely that the 'new' information of a sentence should repeat the 'given' information of a preceding sentence.

4.3.3 Hypothesis 4.3

In addition to the differences between lexical links in Theme and Rheme in their frequency of occurrence, there may also be differences in cohesive power as well. For example, lexical links of Types A and B as presented in Figure 4.1 may tend to create a closer semantic relationship between two sentences

concerned than Types C and D as presented in Figure 4.2. Put simply, my third hypothesis is that

Hypothesis 4.3. T-T links or R-T links will create more coherence than R-R links or T-R links, where coherence is measured in terms of readers' judgements about the text

This seems to agree well with common sense intuition. We all have the experience that the reader normally tries to establish links between two adjacent sentences, even without the aid of explicit lexical links. If two sentences are put together in real-life communication, they will normally be assumed to compose a coherent text, as the reader will try to find continuity of topic between the sentences. The establishment of continuity will certainly be facilitated and reinforced by the repetition of lexical items. If the repetition occurs between a lexical item in the Theme of a sentence and a lexical item in a preceding sentence, the sense of topical continuity will certainly grow. Since Theme is the location for the topical element in the sentence, the repetition of lexical items in this location is likely to reflect the repetition of the topic of the text. If a sentence contains lexical items in its Theme which repeat the lexical items in the Theme of a preceding sentence, there will be a double play of lexical links and thematic links. If the lexical repetition occurs between the Theme of a sentence and Rheme of its preceding sentence, the sense of continuity will also be enhanced, since an important function of Theme is to

serve as a peg to connect a sentence with its preceding text (cf. Halliday 1985a).

4.3.4 Hypothesis 4.4

Conversely, if the link is between the Rheme of a sentence and Theme in the preceding sentence, the sense of continuity will be interrupted by the lexical items in the Theme of the same sentence. It is counter-intuitive that the 'new' information in the Rheme of a succeeding sentence should continue the 'given' information from a preceding sentence of the text. Such links, although not impossible, are very likely to occur by chance, and thus they are likely not to be genuinely useful devices for forming semantic relationships between sentences. Therefore, among the various types of lexical link presented in Figures 1 to 3, the T-R links (Type D) are likely to be the weakest in linking power. Likewise, among the combination types of lexical link in Figure 4.3, Type I might be expected to create the closest relationship, Type J to create the loosest relationship, and Types E to H to create relationships of moderate closeness.

Since Theme usually contains topical elements, it is very likely that a series of sentences with recurrent lexical items in the Theme are dealing with the same 'topic' or 'sub-topic'. Consequently, if there is a repetition of lexical items between the Themes of a series of sentences in the same text, it is reasonable to suppose that these sentences might have a close relationship in meaning, with or without additional links elsewhere. Hence my fourth hypothesis:

Hypothesis 4.4. In the same text a series of sentences with T-T links tend to be closely related in meaning.

According to this hypothesis, it ought to be possible to gather sentences from the source text with repetition of the same lexical items in the Theme to form a sub-text on focused sub-topics. Such a sub-text may however possibly be felt to focus too much on a certain topic without any topical development and therefore may, paradoxically, be felt to be less than satisfactory to some readers.

The four hypotheses together imply that lexical repetition in the Theme is a more useful means of identifying cohesive links between sentences than repetition in the Rheme. Across Theme-Rheme boundaries, the links from a preceding Rheme to a succeeding Theme are being claimed to have much greater semantic linking power than the links from a preceding Theme to a succeeding Rheme.

If the hypotheses can be supported, it will then be possible to exploit the features of lexical repetition in the Theme and Rheme areas for making sub-texts of particular sub-topics from a source text.

In the remainder of this chapter, I will test the validity of the above four hypotheses by examining the different types of lexical link in a sample text taken from a British newspaper.

4.4 The sample text

The text selected for analysis is entitled 'The Invisible Influence of Planet X', taken from the British newspaper 'The Independent on Sunday', published on 18 February, 1990. Because this is the first sample text to be analysed in this thesis, it is encoded as Sample Text A (See Appendix 1). This particular text was selected for analysis for two main reasons. The first reason is that as a newspaper article it is 'domain free', belonging to a genre familiar to the general reader (Bell 1991). The other reason, and a rather more important one, is that, in a pilot study, problems that might distract the reader's attention from the focus of the investigation should be kept to the minimum. Because this is a well analysed text perhaps familiar to many discourse analysts, it suits this purpose very well (e.g. Hoey 1996).

Sample Text A is of moderate length, consisting of 1,171 words. This text belongs to the genre or text type of 'science report', or, to use Bell's (1993) words, 'science popularisation'. In this type of text, according to Bell, papers from scientific journals are turned into news stories. Bell observes that one of the prominent features of such science reports is that 'the paper is given the hard-news flavour of an event, released *today* even though it has been awaiting publication for months' (Bell, 1993: 217, original emphasis). Although the sample text from 'The Independent on Sunday' does not appear as the hard news of 'today', it is indeed related to 'today' in its subtitle and is intended to cater for current public interest. This text presents a striking theory that,

contrary to the common belief that there are nine planets in our solar system, there is very likely to be a tenth plant waiting to be discovered. Although the theory itself is attractive enough, it may appear to be groundless if not supported by convincing details of reasoning. To persuade the readers of the logic of this theory, the text reports on scientific activities carried out by several renowned astronomers, developments of their research projects, and the achievements they have gained through decades of hard work, all of which may further attract the attention of the public. To make the text more attractive, the reporter, or rather the editor, stresses what she/he regards as most 'newsworthy'; that is, what the text is about is explicitly highlighted. As will be shown later from the lexical analysis in Section 4.5 of this chapter, the sample text is 'agent/actor oriented': it repeatedly mentions the names of scientists or institutions with which the scientific activities and achievements are closely connected. In order to attract the widest possible readership the sample text appears as a story, with technical terms replaced by more understandable ones. In addition, it is presented with the visual aid of a photograph of Neptune by the U.S. space probe Voyage 2, to help give a flavour of scientific narrative and convince the readers that it is telling a true story rather than a piece of fiction.

4.5 Methods of analysis

My analysis was carried out in five stages, as follows:

Stage One, text preparation, was further divided into three steps:

1. preparing the text for analysis, that is, restoring the pronouns and ellipses in the original text into full lexical items which carry the meaning of the pronouns and ellipses;
2. dividing the text into Theme and Rheme areas; and
3. making separate wordlists from the Theme area and the Rheme area.

Stage Two, testing Hypothesis 4.1, was carried out in the following three steps:

1. searching for the actual location of link-forming items;
2. counting each type of lexical links; and
3. calculating the ratios between lexical links and lexical items in Theme and Rheme areas.

Stage Three, testing Hypothesis 4.2, was carried out as a single step, namely:

1. calculating the ratios between different types of lexical link in Theme and Rheme areas;

Stage Four, testing Hypothesis 4.3, was carried out in the following four steps:

1. calculating the ratios between different types of lexical link running across Theme and Rheme areas;
2. collecting a series of sentences with one lexical item repeated throughout in the Themes of these sentences, ideally with no other lexical links, or at least with no more than two links between the sentences;

3. preparing questionnaires and conducting informant tests; and
4. analysing the informant-test results.

Stage Five, testing Hypothesis 4.4, was carried out in the following three steps:

1. collecting series of sentences with minimum numbers of lexical repetition in different types, ideally with no other lexical links, or at least with no more than two links between the sentences;
2. placing the sentences together to form two kinds of sub-texts, on the bases of the respective hypotheses; and
3. testing the topicality and readability of the sub-texts.

The remainder of this chapter will report on the procedures of the analysis and discuss the results.

4.6 Preparing the sample text for analysis

4.6.1 Lexicalising pronouns and ellipses

In the first place, to prepare the sample text for analysis, I needed to make explicit cohesive links implied by the pronouns and ellipses by replacing the pronouns and ellipses with the lexical items which carry their meaning in the text. This procedure may be described as 'lexicalising', because semantically nothing was added, except that the sentences containing such pronouns would

be made more independent from the co-text and therefore a fairer picture of the cohesive links would be presented. This follows Hasan's (1984) practice in her analysis of cohesive harmony, when she carries out a lexical rendering - she lexically interprets pronouns and restores references lost through ellipses - before proceeding with her analysis.

Initially, only the simple repetition of lexical items was counted and pronouns were excluded from the analysis, mainly for the purpose of simplifying the procedures of automatic text analysis. After all, it is difficult at present to program computers to recognise what a particular pronoun stands for. But, unfortunately, the omission of pronouns seriously distorted the picture. Although pronouns as grammatical items are often excluded from cohesive analysis (cf. Morris and Hirst 1991), most readers would intuitively agree that they do play a very important role in creating links between sentences, especially adjacent sentences.

Pronouns such as 'it', 'he' and 'that', having been lexicalised, were counted as ordinary lexical items in the analysis. However, the need to include such items may pose a challenge to future automatic text processing. In the analysis, pronouns had to be lexicalised manually before the computer programs were applied for identifying links. At present, it is not yet practicable to teach the computer to find out what the pronouns stand for, though some linguists, such as Langacker (1996) and Woolls (1993), have made attempts to automatically identify the referents of pronouns in text. In the future, it may be possible at

least to program the computer to recognise pronouns and search for potentially relevant lexical items in the environment. If it is capable of doing so, it will presumably recognise only one link per pronoun.

In the analysis, lexicalisation of pronouns was kept to the minimum. Wherever possible, a pronoun was replaced by just one lexical item. This one-pronoun-one-link principle is important for presenting a fair picture of linkage between the sentences. Although Sinclair (1994) observes that a pronoun may be viewed as encapsulating the meaning of a stretch of text, i.e., much more than a single lexical item can do, the fact that a pronoun is used instead of repetition of the whole stretch of the text can be interpreted as suggesting that the author intends to establish only one link between this pronoun and its referent in the preceding text. Indeed, there will be further complexity if the number of links created by pronouns is not well controlled. For example, as seen in Chapter 3 of the present thesis, Winter (1977) observed that a number of lexical items, which he categorised as 'Vocabulary 3', also have the function of referring back to the preceding text for their interpretation, just as pronouns can do. If Sinclair's argument were accepted, the meaning of such lexical items would likewise be retrieved from other lexical items or stretches from the preceding text. To be consistent in our analysis, Vocabulary 3 words were not differentiated and pronouns were treated as ordinary lexical items. Therefore, throughout the analysis, the principle of one link per pronoun applied with very few exceptions.

In the lexicalisation of pronouns, those pronouns with definite referents in the co-text did not cause much difficulty. Although sometimes a pronoun in the sample text might stand for more than one word in a nominal group, only the head word was selected from the nominal group to represent the pronoun. No matter how many words were needed to state what the pronoun denotes, only one link was assumed to exist between this pronoun and its referent. For example, in Sentence 5 of Sample Text A, the pronoun 'he' was lexicalised as 'Harrington' instead of 'Robert Harrington'.

(4) One [astronomer], *Robert Harrington* of the US Naval Observatory in Washington, has begun a new search for 'Planet X'.

(5) He [*Harrington*] is using similar techniques to those [techniques] used by Clyde Tombaugh 60 years ago to discover Pluto.

Grishman (1986) observed that 'the subject is most likely to be the focus of the current discourse, and hence more likely to be referred to anaphorically' (1986: 131). But in our thematic analysis, subject is a compulsory element of the Theme. It follows that there are likely to be many pronouns in Theme. If therefore a lexicalised pronoun was permitted to form more than one link, the number of links would be likely to increase in Theme more than in Rheme. The restriction of the number of links created by pronouns is applied here in order to play safe, because this will make the requirements more rigorous for my hypotheses to be supported.

However, there were a few occasions when it was necessary to replace the pronoun with more than one lexical item:

(26) ~~Its~~ [*Charon*'s] motion showed {26a} that the mass of Pluto is a thousand times too small to influence the giant planets.

(27) ~~This~~ [*Pluto being small*] has become the strongest evidence for the mysterious tenth planet.

The pronoun 'this' had to be replaced by more than one lexical item, because it stands for a proposition rather than a single concept. Even so, one may still argue that the replacement has not restored all the information carried by the pronoun. The pronoun 'this' in Sentence 27 might be thought to stand for the whole of Sentence 26 rather than Clause 26a only. However, in the sample text, this is the only case where we had to lexicalise the pronoun with more than one lexical item.

Adopting an unorthodox approach to the process of understanding text, Sinclair (1992) argues that a sentence contains a single act of reference that encapsulates the whole of the previous text (1992: 10). He says,

> A word of reference like a pronoun should be interpreted exactly like a proper name or a noun phrase. The reader should find a value for it in the immediate state of the text, and not have to retrieve it from previous text unless the text is problematic at that point (1992: 9).

Sinclair's argument could perhaps explain the difficulty in determining how to treat the pronouns. Nevertheless, for quantitative analysis, at the present state of knowledge in text analysis, it is still necessary to replace the pronouns by relevant lexical items from the text, for the purpose of finding the formal links between the pronouns and their referents.

4.6.2 Delimiting Themes

When the pronouns were lexicalised and ellipses recovered, the next step in the text preparation was to delimit the Theme of each main clause. However, in practice, the delimitation of Theme was not without problems. In this section, some remarks will be made about the approaches adopted to the problems.

The delimitation mainly followed Halliday's definition of Theme that it is the starting-point for the clause as message. But because Halliday's practice of delimitation is too limited to account for 'what the text is about', modifications were made mainly following Berry and Ravelli's argumentation, so that the scope of Theme is extended to the 'lexical verb' or the 'Process' (cf. Chapter 2).

In the present analysis, the delimitation of Theme mainly follows Berry (1989); namely, Theme was regarded as everything that precedes the lexical verb of the main clause, including the subordinate clause that precedes the main clause. However, contrary to the proposal by Berry (1996), the main verb was not included as part of the Theme, for reasons given in Section 2.5 of Chapter 2. On the other hand, to account for imperative and interrogative sentences which Ravelli did not include in her exemplification, the domain of Theme was slightly extended up to the main verb. As a result, auxiliary verbs were included in the Theme, while main verbs were not.

The Theme-Rheme delimitation was based on the clause because Theme-Rheme structure is basically a category at clause level. Nevertheless the analysis of the links still had the sentence as the basic unit, since the bonding technique uses the sentence as the basis of analysis (Hoey 1991a: 214). Therefore, in the analysis, all sentence level Themes were analysed.

In his analysis of 'if winter comes, can spring be far behind?' Halliday offers two versions of analysis. One version regards the whole dependent clause of 'if winter comes' as Theme; the other analyses the two clauses separately (1985a: 57-58). He regards both as equally useful. However, Berry adopts a different position: she includes the subordinate clause preceding the main clause as part of Theme, and conducts no further analysis of the subordinate clause (1995:10). Berry's (1995) position was adopted in the present analysis, that dependent clauses preceding the main clause were regarded as part of Theme, while their own thematic structure was left unanalysed. However, if the dependent clause followed the main clause, it was treated as a separate clause.

In the analysis, non-finite clauses were not analysed; only finite clauses were analysed. This is different from Halliday's practice. The reason for this decision is that, since the thematic structure in non-finite clauses is not as prominent as that in finite clauses at the sentence (clause-complex) level, leaving non-finite clauses unanalysed would present a clearer picture of the overall organisation of the text.

Similarly, juxtaposed clauses and projected clauses were regarded as having more than one main clause. Co-ordinated clauses, when their subjects were present, were also treated separately. When a clause occurred in round brackets '( )' in the sample text, it was likewise counted as a separate clause.

According to Halliday, imperatives, having no explicit subject or modal verb, 'may be considered as consisting of Rheme only, the thematic component or request being left implicit' (1985a: 49). Although Halliday also gives an alternative analysis of the imperatives, assigning the lexical verb a thematic status, in the analysis it was decided that imperatives should be regarded as having implicit Theme 'you', but no attempt was made to count the elliptical 'you' as Theme; i.e. no Theme was recorded in such cases. There are two reasons for this decision. One is that since the verb is 'fixed' in its position, its thematic force is rather weak; the other is that it would be more consistent in the analysis not to include lexical verbs of the main clauses in the Theme. In imperatives where 'you' or 'let's' appears, the 'you' or 'let's' is analysed as Theme, since they explicitly point out the concern of the speaker.

Halliday states, 'A form such as 'what the duke gave to my aunt' [in 'what the duke gave to my aunt was that tea pot'] is an instance of a structural feature known as *nominalisation*, whereby any element or group of elements takes on the functions of a nominal group in the clause. Any nominalisation, therefore, constitutes a single element in the message structure'(1985a: 42, original emphasis). This form is also treated under the category of 'thematic equative'

(or 'pseudo-cleft sentence' in formal grammar) (1985a: 43). As a single element, its internal thematic structure was not analysed. When such an element occurred in the first position of sentences in my data, the whole clause was treated as Theme.

When a clause occurs within another clause, but still functions as a constituent of a sentence, i.e. not as part of a nominal group, it is not an embedded clause, but an 'included clause'. Halliday gives an example 'He did eventually get permission, *however reluctantly it was given*, from his father and partner to have leave of absence...' (1985a: 64). In the analysis, the thematic structure of such clauses was analysed separately, while the interrupted clause was analysed as a continuous entry.

Halliday treats the form 'I think...' and 'I don't believe...' as grammatical metaphor, equivalent to 'in my opinion...(not likely)'. Therefore it is regarded as a single entity as interpersonal Theme (1985a: 58-59). However, in 'so he thought...', 'he' is regarded as the Theme and 'thought' the Rheme (1985a: 65). In the analysis, if this form occurred before the projected clause, it was treated as a separate clause. For example in 'he said...', 'he' was regarded as the Theme and 'said' as the Rheme. However, if the form occurred after the projected clause, it was not analysed, since its thematic function in the text was felt to be very weak in this position. Likewise, in the form '...said she', it is unreasonable to regard 'she' as the Theme, since 'she' is not the 'point of departure' in Hallidayan convention. Yet it was equally difficult to assign 'said'

the status of Theme, since it is arguable whether this is 'what the clause is about'. Besides, it would be more consistent not to include lexical verbs of independent clauses in the Theme. Therefore this form was not regarded as an independent clause, but treated as a 'tag' to the main clause, in a way similar to the treatment of minor clauses. The thematic structure of such a form was ignored in the analysis.

In the form 'it + be + ...' as in 'it's love that makes the world go round' (cleft sentence in formal grammar), the internal predication thematises the 'new' information. This structure is treated as predicated Theme by Halliday (1985a: 59). Similarly, in his analysis of a sample text, Halliday treats 'it was only after his departure' in 'it was only after his departure that they discovered...' as Theme. Huang (1996) argues that this structure presents a 'enhanced Theme'. In this analysis, Halliday's approach was adopted, and the element before the main verb of the 'that' clause was regarded as Theme. Likewise, Halliday's treatment of 'there' in 'there + be + ...'(existential clause) as Theme and the remainder of the clause as Rheme (1985a: 65) was followed.

The last few sentences in the sample text may serve as an example of how Theme was delimited in the analysis:

(56) *Dr Harrington* says {56a} *astronomers still do not* understand the outer regions of our solar system.

(57) *He* [*Harrington*] hopes {57a} *Planet X will* help explain the mysterious 'wobble' of Uranus and Neptune.

(58) *'I [Harrington] think we [Harrington, etc.]* have a 50-50 chance of showing {58a} *that the anomalies* are due to another planet orbiting 10 billion miles from the Sun.'

In the above example, Themes are printed in italics, sentences in the sample texts are numbered, and dependent clauses within a sentence, when their Themes are analysed, follow curly brackets with appropriate letters, such as {a}, {b}, {c}, etc.

4.6.3 Making word lists

Once the Theme and Rheme had been clearly demarcated, I set out to obtain statistics with regard to the text and make word lists of various kinds. Using the 'WordList' program in Scott's (1996) 'WordSmith Tools' suite (cf. Section 1.8.2, Chapter 1), all the running words in (a) the full text, (b) the Theme area and (c) the Rheme area were counted respectively, so that an overview of the distribution of words in the sample text could be achieved. Then three lists of lexical items in the full text, the Theme area and the Rheme area, respectively, were made. Grammatical items were excluded from this group of lists with the use of a stopword list. Grammatical items, except those pronouns which had already been lexicalised, were not counted as establishing links between sentences (cf. Stubbs 1986, Zora and Johns-Lewis 1989, Butler 1985: 219-220, and especially Gibson 1993:162 for a discussion of difficulties in differentiating lexical and grammatical words). Finally, the 'Compare Version' function of the WordList program, which is renamed 'Consistency' in the later

versions of WordSmith Tools, was used to compare the lexical items in the three word lists. At this point, only those lexical items which occurred more than once were retained, because they were link-forming items, and those items which occurred only once were dropped, because they did not form links by simple repetition (See Appendix 2).

The first step, as stated, was to count the occurrence of lexical items and all the running words in the Theme and Rheme areas. The results of the counting of word tokens are shown in Table 4.1 below.

	Theme	Rheme	Overall
Grammatical & Lexical tokens	426	745	1,171
Lexical tokens	236	404	640
Lexical tokens that form links	144	259	403

Table 4.1. Word distribution in the sample text

As has been reviewed in Chapter 3, Ure (1971: 445) proposed to differentiate text types by means of lexical density, which she assumed might be a key to the features of the language patterning in English. The lexical density of a particular text is obtained by comparing the number of lexical items with the number of all orthographic words in the text. Table 4.2 below represents the ratios of lexical items to all running words. In addition, the ratios of lexical items that have a minimum frequency of two occurrences in the text and thus are regarded as link-forming items to all running words are also represented in Figure 4.2 below, with Grammatical & Lexical items as 100%. This shows that between Theme and Rheme in the sample text there is no significant difference

of lexical density, either in terms of lexical items or lexical items that form links.

	Theme (%)	Rheme (%)	Overall (%)
Grammatical & Lexical tokens	100.00	100.00	100.00
Lexical tokens	55.40	54.23	54.65
Lexical tokens that form links	33.80	34.77	34.42

Table 4.2. Lexical density in Theme and Rheme

Drawing on Table 4.2 above, with the overall tokens in the whole text as 100%, the ratios of tokens in the Theme and Rheme areas to the full text were obtained, which are around 36% and 63% respectively. The ratios represented as percentage are shown in Table 4.3 below.

	Theme (%)	Rheme (%)	Overall (%)
Grammatical & Lexical tokens	36.38	63.62	100.00
Lexical tokens	36.87	63.13	100.00
Lexical tokens that form links	35.73	64.27	100.00

Table 4.3. Ratios of tokens in Theme and Rheme

Table 4.3 shows that, for the sample text, there is no significant difference between the ratios of running words or lexical items in Theme and Rheme in relation to the full text. On the contrary, grammatical items and lexical items together, lexical items alone, and lexical items that form links are all evenly distributed in the Theme and Rheme areas in the sample text. At first glance, the figures suggest that the links might also be evenly distributed in the Theme and Rheme areas. However, as will be shown later, this assumption proved unfounded. The focus of our investigation in this chapter is the distribution of links between sentences in the text, not merely the distribution of words. If the

proportions of words and links in Theme and Rheme are different, there must be something worth investigating.

4.7 Analysis

Now that the text was prepared and word lists were made, I was in a position to test the hypotheses. This was done step by step.

4.7.1 Testing Hypothesis 4.1

Hypothesis 4.1 of this chapter is that 'proportionally, there are more T-T links than R-R links between sentences in a text'. In the first step of testing Hypothesis 4.1, I needed to find the different types of links as mentioned in Section 4.2 of this chapter.

For this purpose, a computer program was used, because manual counting is liable to mistakes and inconsistency, while the results obtained by computer programs would be consistent and reliable. The 'Concordancer' program in the 'WordSmith Tools' suite is a very powerful tool, well suited for searching for the locations of the link-forming items in the sample text. During the search, the wildcard symbol '*' was used to replace prefixes or suffixes as a way of lemmatising the lexical items. For example, the lexical items 'accuracy' and 'accurately' were lemmatised into one item, which occurs three times in the

sample text. Table 4.4 below represents a fraction of the search results. For the complete table showing the locations of all the lexical items which form links in Theme and Rheme of the sample text, see Appendix 3.

Lexical tokens	Theme	Rheme
accuracy/ately	32	11, 47
area	54	46
astrometry/nomer/s	3, 4, 6, 10, 11, 12, 13, 18, 19, 21, 23, 28, 56	2

Table 4.4. Sentence locations of three link-forming lexical items in Theme and Rheme

Table 4.4 shows that the lemma 'accuracy/ately', which stands for the lexical items 'accuracy' and 'accurately', occurs in three places of the sample text: the Theme of Sentence 32, and the Rhemes of Sentences 11 and 47. This suggests that these items form three links in the sample text, involving three sentences, one between the Rheme of Sentence 11 and the Theme of Sentence 32 (Type B, an R-T link), one between the Theme of Sentence 32 and the Rheme of Sentence 47 (Type D, a T-R link), and one between the Rhemes of Sentences 11 and 47 (Type C, an R-R link).

Although the locations of lexical items are important, the proportion of different types of links is what we are more interested in at this time. However, this feature is not very clearly shown in tables such as Table 4.4. So another table was made to indicate the relations of sentences by these different types of links. Table 4.5 below is a sample of this table, which shows different types of links between the sentences. Except those in the square brackets, figures in Table 4.5 are all sentence numbers. Reading horizontally, we can see that

Sentence 1 has no Type A (T-T) links with any other sentences in the sample text, but it has one Type B (R-T) link with each of Sentences 2, 7, 14, etc., one Type C (R-R) links with Sentences 3, 4, 11, etc., two Type C (R-R) links with Sentences 23 and 34, which is marked by a [×2] symbol, three Type C (R-R) links with Sentences 42 and 56, which is marked by a [×3] symbol, and four Type C (R-R) links with Sentence 9, which is marked by a [×4] symbol. The remainder of the table is read in the same way. For the complete table of links between the sentences of Sample Text A, see Appendix 4.

Sentence	A (T-T)	B (R-T)	C (R-R)	D (T-R)
1		2, 7, 14, 37, 39, 45, 55, 57	3, 4, 9[×4], 11, 12, 13, 15, 16, 18, 23[×2], 26, 27, 34[×2], 42[×3], 43, 44, 48, 56[×3], 58	
2	37, 39, 45, 48, 55, 57	6, 8, 10, 19[×2], 21[×2], 24, 28, 56	50	15, 16, 27, 32, 34, 42, 43, 44, 58

Table 4.5. Basic types of links between sentences of the sample text

On the basis of the table in Appendix 4, a count was carried out of the number of pairs of sentences with the four basic types of lexical link mentioned in Figures 1 and 2 of this chapter, namely Theme-Theme links (Type A), Rheme-Theme links (Type B), Rheme-Rheme links (Type C), and Theme-Rheme links (Type D). In my counting of the four basic links, if a sentence had more than one type of link, for example a Theme-Theme link together with a Rheme-Theme link, then neither link was included in this counting. They were counted later among the combination types. The distribution of the four basic types of lexical link is presented in Table 4.6 below.

Types→	*A (T-T)*		*B (R-T)*		*C (R-R)*		*D (T-R)*	
Links↓	*S*	*L*	*S*	*L*	*S*	*L*	*S*	*L*
1 link	166	166	136	136	168	168	61	61
2 links	11	22	21	42	36	72	4	8
3 links	4	12	4	12	11	33	3	9
4 links	0	0	0	0	1	4	0	0
Total	*181*	*200*	*161*	*190*	*216*	*277*	*68*	*78*

Table 4.6. Distribution of four basic types of links in the sample text

In Table 4.6, figures in the shaded 'S' columns indicate the number of sentence pairs involved and figures in the clear 'L' columns indicate the number of links. For example, under the heading of Type A, 166 sentence pairs are involved at the level of 1 link per pair; 11 sentence pairs are involved at the level of 2 links per pair; and 4 sentence pairs are involved at the level of 3 links per pair. The total number of sentence pairs involved is 181, while there are 200 links in total. The number of sentence pairs is much greater than the number of all the sentences in the original text. This is because of multiple relationships amongst the sentences. Any two sentences in the 58 sentences of the sample text have the potential to form a pair. So, theoretically, there could be 1,653 sentence pairs in the text (i.e. $58 \times (58\text{-}1) / 2$) .

Another count was carried out of the number of the six combination types of links, i.e. Theme-Theme plus Rheme-Rheme links (Type E), Rheme-Theme plus Theme-Rheme links (Type F), Theme-Theme plus Theme-Rheme links (Type G), Rheme-Theme plus Rheme-Rheme links (Type H), Theme-Theme plus Rheme-Theme links (Type I), and Rheme-Rheme plus Theme-Rheme links (Type J). The number of combination types of links is presented in Tables

4.7 to 4.12 below. The leftmost column in the tables show how these combination types are formed by the number of links of the first basic type plus the number of links of the second basic type. For example, in Table 4.7, '1 + 2' means one T-T link plus two R-R links; '1 + 3' means one T-T link plus three R-R links, etc.

Types→ Links↓		Type E (T-T + R-R)		
	S	T-T	R-R	Total
1 + 1	60	60	60	120
1 + 2	26	26	52	78
1 + 3	4	4	12	16
1 + 4	2	2	8	10
2 + 1	6	12	6	18
2 + 2	2	4	4	8
4 + 1	1	4	1	5
Total	*101*	*112*	*143*	*255*

Table 4.7. Number of Type E (T-T + R-R) links

Types→ Links↓		Type F (T-R + R-T)		
	S	T-R	R-T	Total
1 + 1	8	8	8	16
1 + 2	1	1	2	3
2 + 1	4	8	4	12
2 + 2	1	2	2	4
Total	*14*	*19*	*16*	*35*

Table 4.8. Number of Type F (T-R + R-T) links

Types→ Links↓		Type G (T-T + T-R)		
	S	T-T	T-R	Total
1 + 1	9	9	9	18
1 + 2	1	1	2	3
2 + 1	1	2	1	3
Total	*11*	*12*	*12*	*24*

Table 4.9. Number of Type G (T-T + T-R) links

Types→		Type H (R-T + R-R)		
Links↓	*S*	R-T	R-R	Total
1 + 1	19	19	19	38
1 + 2	5	5	10	15
1 + 3	1	1	3	4
2 + 1	3	6	3	9
2 + 2	2	4	4	8
2 + 3	1	2	3	5
3 + 1	1	3	1	4
4 + 1	1	4	1	5
Total	*33*	*44*	*44*	*88*

Table 4.10. Number of Type H (R-T + R-R) links

Types→		Type I (T-T + R-T)		
Links↓	*S*	T-T	R-T	Total
1 + 1	16	16	16	32
1 + 2	9	9	18	27
1 + 3	4	4	12	16
2 + 1	1	2	1	3
Total	*34*	*31*	*47*	*78*

Table 4.11. Number of Type I (T-T + R-T) links

Types→		Type J (T-R + R-R)		
Links↓	S	T-R	R-R	Total
1 + 1	11	11	11	22
1 + 2	2	2	4	6
1 + 4	1	1	4	5
2 + 1	2	4	2	6
4 + 1	1	4	1	5
Total	*17*	*22*	*22*	*44*

Table 4.12. Number of Type J (T-R + R-R) links

There remain a few instances of more complicated combinations of types. They are listed in Table 4.13 below.

Types→ Links↓	S	Others T-T	R-T	R-R	T-R	Total
1 + 1 + 1 + 0	4	4	4	4	0	12
1 + 2 + 1 + 0	1	1	2	1	0	4
1 + 1 + 2 + 0	3	3	3	6	0	12
1 + 1 + 3 + 0	1	1	1	3	0	5
1 + 2 + 2 + 0	1	1	2	2	0	5
1 + 1 + 0 + 1	1	1	1	0	1	3
1 + 0 + 1 + 1	2	2	0	2	2	6
1 + 0 + 2 + 1	2	2	0	4	2	8
0 + 1 + 1 + 1	3	0	3	3	3	9
0 + 1 + 1 + 2	1	0	1	1	2	4
Total	*19*	*15*	*17*	*26*	*10*	*68*

Table 4.13. Number of other types of links

The total number of links in the sample text amounts to 1,321 (See Appendix 5). However, at this stage, the links in the Theme and Rheme areas needed to be counted separately, so that the number of links and the number of lexical items in the Theme and Rheme areas could be compared. Links across Theme and Rheme areas were not considered. The results of this counting are presented in Table 4.14 below.

Types	A	C	E	G	H	I	J	Other	Total
T-T	200		111	12		31		8	362
R-R		277	142		40		21	14	494

Table 4.14. Number of links within the Theme and Rheme areas

Table 4.14 shows that there are 362 links within the Theme area and 494 links within the Rheme area, apart from links across Theme and Rheme areas,

The absolute number of links in Theme is smaller that that in Rheme, but this is a direct consequence of there being fewer words in Theme than in Rheme. Proportionally, however, the lexical items in Theme form more links between

sentences than those in Rheme. This will be shown by a simple calculation. If calculated on the basis of both grammatical and lexical items (see Table 4.1 above), an item on average forms 0.85 links in the Theme area (i.e. 362 links divided by 426 items), as opposed to only 0.66 links in the Rheme area (494/745). That is, there is a 29% greater likelihood that an item will form links in the Theme area than in the Rheme area. If calculated on the basis of lexical items, a lexical item on average forms 1.53 links in the Theme area (362/236), as opposed to only 1.22 links in the Rheme area (494/404). That is, there is a 25% greater likelihood that a lexical item will form links in the Theme area than in the Rheme area. If calculated on the basis of the link-forming lexical items, an item on average forms 2.51 links in the Theme area (362/144), whilst a similar item forms only 1.90 links in the Rheme area (494/259). That is, there is a 32% greater likelihood that the link-forming lexical item will form links in the Theme area than in the Rheme area. Table 4.15 represents the results of the above calculation.

	Theme	*Rheme*	*(Th-Rh)/Rh*
Grammatical & Lexical tokens	0.85	0.66	0.29
Lexical tokens	1.53	1.22	0.25
Lexical tokens that form links	2.51	1.90	0.32

Table 4.15. Link/item ratios in the Theme and Rheme of the sample text

Intuitively, the differences between link/item ratios in Theme and Rheme seem to be significant. However, this needs to be tested by a statistical calculation. In order to find out whether this result is statistically reliable, a Chi-square test of

the above data was carried out. The Chi-square test is computed with Yate's correction factor (see Hatch & Farhady 1982: 171). The formula is as follows:

$$\chi^2 = \frac{N\left(\left[ad - bc\right] - \frac{N}{2}\right)^2}{(a + b)(c + d)(a + c)(b + d)}$$

where a = lexical items in Theme, b = lexical items in Rheme, c = links in Theme, d = links in Rheme, and N = the total. The result of the Chi-square test at the degree of freedom (d.f.) of 3 is 5.17, greater than the critical values of χ^2 of 3.84 (P = 0.05, d.f = 3). This result suggests that the difference in the distribution of links between the Theme and Rheme areas is statistically significant.

At this point, it may be concluded that lexical items in Theme do create proportionally more links than those in Rheme. In other words, Hypothesis 4.1 is supported by the analysis.

4.7.2 Testing Hypothesis 4.2

Hypothesis 4.2 is that 'proportionally, there are more R-T links than T-R links between sentences in a text', which concerns links across the Theme-Rheme areas. To test this hypothesis, I needed to count the different types of links that run across the Theme and Rheme areas. Data in Table 4.16 and 4.17 are based

on Tables 4.7 to 4.13 in the last section. From Table 4.16 and 4.17 we can see the distribution of links in the different directions.

Types	*B*	*F*	*H*	*I*	*Other*	*Total*
Links	190	16	44	47	17	314

Table 4.16. Number of R-T links

Types	*D*	*F*	*G*	*J*	*Other*	*Total*
Links	78	19	12	20	10	139

Table 4.17. Number of T-R links

Given that the links between Theme and Rheme, whether they are R-T links or T-R links, are all created by the same number of lexical items in the Theme and Rheme areas, it is obvious from Tables 4.16 and 4.17 that there are far more R-T links than T-R links. This phenomenon suggests that there is a much greater likelihood that the lexical item in Theme will form links with items in the Rheme of the preceding text, whereas the possibility of a lexical item in the Rheme forming links with items in the Theme of the preceding text is comparatively slight. This supports the argument that Theme is more closely associated with given information , which is recoverable from the preceding text, whereas Rheme is more closely associated with new information, which may not be repetition of information in the preceding text.

Although the difference is visible in the raw numbers, a chi-square test is still necessary for the claim to be established. The resulting P value of the chi-square test is 0.035, smaller than the commonly accepted value of 0.05. This means that there is only 3.5% of probability that the difference between the

quantities of R-T links and T-R links is by chance. In other words, it means that the difference between T-R links and R-T links is statistically significant.

It may be concluded here that across the Theme-Rheme boundary the lexical items form more links from Rheme to Theme than in the reverse direction. In other words, Hypothesis 4.2 is supported by the analysis.

Based on the discovery that there are proportionally more links in the Theme area than in the Rheme area, and there are more links from Rheme to Theme than from Theme to Rheme, a question that naturally follows would be whether the links in Theme have more linking power than those in Rheme. This question will be answered in the succeeding sections.

4.7.3 Testing Hypothesis 4.3

Hypothesis 4.3 is that 'T-T links or R-T links will create more coherence than R-R links or T-R links, where coherence is measured in terms of readers' judgements about the text'. That is to say, lexical links of Types A and B as presented in Figure 4.1 of Section 4.2 may create a closer semantic relationship between the two sentences concerned than Types C and D in Figure 4.2.

An ideal way of testing this hypothesis seems to be to carry out informant tests to discover native English speakers' intuitions about the coherence of all the sentence pairs with different types of links. But this is not practical, owing to

the complexity of the network of cohesive links and the enormous number of sentence pairs to be examined (see the statistics in Section 4.7.1). In addition, there were two other major obstacles to using all the sentence pairs in the informant test. First, as there are multiple relationships between the sentences, a sentence may be paired with a large number of other sentences in the same text, and too many repetitions of the same sentences would lead the reader to imagine a context which would otherwise be non-existent, thus distorting the results. Secondly, the reader might get too bored to give reliable judgements after reading so many repeated sentences (cf. Leech & Candlin 1986).

Therefore, I carried out a minimum informant test to examine only two of the basic types of links: Type A (T-T links), presumably the strongest of the basic types of lexical link, and Type D (T-R links), the presumably weakest type of the basic links. For the purpose of my analysis, I limited my informant questionnaire to eight pairs of sentences containing only two links each, which are below the threshold set by Hoey (1991a: 36) for a pair of sentences to bond. The aim was to test how many of the sentence pairs which according to Hoey were not bonded could be regard as coherent and how the different types of links influenced the informants' judgement. Therefore, those sentence pairs which have three links or more, having met the requirement of the minimum number of three links to make them bond, were not analysed in this connection.

For each sentence pair, the informants were required to decide whether it was very coherent, slightly coherent, slightly incoherent, or incoherent. In addition,

the informants were invited to give reasons for their decision wherever possible.

The informants were expected to answer the questions independently. For this purpose, two similar questionnaires were designed. The two questionnaires contain almost identical sets of eight sentence pairs, with four Type A linked pairs and four Type D linked pairs each. The sentences were taken from ten different texts at random. Because the 'Planet X' text was of particular interest, two pairs were selected from this text to represent both the two types of links; one (with Type A links) was included in Questionnaire A, the other (with Type D links) in Questionnaire B. The sentence pairs were arranged in such an order that in Questionnaire A, a sentence pair with Type A links was followed by a sentence pair with Type D links then followed by a pair with Type A links and then followed by a pair with Type D links again, and so on. In Questionnaire B, the order was reversed, starting with a sentence pair of Type D links. All this arrangement was intended to prevent discussions among the informants and to ensure that the informants would answer the questionnaire independently, so as to ensure the reliability of the results. For the sample questionnaires, see Appendices 6 and 7.

Two informant tests were carried out, at first on 14 August 1996 with a group of 10 native English teacher trainees and later on 19 September 1996 with 6 native English teacher trainees. Being native English speakers preparing to be teachers of English as a foreign language, the informants had some knowledge

of basic linguistic theories as well as a good command of standard English. It was assumed that they would be capable of identifying meaning relations between the sentence pairs and of describing their intuitions unambiguously. Indeed they read every pair of sentences very carefully before they made their decisions. All of them wrote down comments to explain why they regarded certain pairs as more coherent than others. The results of the informant tests are represented in Table 4.18 below:

	1st informant test (S = 10; N = 80)		2nd informant test (S = 6; N = 48)	
	Type A	Type D	Type A	Type D
Very coherent	18	8	12	3
Slightly coherent	10	13	8	3
Slightly incoherent	3	12	1	13
Incoherent	4	12	3	5
Total	*35*	*45*	*24*	*24*

Table 4.18. Data collected from the informant tests

In Table 4.18, the letter 'S' stands for the number of subjects, i.e., in the first informant test there were 10 informants and in the second informant test there were 6 informants. The letter 'N' stands for the number of items answered, i.e., in the first informant test 80 pairs of sentences were commented on and in the second informant test 48 pairs of sentences were commented on. Note that in the first informant test there were 10 more pairs of sentences with Type D links than those with Type A links.

In their comments, the informants showed that they used real world knowledge to establish links between the sentence pairs ('Recent news'; 'The topical content meant that I knew (a) was about abortion before reading (b)...'). Many

of them regarded coherence as continuation of topic ('Both sentences centre on the same point about family background'; 'Text follows on about planet'). Most of them noticed the relation between lexical repetition and topics ('"Dr Harrington" links the two sentences'; 'Because the "Call" is mentioned in (a) and expanded on in (b)...').

It was found very interesting that the informants tended to notice the lexical repetition if it occurred in the Theme area, whereas they tended, if the repetition occurred in the Rheme area, either to neglect it ('There is no obvious link between the sentences'; 'The content of each sentence is totally different — no similar words or meanings'), or to regard the repetition as 'unrelated' ('Totally unrelated'; 'Both mention "house prices", but do not correlate clearly in any other sense'). Nevertheless, they were very cautious not to categorise such sentence pairs as 'incoherent', still trying to establish links between them. Perhaps adjacency itself suggests to the reader a close relationship between the sentences, whatever the nature of links between them, since readers normally endeavour to establish links between adjacent sentences, even if there are no explicit lexical links. This may probably explain why sentence pairs with Type D links, though topically unconnected, were not given a very low grading on the coherence scale.

Summarising the two informant tests, it is clear that an overwhelming proportion of sentence pairs with Type A links were felt to be either 'very coherent'(50.8%) or 'slightly coherent'(30.5%). Only a small percentage were

regarded as either 'slightly incoherent'(6.8%) or 'incoherent'(11.9%). On the other hand, with Type D links, the grading is tilted to 'slightly incoherent' (36.2%) and 'incoherent' (24.6%). See Table 4.19 below.

	Type A (%)	*Type D (%)*
Very coherent	50.8	15.9
Slightly coherent	30.5	23.2
Slightly incoherent	6.8	36.2
Incoherent	11.9	24.6
Total	*100.0*	*100.0*

Table 4.19. Summary of the results of the informant tests

Figure 4.4 shows that the informants had a strong tendency to find coherence between sentence pairs related by Type A links, while they were cautious in discarding sentence pairs related by Type D links as incoherent.

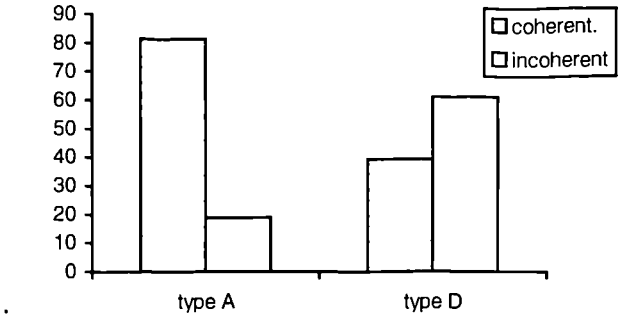


Figure 4.4. Comparison of the results of the informant test

Again a chi-square test was carried out to check whether the results are statistically reliable. With Type A links, the result of the chi-square test is 28.39 at the degree of freedom (d.f.) of 3. Compared with the critical value of 7.81 at the probability level of 0.05, this result indicates that there is a statistically significant tendency for Type A links to create a close relationship

between sentences. On the other hand, the result with Type D links is 5.84, which suggests that Type D links do not have significant influence on readers' judgements regarding coherence between sentences.

A tentative conclusion may be drawn here that sentence pairs with Type A links are, on the whole, felt to be more coherent than those with Type D links, in spite of the variation of opinion among individual informants. This may be interpreted as meaning that lexical links in the Theme area create a closer relationship between the sentences than those in the Rheme area. There seems to be a significant difference of linking power between lexical links in the Theme and those in the Rheme. The Theme area seems to be more important than the Rheme area when it comes to influencing readers' intuitions regarding coherence between two sentences. In other words, Hypothesis 4.3 is supported by the analysis.

4.7.4 Testing Hypothesis 4.4

Hypothesis 4.4 is that 'in the same text a series of sentences with T-T links tend to be closely related in meaning.' Before this hypothesis could be tested, it will be helpful to review some arguments about the unmarked conflation between Theme and Subject.

Davies (1988) argues that Subject is 'an obligatory element in Theme'; '...the repeated occurrence or recurrence of the same topical element or a related

topical element as Subject is seen not only to specify topic,... but also to be the primary means by which the continuity of coherent discourse is achieved'(1988: 3).

This argument may be understood as having two significant points: firstly, that Theme normally contains grammatical subject, and secondly, that recurrence of the same element in Theme specifies topic and helps achieve continuity of discourse. In other words, in a series of sentences, the recurrence of the same or related lexical items in Theme may help to achieve a coherent text on a specific topic.

I therefore selected sentences with the lexical item 'Harrington' functioning as the topical element in the Theme to make a sub-text concerning this person. These sentences are: Sentences 4, 9, 12, 41, 48 and 53. In addition to the lexical item 'Harrington' in the Theme to link these sentences, Sentences 4 and 9, 9 and 12 are also linked by the lexical item 'planet(s)' in the Rheme. Sentences 12 and 41 are linked by 'began/begun' in the Rheme.

The number and type of links between the sentences are shown in Figure 4.5 below.

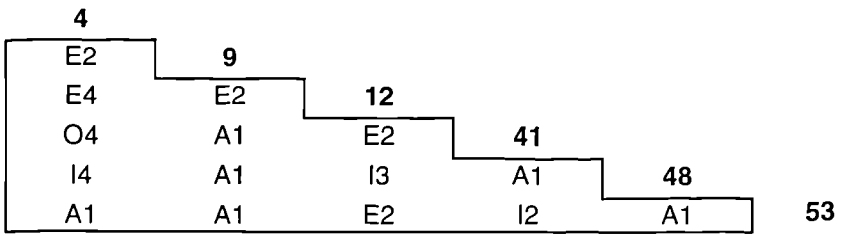


Figure 4.5. Types of links between the six sentences with 'Harrington' in Theme

In Figure 4.5, the numerals in bold represent sentence numbers, the capital letters represent linkage types ('O' stands for 'other type of links', cf. Table 4.13), and the numerals immediately following the capital letters stand for the number of links between the sentences marked on the relevant row and column. To read Figure 4.5, first start from a numeral in bold and move down vertically, then start from another numeral in bold and move towards the left horizontally, the cell where two bold face numbers meet shows the type and number of links between the two sentences thereby presented. For example, sentence pairs 4 and 9, 9 and 12, and 12 and 41 are all in Type E relationships, each having only 2 links between them. Sentences 41 and 48, and 48 and 53 are in a Type A relationship, each having only one link between them. Therefore, by Hoey's (1991a) criteria, which require a minimum of three links, none of these pairs of sentences are bonded.

Although Sentence 4 has four links with Sentences 9, 12 and 41 respectively, and Sentence 12 has three links with Sentence 48, the intrusion of sentences with fewer than three links would presumably, according to Hoey (1991a), reduce the coherence of this series of sentences. However, the hypothesis

examined at the beginning of this chapter is that the link between Themes, especially when the links in the Themes of the series of sentences are created by one and the same lexical item, should be strong enough to enhance the feeling of coherence and create a coherent sub-text.

Now take a look at this series of sentences:

(4) An astronomer, Robert **Harrington** of the US Naval Observatory in Washington, has begun a new search for '*Planet* X'.

(9) Dr **Harrington** and other sceptics say Pluto is too small to explain the orbits of the *planets* in the outer regions of the solar system, such as Uranus and Neptune.

(12) Dr **Harrington** is continuing a tradition of planet-searching which *began* thousands of years ago, when ancient astronomers discovered the five *planets* visible to the naked eye; Jupiter, for instance, is the brilliant object high up in the southern sky, and Venus can be easily seen in the east at sunrise.

(41) Dr **Harrington** has now *begun* work on two fronts, running new computer calculations in Washington and making fresh observations in New Zealand.

(48) Dr **Harrington** says the most remarkable feature predicted for Planet X is that its orbit is tilted 300 degrees away from the ecliptic, the main plane of the solar system, where all previous searches have concentrated.

(53) Using a blink comparator, a device that compares two photographs, Dr **Harrington** hopes to locate any faint object that has moved during the interval between the two pictures.

This sub-text sounds a little monotonous, because the repetition of the same item in the same position in adjacent sentences lacks variation, which is not recommended rhetorically. However, the sentences are closely related by the

same topical element, and they answer relevant questions the readers might have asked.

Hoey says, 'monologues may be projected into dialogue and such projection may clarify the monologue's organisation'(1983: 30). In order to clarify the relations between the sentences, we can recreate a dialogue between the writer and the reader by inserting questions. In this way, the sub-text may read more coherently:

(4) An astronomer, Robert **Harrington** of the US Naval Observatory in Washington, has begun a new search for 'Planet X'.

Why does he search for Planet X?

(9) Dr **Harrington** and other sceptics say Pluto is too small to explain the orbits of the planets in the outer regions of the solar system, such as Uranus and Neptune.

What are his guiding principles in the search?

(12) Dr **Harrington** is continuing a tradition of planet-searching which began thousands of years ago, when ancient astronomers discovered the five planets visible to the naked eye; Jupiter, for instance, is the brilliant object high up in the southern sky, and Venus can be easily seen in the east at sunrise.

How does Dr Harrington actually search for Planet X?

(41) Dr **Harrington** has now begun work on two fronts, running new computer calculations in Washington and making fresh observations in New Zealand.

What features does he expect to find of Planet X,?

(48) Dr **Harrington** says the most remarkable feature predicted for Planet X is that its orbit is tilted 300 degrees away from the ecliptic, the main plane of the solar system, where all previous searches have concentrated.

What are his techniques in searching for Planet X?

(53) Using a blink comparator, a device that compares two photographs, Dr **Harrington** hopes to locate any faint object that has moved during the interval between the two pictures.

This series of sentences could be regarded as a summary of the sub-topic 'Dr Harrington' in the original text. It answers questions concerning his principles, aims, techniques in research, as well as the theory and rationale behind his actions.

Another sub-text takes the lexical item 'astronomer' in the Themes of the component sentences. They are Sentences 3, 6, 10, 13, 18, 19, and 23. The number and type of links between the sentences are shown in Figure 4.6 below.

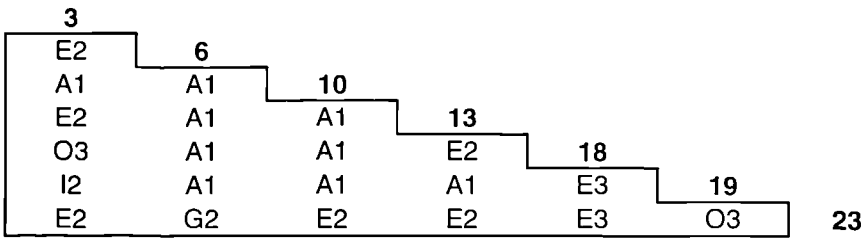


Figure 4.6. Types of links between the six sentences with 'astronomers' in Theme

Except Sentences 18 and 19, 19 and 23, which have respectively reached the threshold of three links for them to be bonded, the other sentences have only one or two links between them, and according to Hoey (1991a) should not be regarded as bonded. In addition, the lexical item 'astronomers' has different referents in the real world, which creates a looser relationship than a lexical item with the same referent in the real world. However, because of the hypothesised strength of links in the Themes, this series of sentences should

serve as a summary of what the group of people known as 'astronomers' have done and/or are doing.

The sentences are as follows:

(3) Some **astronomers** have continued to suspect there may be a tenth planet lurking even further away which has somehow escaped detection.

(6) The young **astronomer** Mr Tombaugh had detected a sure sign of an object orbiting the Sun by comparing two photographs showing that a speck of light had shifted position against the stars.

(10) Most **astronomers** nowadays work on such exotic problems as the origin of our universe or the properties of black holes, but Dr Harrington is cast in a traditional mould.

(13) In 1781, the British **astronomer**, Sir William Herschel found a new planet, Uranus, setting off feverish planet-hunting.

(18) This century, Percival Lowell, an American **astronomer**, who achieved notoriety for suggesting Mars had canals and life, urged a search for a further planet using wide-angle cameras.

(19) For 20 years, **astronomers** at Lowell Observatory searched the skies.

(23) Still a professional **astronomer** at 83, Mr Tombaugh said last week: 'The moving speck of light was much fainter than we expected, and I carried on searching, just in case there was another one, for 14 years, until May 1943.

It could not be claimed that this series of sentences is coherent in the normal sense. One possible cause may perhaps be the presence of the adjuncts of time. The jumbled sequence of time confuses the readers. However, it could be argued that these sentences are relevant around a certain topic, as they all deal with a category of professionals. Further, if we rearrange the sequence of sentences in the order of a more natural time sequence, the sub-text reads much

more coherently. For ease of reference, adjuncts to time in the following sentences are printed in italics.

(13) *In 1781*, the British **astronomer**, Sir William Herschel found a new planet, Uranus, setting off feverish planet-hunting.

(18) *This century*, Percival Lowell, an American **astronomer**, who achieved notoriety for suggesting Mars had canals and life, urged a search for a further planet using wide-angle cameras.

(19) *For 20 years*, **astronomers** at Lowell Observatory searched the skies.

(6) The *young* **astronomer** Mr Tombaugh had detected a sure sign of an object orbiting the Sun by comparing two photographs showing that a speck of light had shifted position against the stars.

(23) Still a professional **astronomer** *at 83*, Mr Tombaugh said last week: 'The moving speck of light was much fainter than we expected, and I carried on searching, just in case there was another one, *for 14 years, until May 1943.*

(3) Some **astronomers** *have continued* to suspect there may be a tenth planet lurking even further away which has somehow escaped detection.

(10) Most **astronomers** *nowadays* work on such exotic problems as the origin of our universe or the properties of black holes, but Dr Harrington is cast in a traditional mould.

However, the re-sequencing of sentences requires further study which goes beyond the scope of the present thesis. This point will not be discussed further here. For more details about re-sequencing sentences, see Hoey (1983: 3 ff.).

Hypothesis 4.4 seems to be supported to a certain extent by the above analysis. Although the sentences with the same lexical items in the Theme may not readily form sub-texts, they do revolve around a closely related topic, and still provide some clues as to the sub-topics which the text may accommodate.

4.8 Conclusion

Up to this point, several provisional conclusions may be drawn from the analyses. In the first place, the study shows that there is evidence to support the hypothesis that lexical items tend to form more links in the Theme area than in the Rheme area. Secondly, the probability of lexical items in Theme forming links with lexical items in the Rheme of the preceding text is much greater than the probability of lexical items in Rheme forming links with those in Theme of the preceding text. Thirdly, there is evidence to support the hypothesis that the repetition of lexical items in the Theme create a closer relationship between the sentences concerned than the repetition of lexical items in the Rheme. The semantic closeness of sentences is related with the position of the lexical links in the sentences. Further, it is observed that a series of sentences which have repeated lexical items between their Themes are topically close and may be picked out to form subtexts concerning certain topics; in other words, Hypothesis 4.3 is supported to a certain extent by the analysis. However, this is only a preliminary investigation in this area. Much more has to be done. It is hoped that if the results could be duplicated with a wider variety of text types and more informant tests, so as to yield more convincing evidence, the features of lexical links as reported in this chapter may be used to improve the quality of automatic abridgement in the future.

From the analysis, it may be assumed that a normal text would contain sentences with a combination of T-T links and R-T links. Therefore, it may be

further assumed that a collection of sentences with a combination of T-T links and R-T links abstracted from a source text will produce a readable sub-text. This assumption draws on Hoey's (1991a) method of collecting central sentences with an above average number of links and bonds between them, but it also takes into consideration the hypothesised features of Theme and Rheme links. However, this assumption is largely intuitive and not readily testable, though the sample text itself manifests this feature to a certain extent.

This chapter, in answer to the first general research question raised in Section 1.5 of Chapter 1, has confirmed that there is indeed a close relationship between lexical patterning and the Theme-Rheme system in the text, and it is manifested in the form of a concentration of lexical links in the Theme area of the text. The four hypotheses intended to test the assumed relationships have all been supported by the analysis in this chapter. The next chapter will move on to answer the second general research question, examining whether there is a relationship between key words and the Theme-Rheme system, and if there is, how it is manifested in the text.

Chapter 5. Key Words in Theme and Rheme

5.1 Introduction to this chapter

Chapter 4 answered the first general research question raised in Section 1.5 of Chapter 1, concerning the relationship between lexical patterning and the Theme-Rheme system and its manifestation in the text. It investigated the nature of lexical patterning in the Theme and Rheme areas and confirmed that there is a close relationship between lexical patterning and the Theme-Rheme system. This relationship is manifested in the form of a concentration of lexical links in the Theme area of the text.

The present chapter moves on to answer the second general research question, 'Is there a relationship between key words and the Theme-Rheme system and, if there is, how is it manifested in the text?' To answer this question, it is necessary to investigate the distribution of key words and keyness in the Theme-Rheme areas of the text. Although the notions of key words and keyness were very briefly introduced in Section 1.5 of Chapter 1, these notions

are so essential to our analysis in this chapter that they merit a more detailed description in the next subsection.

5.1.1 The notions of key words and keyness

Key words are popularly regarded as 'important' words. For example, Carter (1995) uses the term to mean words essential in literary study. In the present thesis, the concept of key words is slightly different from the popular sense. The term 'key words' used in this chapter refers to words which have 'special significance for the meaning of the particular text' (Halliday 1994: 310). They are 'words or phrases which in some sense characterise, represent, typify or identify a text' (Collins and Scott 1997: 3).

Specifically, here, key words are those words generated from a computer program 'KeyWords', which is a component part of the software package 'WordSmith Tools' by Scott (1996). The 'KeyWords' program is designed to extract key words from naturally occurring texts. This program computes key words on the principle of comparing frequency of occurrence of certain words in the sample texts with frequency of occurrence of the same words in a reference corpus. It is claimed to be effective in identifying intratextual relations of lexical items, especially with texts above a minimum length, e.g. 300 words, and intertextual relations with a great number of texts.

Scott (1997a) states that a key word is recognised by identifying a word which occurs with unusual frequency in a given text. He emphasises, 'this does not mean high frequency, but unusual frequency, by comparison with a reference corpus of some kind' (1997a: 236). In the on-line manual to the computer program of WordSmith Tools, Scott (1996) explains that key words are identified on a mechanical basis by comparing patterns of frequency. A word is said to be 'key' if 'its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p-value specified by the user.' So it is possible to have positive key words which are unusually frequent, and negative key words which are unusually infrequent, when compared with a reference corpus.

The key word is a comparative concept. In a sense, it may be regarded as a 'text specific' word, specifying the uniqueness of a particular text in comparison with other texts, or even with the language in general. A word may be a key word in one specific text, but not in another. Further, even in one text, not all key words are 'key' to the same extent. The differences in extent can be indicated by different degrees of 'keyness', a measure of the significance of a certain key word in the text from which the key word is derived. A key word has a higher keyness if its frequency of occurrence in the text is extremely unusual; it has a lower keyness if its frequency of occurrence is less unusual.

For example, suppose the word 'apple' occurs 10 times in a sample text of 1,000 words, which means that the percentage of its occurrence amounts to 1% of all words in the sample text. Then suppose the same word occurs 1,000 times in a one million word reference corpus, equivalent to 0.1%. The difference between the frequency of occurrence in the sample text and that in the reference corpus is 1% divided by 0.1%, which equals 10. This indicates that the word 'apple' in the sample text is 10 times as frequent as it is in the reference corpus. In comparison, if the word 'pear', which also occurs 10 times in the same sample text of 1,000 words, occurs 100 times in the one million words reference corpus, approximating 0.01%, the difference between the frequency of occurrence of 'pear' in the sample text and that in the reference corpus is 1% divided by 0.01%, equal to 100. This indicates that the word 'pear' in the sample text is 100 times as frequent as it is in the reference corpus. If only one word could be selected as key word of the sample text, then 'apple' must be excluded, or alternatively, if both words are included as key words, the keyness of 'apple' is obviously much lower than that of 'pear'. This is because the difference between the frequency of occurrence in the sample text and that in the reference corpus for the word 'pear' is 10 times greater than the difference between the frequency of occurrence in the same two corpora of the word 'apple'.

The above example only gives a very rough idea of the notions of key word and keyness. The actual computation is more complex than exemplified. But this is beyond the scope of the present study and will not be pursued further.

The basic assumption for the present chapter is that the key words in a text, viewed together, may be indicative of the main topic of that text, or what that text is about. The premise of this chapter is that, if both key words and Theme have the function of indicating the 'aboutness' of the text, then there should be some kind of correlation between the two. That is, if the text is concerned with some topic, that topic will be reflected both by the patterns of key words and by the thematic choices.

5.2 Hypotheses

As a starting point, we shall consider some features shared by key words and Theme-Rheme structure. In the first place, we shall recall the notion of Theme, which is reviewed in Chapter 2 of the present thesis. Basically, Theme is a category on the clause level. It is 'what the clause is about' (Halliday 1985a: 38). Furthermore, all the clause Themes taken together in a text realise 'discourse Theme_{M}', which is the prioritised meaning of the text (Berry 1996). On the other hand, key words are also claimed to reflect the 'aboutness' of the text (Scott 1997a). Therefore, we can hypothesise that, in a text, key words will concentrate in the Theme area, rather than in the Rheme area. For this hypothesis to be testable, it may be formulated as follows,

Hypothesis 5.1. A higher proportion of key words will occur in Theme than in Rheme.

Because key words are actually text specific words, it may be assumed that generally a key word that is more central to what the text is about will appear higher on the scale of 'keyness' than one that is less central to what the text is about. From this assumption, it may be further assumed that words in the Theme area will have a higher 'keyness' than those in the Rheme area, since what the text is about is reflected in the discourse $Theme_M$, which in turn is a reflection of the cumulative force of all the clause Themes taken together. This leads us to Hypothesis 5.2.

Hypothesis 5.2. There will be a higher overall 'keyness' in Theme than in Rheme.

Moreover, comparing the notion of 'key words' with that of lexical link, if a lexical item within the Theme area is more likely to be reiterated than one in the Rheme area, and thus more lexical links are formed in the Theme than in the Rheme (cf. Chapter 4 of the present thesis), then this should also be true of the key words, because key words are by definition closer to the 'aboutness' of a text than ordinary lexical words. So, we may assume that a key word in Theme will form more links than a key word in Rheme. Then Hypothesis 5.3 for the present chapter is that

Hypothesis 5.3. On average, a key word in Theme will be reiterated more than a key word in Rheme, thus forming more links in Theme than in Rheme.

Before we set out to test these hypotheses, it is necessary to make some remarks about the computer programs which are used in the analysis to generate key words.

5.3 The computer programs used in the analysis

'WordSmith Tools' designed by Scott (1996) is the suite of computer programs used in the present study. Of its several component tools, the 'WordList' and the 'KeyWords' tools are essential for the analysis of key word patterns in the sample texts. The 'WordList' program is used to make word lists of the sample texts. The resulting lists can be displayed either alphabetically or in order of frequency. This tool has a 'Consistency' function which can compare different word lists and display the results in 'simple' or 'detailed' format.

The 'KeyWords' program is designed on the assumption that key words are those whose frequency is unusually high or low in comparison with some norm. For our purpose, we focus on the high frequency key words. A word list saved on a corpus of 95 million words from the British newspaper 'The Guardian' published between 1991-1994 is provided by the author of 'WordSmith Tools' to serve the purpose of a reference 'norm'. The mechanics of the 'KeyWords' program requires two wordlists: one from the text under consideration, the other from the reference corpus. As indicated previously

(Section 5.1.1), the key words are those whose frequency in the sample text is unusual when compared with their frequency in the reference corpus. More specifically, a word is included in the key word list when the statistical probability of its occurrence in the sample text and the reference corpus as computed by an appropriate procedure is smaller than or equal to a p-value of 0.000001. The default 'appropriate procedure' to obtain the p-value of the key words used to be a chi-square comparison, but in the latest version of the KeyWords software, Dunning's (1993) Log Likelihood is used. The 'keyness' is calculated by comparing the frequency of each word in the wordlist derived from the sample text with the frequency of the same word in the reference corpus.

5.3.1 Key words by human and by machine

To evaluate the performance of the KeyWords program, one approach would be to compare its output with the output of an independent device, either from another computer program which is recognised as reliable, or from human resources. At present, however, no other computer programs available have yet achieved the necessary status of complete reliability, so we have to turn to human resources, where key words selected by human subjects may serve as the basis for our evaluation.

József Andor (1989) reports on a very interesting experiment, in which he asks students to find key words (or, as he explains it, words which have dominance

over the occurrence of others) in a short sample text. Since the sample text is not very long, it is convenient to reprint it:

> A Thief in the Night
>
> During the recent transit strike, a young man was walking home from work through the park. It was late and he was alone. In the middle of his treck [sic] he saw someone approaching him on the path. There was, of course, a spasm of fear: He veered, the stranger veered. But since they both veered in the same direction, they bumped in passing.
>
> A few moments later the young man realized that this could hardly have been an accident, and felt for his wallet.
>
> It was gone.
>
> Anger triumphed and he turned, caught up with the pickpocket and demanded his wallet. The man surrendered it.
>
> When he got home, the first thing he saw was his wallet lying on the bed. There was no way of avoiding the truth: He had mugged somebody.
>
> (*New York*, September 1, 1980, reprinted from Charolles 1989: 11)

The key words identified by Andor's subjects (n = 80) are presented in Table 5.1 below, in the order of their first occurrence in the sample text.

word	*n*	*%*	*word*	*n*	*%*
thief	80	100.00	bump	57	71.25
night	80	100.00	realize	42	52.50
transit strike	12	15.00	wallet	80	100.00
park	80	100.00	pickpocket	80	100.00
late	63	78.75	demand	48	52.50
alone	46	57.50	surrender	61	76.25
path	80	100.00	mug	80	100.00
stranger	68	85.00			

Table 5.1. Key words identified by student subjects (after Andor 1989: 32)

Andor argues, 'Beyond any doubt, a thematically arranged lexical make-up can greatly contribute to the coherence of a text or discourse through the stimulating capacity of the elements'(1989: 34). He labels these words as 'text organisers', emphasising that 'the frame triggering capacity of particular lexical items that serve as organisers in a passage highly contributes to the conditions of coherence revealed by the interpreter...' (1989: 35). As he observes, the title

of the sample text contains two such organisers, and 'thus clearly reflects the core of the contents of the story' (1989: 34).

Andor's analysis sheds some light on our understanding of the coherence of text. The 'frame triggering capacity of particular lexical items' identified in his analysis plays a crucial role precisely in the activation of normal interpretation. A reader will be able to interpret the text more easily if it is connected with a frame in his mind. This agrees very well with common intuition.

Comparison of the output of Andor's test and that of the KeyWords program may serve to test the reliability of the KeyWords program, checking how it matches human intuition. However, it should be noted that human intuition may not be fully reliable. For example, when Collier (1998) asked his informants to select concordance lines from the COBUILD corpus to compare with the concordance lines selected by his software, he found not only that their selection did not match the selection of his software, but that the informants did not agree with each other among themselves. Their selection was close to a random selection. Likewise, when asking his 20 informants to manually abridge a newspaper text by selecting sentences from the text, Hoey (personal communication) also found that they selected sentences close to random selection. Nevertheless, when there is no other source for our comparison of the computer program, human intuition may serve as a useful reference.

The KeyWords program was originally designed to process texts of at least 300 words. But the sample text in consideration is composed of only 138 words, of 85 word types, in which lexical items are unlikely to recur. However, it is possible to adjust the frequency and p-value settings of the KeyWords program to generate a key word list of the short text, though the results may be less reliable than those with longer texts. Considering the extreme shortness of the sample text, the minimum frequency is set at 1, so that every word is a potential candidate for the key word list. The maximum p-value for the computation to be statistically acceptable is 0.05, but it should be set as low as possible. Another consideration is that the key word list should be as close to Andor's list in length as possible, so that it is conveniently compatible. Table 5.1 shows that Andor's list has 15 items, with 'transit strike' selected by Andor's students as one item. After several trials, the maximum p-value was set at 0.001.

Out of the 85 word types in the original text, 17 were selected in the automatic key word list, which equals 20% of the total. If counted by token, then 36 of the 138 tokens in the original text were selected in the key word list, which equals 26% of the total. The key word list generated by the 'KeyWords' program (frequency = 1, P = 0.001) is presented below in Table 5.2, in decreasing order of 'keyness'.

Word	*Freq.txt*	*%txt*	*Freq.ref*	*%ref*	*Keyness*	*P-value*
veered	3	2.17	159		50.8	0.000000
wallet	3	2.17	342		46.3	0.000000
treck	1	0.72	0		26.9	0.000000
realized	1	0.72	30		18.1	0.000021
pickpocket	1	0.72	32		17.9	0.000023
he	7	5.07	574,604	0.60	17.7	0.000025
man	3	2.17	55,702	0.06	15.9	0.000067
spasm	1	0.72	121		15.3	0.000092
was	7	5.07	701,712	0.74	15.3	0.000092
mugged	1	0.72	186		14.4	0.000145
saw	2	1.45	16,558	0.02	13.8	0.000207
bumped	1	0.72	255		13.8	0.000203
triumphed	1	0.72	321		13.3	0.000259
transit	1	0.72	535		12.3	0.000446
thief	1	0.72	531		12.3	0.000443
surrendered	1	0.72	634		12.0	0.000535
stranger	1	0.72	999		11.1	0.000872

Table 5.2. Key words generated by the computer program

It should be remembered that computer generated key word list and Andor's list of key words, or 'text organizers' as he calls it, are made on different working principles. The computer program works purely on the statistical principle, whereas the human subjects select key words from their interpretation of the meaning of the source text.

No doubt, it is unlikely that any two key word lists generated through different procedures would be identical. However, the comparison reveals some common features shared by them. For convenience of comparison, the two key word lists are represented below in alphabetical order side by side in Table 5.3, with those items that occur in both lists printed in bold.

Human	*Computer*
alone	—
bump	**bumped**
demand	—
—	he
late	—
—	man
mug	**mugged**
night	—
park	—
path	—
pickpocket	**pickpocket**
realize	**realized**
—	saw
—	spasm
stranger	**stranger**
surrender	**surrendered**
thief	**thief**
transit strike	**transit**
—	treck
—	triumphed
—	veered
wallet	**wallet**
—	was

Table 5.3. Key words selected by human subjects and the computer program

A quick comparison of the lists in Table 5.3 shows that there are nine words selected both by the KeyWords program and Andor's subjects. But what does this mean to us? To answer this question, we need to compare the lists from two perspectives. One is statistical; the other is semantic. A commonly used statistical test is to adopt two performance matrices of recall and precision (cf. Marcu 1997, Berber Sardinha 1997). The recall rate is obtained by dividing the number of key word types concurrently selected by the KeyWords program and Andor's student subjects by the number of key word types identified only by the student subjects. In this case, it is 9 divided by 15, which equals 60.00%. Likewise, the precision rate is obtained by dividing the number of key word

types identified both by the KeyWords program and the student subjects by the total number of key word types only recognised by the KeyWords program. This time, it is 9 divided by 17, resulting in 52.94%.

The recall and precision rates between the two word lists may be regarded as reasonably high, and it may be claimed that there is some correlation between these two word lists. To check the validity of this claim, a test of the correlation between the key words and a randomly generated word list was conducted. The Random Number Generation analysis tool in MS Excel was used to generate 17 numbers out of a total of 85, corresponding to the 17 types of key words out of the 85 word types from the sample text. Then the numbers were applied to the total word list in alphabetic order. The words picked up by the random procedure are as follows: 'a, an, bumped, caught, could, during, for, middle, mugged, on, path, saw, since, up, veered, was, when'.

Only five items in the randomly generated word list correlate with items in the computer generated list and only three items with those in Andor's list. The recall rates of the randomly generated word list against the KeyWords program generated word list and Andor's list are 29.41% and 20.00% respectively, whereas the precision rates are 29.41% and 17.65% respectively. This provides evidence that the KeyWords program does not select key words at random, but its selection is significantly close to the human selection of key words.

Now, we are in a position to examine the semantic contents of the two lists made through the different procedures.

Both lists select the nine words 'bump, mug, pickpocket, realize, stranger, surrender, thief, transit' and 'wallet'. These are indeed essential to the topic of the text. Of the nine key words, three refer to the participants: 'pickpocket, stranger, thief', four refer to the processes of the event 'bump, mug, realize, surrender', one refers to the object 'wallet', and one is used to describe the setting 'transit (strike)'. Intuitively, the reader can use these key words to build a framework of what is being talked about in the story. The item 'transit strike' may need a brief comment here. While the computer program regarded each set of letters bounded by spaces as a separate word, the human subjects were able to recognise multi-word items. So, strictly speaking, this item can only be regarded as an incomplete match. In the next paragraph we shall see another interesting phenomenon of the cumulative effect of collocates recognised by human subjects which no computer program is able to achieve at present.

The words in Andor's list which the KeyWords program fails to capture are 'night, park, late, alone, path, demand' and 'strike'. Based on our socio-cultural knowledge, we admit that most of these words are scene-setting items, contributing to the creation of a gloomy atmosphere. Moreover, it is their operation in combination that produces the gloomy effect, something which humans are sensitive to but no computer program can take into account. The computer program's failure to detect these words may be due to the high

frequency of these words in the reference corpus. In other words, the meanings carried by these words are not particularly unusual in the western culture, at least as reflected by the Guardian newspaper between 1991-1994, from which the 95-million-word reference corpus for the KeyWords program was made.

On the other hand, there are eight key words which do not occur in Andor's list. The word 'veer' is used in the sample text to describe the action of both the young man and the alleged thief. Instead of 'curve, depart, deviate, digress, diverge, swerve, turn, walk aside', etc., the word is used in the sample text several times with an implication of avoidance of something unpleasant. The word 'spasm' also carries the implication of something involuntary and unpleasant. The word 'triumph' normally denotes noteworthy or spectacular success, but in this sample text it is used metaphorically with the agent of 'anger', again something unpleasant. These words are typical of this text at least in stylistic terms, and it may be argued that they also contribute to the creation of the coherence of the text. 'He' and 'man' are interesting cases. The mechanics of the KeyWords program might have discarded such general words, since they have a very high frequency in the reference corpus. But they occur in the sample text in such a high frequency that they were included in the key word list by the computer program. It is worth noticing that 'he' is a pronoun and 'man' is a superordinate noun, both being used to refer to the 'young man' and the 'thief', the two main characters in the story. Naturally, they are referred to frequently. However, the human readers failed to select these words, perhaps because they did not regard the high frequency of these

words as anything unusual, given the assumption that a story must be about somebody.

Finally, there are some spurious cases. The word 'treck' was selected by KeyWords program because of its low frequency in the reference corpus. In fact, it does not occur at all in the 95-million-word reference corpus, nor can it be found in a desktop dictionary. So, it is suspected to be a misprint of the word 'trek'. When the KeyWords program was re-run on the sample text with the spelling of this word altered, the word 'trek' was still in the key word list, only with the keyness reduced from 26.9 to 12.4. The words 'saw' and 'was' may also be regarded as spurious cases; they are the effect of the extreme shortness of the sample text, which forced the program to lift its protective threshold of minimum requirement of frequency and p-value.

From this comparison it may be concluded that for our analysis of text organisation the KeyWords program is a reliable tool, which can make key word lists that not only reveal the content but also reflect the stylistic features of the text under analysis.

5.4 Group B Sample Texts

The sample texts analysed in this chapter are eleven science reports taken from the British newspaper 'The Independent on Sunday' between March 1995 and

June 1996. They are called 'Group B Sample Texts' in the present thesis. These texts are on average 1,517 words in length, with the longest being 2,122 words, and the shortest being 1,198 words. The eleven Group B Sample Texts with their date of publication and length in words and sentences are listed below in Table 5.4.

Texts	*Date*	*Words*	*Sentences*
B1	19/03/1995	1,559	75
B2	14/05/1995	1,382	68
B3	09/07/1995	1,470	74
B4	20/08/1995	1,203	61
B5	03/09/1995	2,122	78
B6	10/12/1995	1,563	85
B7	07/04/1996	1,198	64
B8	19/05/1996	1,400	69
B9	26/05/1996	1,206	63
B10	23/06/1996	1,770	84
B11	30/06/1996	1,815	90

Table 5.4. Date and length of Group B Sample Texts

Group B Sample Texts were selected mainly on the basis that the analysis in this chapter is a continuation of the research on the correlation of lexical patterns and thematic organisation in the written text reported in Chapter 4 of this thesis. If the results of analysis in Chapter 4 are reliable, and hypotheses raised in this chapter are logically connected with those in Chapter 4, then analyses using extended data should further reveal the features of text organisation. Therefore, it is considered plausible to keep the factors compatible: Group B Sample Texts used in this chapter should be of the same genre or text type and of similar length to Sample Text A used in Chapter 4. At this stage of research, it is not intended to claim that the assumed correlation between thematic organisation and patterns of lexis is universal to all genres or

text types. It is only hypothesised that the correlation will exist in the text type of the popular science report as found in British newspapers.

5.5 Method of analysis

I shall now report on my analysis of the eleven Group B Sample Texts for the purpose of testing the hypotheses listed at the beginning of this chapter. The approach to the analysis was basically the same as that in Chapter 4. The analysis was carried out in the following steps:

The first step was to mark out the Theme of each clause. The delimitation of Theme was similar to that described in Chapter 4.

In the second step, the sample texts were divided into Theme and Rheme to make separate corpora. All the words in the Theme of a sample text were extracted to make a separate Theme corpus; the remainder were kept to make a Rheme corpus. As a result, eleven Theme corpora, eleven Rheme corpora, plus eleven corpora of complete texts, totalling thirty-three corpora in all, were created out of the eleven Group B Sample Texts.

The number of words in the Themes and Rhemes of each sample text is presented in Table 5.5 below.

Texts	*Theme*		*Rheme*		*Overall*
	words	*%*	*words*	*%*	*words*
B1	554	35.54	1,005	64.46	1,559
B2	552	39.94	830	60.06	1,382
B3	507	34.49	963	65.51	1,470
B4	375	31.17	828	68.83	1,203
B5	796	37.51	1,326	62.49	2,122
B6	639	40.88	924	59.12	1,563
B7	436	36.39	762	63.61	1,198
B8	509	36.36	891	63.64	1,400
B9	474	39.30	732	60.70	1,206
B10	750	42.37	1,020	57.63	1,770
B11	640	35.26	1,175	64.74	1,815
Total	*6,232*	*37.34*	*10,456*	*62.66*	*16,688*

Table 5.5. Number of words in Theme and Rheme of Group B Sample Texts

In order to show the distribution of words across Theme and Rheme in the sample texts, I have included in Table 5.5 percentages of Theme and Rheme in the texts. Table 5.5 shows that words in Theme and Rheme are distributed roughly in the proportion of 1:2, either in the group as a whole or in individual texts. In other words, Rhemes are about twice as long as Themes in these sample texts.

The third step was to make word lists out of the separate Theme and Rheme corpora, using the WordList program from the WordSmith Tools program suite. To reduce the length of the word lists and make the lexical words prominent, a stopword list was used to eliminate functional words (grammatical items). This is only for the purpose of clearer comparison of the raw wordlists and key word lists later. For the making of key word lists, however, there is no need to use the stopword list, since most of the

grammatical items will be automatically excluded from the key word lists due to their very high frequency in the reference corpus.

First, three separate word lists were made out of each text on the basis of the corresponding Theme corpus, Rheme corpus and the corpus of the complete text. Then, using the 'Consistency (detailed)' function of the WordList program, the three word lists were placed side by side for comparison. In this way, the word profile of each sample text emerged. The 'Consistency (detailed)' word list of Sample Text B1 is represented below in Table 5.6, in order of frequency. For lack of space, I have truncated the word list by leaving out words with frequency of lower than three occurrences in the text, because those words would not be selected by the KeyWords program in the setting of minimum frequency of three occurrences. For fuller word lists of all the eleven Group B Sample Texts, see Appendix 9.

Words	*Theme*	*Rheme*	*Overall*
1. Intelligence	**7**	**9**	**16**
2. Apes	**5**	**5**	**10**
3. Gorillas	**4**	**3**	**7**
4. Ape	**4**	**2**	**6**
5. Human	**4**	**2**	**6**
6. Orangutans	**4**	**2**	**6**
7. Social	**3**	**6**	**9**
8. Chimps	**3**	**4**	**7**
9. Ancestors	**3**	**2**	**5**
10.Primates	**3**	**2**	**5**
11.Orangutan	**3**	**1**	**4**
12.Groups	**2**	**4**	**6**
13.Species	**2**	**4**	**6**
14.Humans	**2**	**3**	**5**
15.Live	**2**	**3**	**5**
16.Primate	**2**	**3**	**5**
17.Great	**2**	**2**	**4**
18.Living	**2**	**2**	**4**
19.Monkeys	**2**	**2**	**4**
20.Sheep	**2**	**2**	**4**
21.Borneo	**2**	**1**	**3**
22.Brain	**2**	**1**	**3**
23.Says	**1**	**5**	**6**
24.Time	**1**	**5**	**6**
25.Behaviour	**1**	**4**	**5**
26.Animals	**1**	**3**	**4**
27.See	**1**	**3**	**4**
28.Evolution	**1**	**2**	**3**
29.Food	**1**	**2**	**3**
30.Forest	**1**	**2**	**3**
31.Imitation	**1**	**2**	**3**
32.Mental	**1**	**2**	**3**
33.Particular	**1**	**2**	**3**
34.Things	**1**	**2**	**3**
35.Byrne	*9*	*0*	*9*
36.Department	*3*	*0*	*3*
37.University	*3*	*0*	*3*
38.Complex	0	6	6
39.Understanding	0	5	5
40.Believes	0	4	4
41.Requires	0	4	4
42.Way	0	4	4
43.Evolved	0	3	3
44.Range	0	3	3
45.World	0	3	3

Table 5.6. Word list of Sample Text B1 (frequency => 3)

Table 5.6 represents a list of candidates for the key words of Sample Text B1. This list is divided into three distinct sectors. Sector 1, which is printed in bold,

contains words occurring both in Theme and Rheme; Sector 2, which is printed in italics, contains words that are found only in Theme; and Sector 3, in regular printing, contains words only in Rheme. A glance of the three sectors reveals that words in Sector 1, especially those on the top of this sector, are mostly nominals which give a good hint of what the text is about. Words in Sector 2 mostly refer to researchers, institutions, and subjects of research which are also related to the topics of the text, but are slightly less central, whereas words in Sector 3 contain a comparatively large portion of verbs and adjectives that give little specific hint of what the text is talking about. Intuitively, words in Sectors 1 and 2 will contain more key words than Sector 3, since key words are claimed to be indicative of what the text is about.

Therefore, in the next step, the third step, a key word list was produced. One of its uses is to check the above intuition. The key word list was made by comparing the word list of Sample Text B1 with the 95-million-word reference corpus, using the KeyWords program of WordSmith Tools, as described in Section 5.1.1. Using the default setting of the program, only words with a minimum frequency of three occurrences in the sample text were selected. This was to ensure that rare words, such as proper nouns, would not be categorised as key words, unless adequate frequency in the sample text indicated that they were really key to the central topic. The key word list of Sample Text B1 is presented below in Table 5.7, with statistics including the frequency of occurrence of the key words and their percentage on the basis of the sample text, the frequency and percentage of the key words on the basis of the

reference corpus, and their keyness and p-value. The key words are listed in descending order of their frequency in the sample text.

Key words	*Freq.B1*	*%B1*	*Freq.ref*	*%ref*	*Keyness*	*P-value*
1. Intelligence	16	1.03	6,718		127.7	0.000000
2. Apes	10	0.64	145		146.3	0.000000
3. Byrne	9	0.58	602		104.6	0.000000
4. Social	9	0.58	31,973	0.03	34.3	0.000000
5. Gorillas	7	0.45	85		104.8	0.000000
6. Chimps	7	0.45	101		102.4	0.000000
7. Orangutans	6	0.38	3		120.8	0.000000
8. Ape	6	0.38	205		77.7	0.000000
9. Species	6	0.38	3,322		44.5	0.000000
10.Complex	6	0.38	5,931		37.7	0.000000
11.Groups	6	0.38	14,225	0.01	27.5	0.000000
12.Human	6	0.38	17,036	0.02	25.4	0.000000
13.Primates	5	0.32	69		73.6	0.000000
14.Primate	5	0.32	135		67.1	0.000000
15.Ancestors	5	0.32	485		54.4	0.000000
16.Humans	5	0.32	1,125		46.1	0.000000
17.Understanding	5	0.32	4,864		31.6	0.000000
18.Behaviour	5	0.32	6,352		28.9	0.000000
19.Orangutan	4	0.26	4		77.1	0.000000
20.Monkeys	4	0.26	269		46.4	0.000000
21.Sheep	4	0.26	1,901		30.9	0.000000
22.Requires	4	0.26	2,833		27.7	0.000000
23.Animals	4	0.26	4,251		24.5	0.000001
24.Borneo	3	0.19	121		37.9	0.000000
25.Imitation	3	0.19	450		30.0	0.000000
26.Evolved	3	0.19	695		27.5	0.000000
27.Evolution	3	0.19	832		26.4	0.000000

Table 5.7. The key word list of Sample Text B1

In Table 5.7, key words of Sample Text B1 are listed in the first column. The second and third columns provide actual occurrences of the key words and their percentages of frequency in the sample text. The fourth and fifth columns represent the actual occurrences of the key words and percentages of frequency (if greater than 0.01%) in the reference corpus. The last two columns indicate the keyness of the key words and probability value of the statistical calculation as a result of comparison of the sample text and the reference corpus.

If we compare the key word list in Table 5.7 with the word list in Table 5.6, we find that, of the twenty-seven key words in Sample Text B1, there are twenty-two drawn from the thirty-four candidates in Sector 1 of Table 5.6, four from the eight candidates in Sector 3, and only one from the three candidates in Sector 2. This supports our intuition that words in Sector 1 are more likely to be selected as key words by the computer program, if they are close to the central topic of the sample text. However, it disproves our intuition that words in Sector 2 are also important to the topical content, since only one word was selected by the computer program. This phenomenon is interesting, but we do not have much to say since the data are extremely limited at this stage. So we shall leave it for later investigation.

As a reminder, the above four steps of the analysis are summarised as follows:

1. marking out the Theme of each clause;
2. making separate Theme and Rheme corpora;
3. making word lists out of the separate Theme and Rheme corpora, using WordList program; and
4. producing key word lists, using KeyWords program.

Now that the key word lists are ready, we are in a position to test the hypothesis raised in Section 5.2 of this chapter. This is the next step of the analysis, which is the most important part of the present chapter, namely, testing the hypotheses by checking the computed results. The remainder of this chapter will report on the procedures of the analysis and discuss the results.

5.6 Testing the hypotheses

5.6.1 Testing Hypothesis 5.1

To test Hypothesis 5.1, that 'a higher proportion of key words will appear in Theme than in Rheme', it was necessary to compare the proportions of key words in the two areas. This was done by comparing the ratios of key words to all words in the Theme area with those in the Rheme area. Of the 1,559 words in Sample Text B1, there are 554 words in the Theme and 1,005 words in the Rheme. The ratios of key words were obtained by dividing the occurrence of key words by all words in the respective areas. For example, the key word 'intelligence' occurs sixteen times in the sample text, with seven occurrences in Theme and nine occurrences in Rheme. The ratio of occurrence of this particular key word in the whole text is 0.01026, which is the number of key words (16) divided by the number of all words (1,559) in the whole text. The ratios of occurrence of the same key word in Theme and in Rheme were achieved in the same way; that is, by dividing its 7 occurrences by the total of 554 words in Theme, and 9 occurrences by the total of 1,005 words in Rheme, resulting in 0.01264 and 0.00896 respectively. The ratios of key words to all words in Sample Text B1 is presented in Table 5.8 below. For legibility, these ratios are multiplied by 1,000. The ratios of key words to all words in all the Group B Sample Texts are presented in Appendix 13.

Key words	*Frequency*			*Key words (×1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Intelligence	7	9	16	12.64	8.96	10.26
2. Apes	5	5	10	9.03	4.98	6.41
3. Byrne	9	0	9	16.25	0.00	5.77
4. Social	3	6	9	5.42	5.97	5.77
5. Gorillas	4	3	7	7.22	2.99	4.49
6. Chimps	3	4	7	5.42	3.98	4.49
7. Orangutans	4	2	6	7.22	1.99	3.85
8. Ape	4	2	6	7.22	1.99	3.85
9. Species	2	4	6	3.61	3.98	3.85
10.Complex	0	6	6	0.00	5.97	3.85
11.Groups	2	4	6	3.61	3.98	3.85
12.Human	4	2	6	7.22	1.99	3.85
13.Primates	3	2	5	5.42	1.99	3.21
14.Primate	2	3	5	3.61	2.99	3.21
15.Ancestors	3	2	5	5.42	1.99	3.21
16.Humans	2	3	5	3.61	2.99	3.21
17.Understanding	0	5	5	0.00	4.98	3.21
18.Behaviour	1	4	5	1.81	3.98	3.21
19.Orangutan	3	1	4	5.42	1.00	2.57
20.Monkeys	2	2	4	3.61	1.99	2.57
21.Sheep	2	2	4	3.61	1.99	2.57
22.Requires	0	4	4	0.00	3.98	2.57
23.Animals	1	3	4	1.81	2.99	2.57
24.Borneo	2	1	3	3.61	1.00	1.92
25.Imitation	1	2	3	1.81	1.99	1.92
26.Evolved	0	3	3	0.00	2.99	1.92
27.Evolution	1	2	3	1.81	1.99	1.92
Total	*70*	*86*	*156*	*126.35*	*85.57*	*100.06*

Table 5.8. Ratios of key words to all words in Theme and Rheme of Sample Text B1

Table 5.8 shows that in spite of the seemingly irregular ratios of the individual key words, the overall ratio of key word occurrence in Theme is higher than that in Rheme. The same method was applied to all the eleven Group B Sample Texts. For key words of all the Group B Sample Texts see Appendix 11. The summarised results of the calculation are shown in Table 5.9 below. For legibility, these ratios are also multiplied by 1,000.

Texts	*Key words*			*All words*			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
B1	70	86	156	554	1,005	1,559	126.35	85.57	100.06
B2	74	82	156	552	830	1,382	134.06	98.80	112.88
B3	63	64	127	507	963	1,470	124.26	66.46	86.40
B4	43	54	97	375	828	1,203	114.67	65.22	80.63
B5	73	109	182	796	1,326	2,122	91.71	82.20	85.77
B6	49	59	108	639	924	1,563	76.68	63.85	69.10
B7	57	48	105	436	762	1,198	130.73	62.99	87.65
B8	42	37	79	509	891	1,400	82.52	41.53	56.43
B9	46	70	116	474	732	1,206	97.05	95.63	96.19
B10	43	32	75	750	1,020	1,770	57.33	31.37	42.37
B11	114	101	215	640	1,175	1,815	178.13	85.96	118.46
Total	*674*	*742*	*1,416*	*6,232*	*10,456*	*16,688*	*108.15*	*70.96*	*84.85*

Table 5.9. Ratios of key words to all words in Group B Sample Texts

Table 5.9 shows that, in the eleven Group B Sample Texts, the ratios of key words to all words in the Theme area are invariably higher than those in the Rheme area, though the degrees of difference vary. This is more clearly represented in Figure 5.1 below.

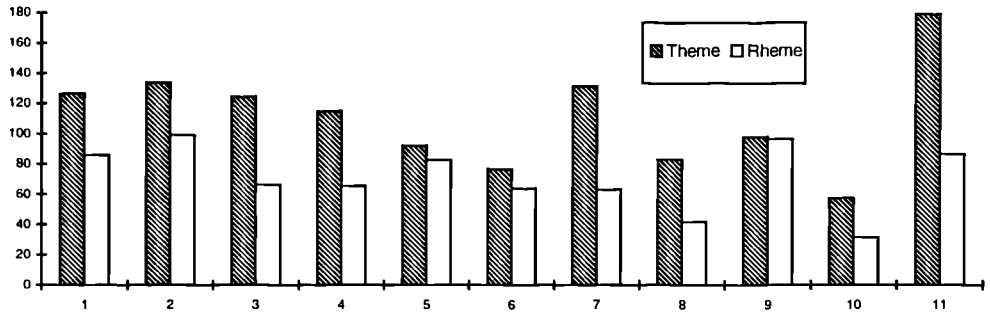


Figure 5.1. Ratios of key words to all words in Theme and Rheme of Group B Sample Texts

In addition, a statistical test was carried out on the results of Table 5.9, to check the validity of this finding. The statistical test used was the *t*-test (two-sample assuming equal variances), which gave a significance value of 0.0037 (P(T<=t)

two-tail), much smaller than the significance value 0.01. This means that the difference of key word distribution between Theme and Rheme is statistically highly significant.

The results of the analysis show that, on the whole, there is a higher proportion of key words in the Theme area than in the Rheme area. In other words, Hypothesis 5.1 is supported by the analysis.

5.6.2 Testing Hypothesis 5.2

Hypothesis 5.2 of this chapter is that 'there will be a higher overall "keyness" in Theme than in Rheme.'

The concept of 'keyness' proposed by Scott (1996) is based on word frequency and is claimed to reflect the 'aboutness' of the text from which the key words are derived. The KeyWords program in the WordSmith Tools is designed to show how the key word is 'key' to the main topic of the text in question through the presentation of the keyness of each key word.

Keyness is achieved by comparing the frequency of a key word in the sample text and its frequency in the reference corpus. In the analysis, the overall keyness in Theme and Rheme areas in the sample text needed to be compared. Many of the key words occur both in Theme and Rheme, so the first step was

to divide the value of keyness of each key word according to their frequency of occurrence in the respective areas.

Take the word 'intelligence', for example. With 16 occurrences in Sample Text B1 and 6,718 occurrences in the 95-million-word corpus, the keyness of this word is 127.7 in the whole text (See Table 5.7 above). Because it occurs seven times in Theme and nine times in Rheme, this figure is split in the proportion of 7:9, resulting in 55.87 and 71.83 respectively. This means that the keyness of the word 'intelligence' in the sample text is 55.87 in Theme and 71.83 in Rheme. The keyness of key words in Sample Text B1 is represented in Table 5.10 below. For the keyness of the key words in the Theme and Rheme areas of all the eleven Group B Sample Texts, see Appendix 14.

	Key words			*Keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Intelligence	7	9	16	55.87	71.83	127.7
2. Apes	5	5	10	73.15	73.15	146.3
3. Byrne	9	0	9	104.60	0.00	104.6
4. Social	3	6	9	11.43	22.87	34.3
5. Gorillas	4	3	7	59.89	44.91	104.8
6. Chimps	3	4	7	43.89	58.51	102.4
7. Orangutans	4	2	6	80.53	40.27	120.8
8. Ape	4	2	6	51.80	25.90	77.7
9. Species	2	4	6	14.83	29.67	44.5
10.Complex	0	6	6	0.00	37.70	37.7
11.Groups	2	4	6	9.17	18.33	27.5
12.Human	4	2	6	16.93	8.47	25.4
13.Primates	3	2	5	44.16	29.44	73.6
14.Primate	2	3	5	26.84	40.26	67.1
15.Ancestors	3	2	5	32.64	21.76	54.4
16.Humans	2	3	5	18.44	27.66	46.1
17.Understanding	0	5	5	0.00	31.60	31.6
18.Behaviour	1	4	5	5.78	23.12	28.9
19.Orangutan	3	1	4	57.83	19.28	77.1
20.Monkeys	2	2	4	23.20	23.20	46.4
21.Sheep	2	2	4	15.45	15.45	30.9
22.Requires	0	4	4	0.00	27.70	27.7
23.Animals	1	3	4	6.13	18.38	24.5
24.Borneo	2	1	3	25.27	12.63	37.9
25.Imitation	1	2	3	10.00	20.00	30.0
26.Evolved	0	3	3	0.00	27.50	27.5
27.Evolution	1	2	3	8.80	17.60	26.4
Total	*70*	*86*	*156*	*796.62*	*787.18*	*1,583.8*

Table 5.10. Keyness of key words in Theme and Rheme of Sample Text B1

Table 5.10 shows a variety of keyness for the individual key words similar to that of the key word ratios, but the overall keyness is only slightly higher in Theme than in Rheme. However, this is a distorted picture, because the keyness of a word is based on its frequency in the context and our splitting of keyness did not take account of the factor of length of Theme or Rheme context. A fairer picture may be obtained by comparing the keyness of individual key words with all the words in corresponding areas of the text.

As shown in Table 5.2, out of the total number of 1,559 words in the first sample text, 554 words occur in Theme and 1,005 in Rheme. To obtain the ratios of keyness to all words in Theme and Rheme, the keyness of each key word was divided by the total number of words in the Theme and Rheme areas. The results of this calculation are presented in Table 5.11 below. For legibility, the numbers are multiplied by 100.

Key words	*Keyness*		
	Theme (x100)	*Rheme (x100)*	*Overall (x100)*
1. Intelligence	10.08	7.15	8.19
2. Apes	13.20	7.28	9.38
3. Byrne	18.88	0.00	6.71
4. Social	2.06	2.28	2.20
5. Gorillas	10.81	4.47	6.72
6. Chimps	7.92	5.82	6.57
7. Orangutans	14.54	4.01	7.75
8. Ape	9.35	2.58	4.98
9. Species	2.68	2.95	2.85
10. Complex	0.00	3.75	2.42
11. Groups	1.66	1.82	1.76
12. Human	3.06	0.84	1.63
13. Primates	7.97	2.93	4.72
14. Primate	4.85	4.01	4.30
15. Ancestors	5.89	2.17	3.49
16. Humans	3.33	2.75	2.96
17. Understanding	0.00	3.14	2.03
18. Behaviour	1.04	2.30	1.85
19. Orangutan	10.44	1.92	4.95
20. Monkeys	4.19	2.31	2.98
21. Sheep	2.79	1.54	1.98
22. Requires	0.00	2.76	1.78
23. Animals	1.11	1.83	1.57
24. Borneo	4.56	1.26	2.43
25. Imitation	1.81	1.99	1.92
26. Evolved	0.00	2.74	1.76
27. Evolution	1.59	1.75	1.69
Total	*143.79*	*78.30*	*101.59*

Table 5.11. Ratios of keyness to all words in Theme and Rheme of Sample Text B1

Table 5.11 shows that, although there are variations for individual words, the overall ratio of keyness to all words is higher in Theme than in Rheme. A

similar analysis was carried out for the other sample texts. Detailed results of the analysis of all the eleven Group B Sample Texts are presented in Appendix 15. Table 5.12 below is a summary of the ratios of keyness to all words in all the eleven Group B Sample Texts.

Texts	*Theme*	*Rheme*	*Overall*
B1	143.79	78.33	101.59
B2	125.38	92.40	105.57
B3	123.93	66.28	86.16
B4	125.76	71.52	88.43
B5	68.48	61.39	64.05
B6	70.63	58.81	63.64
B7	187.10	90.15	125.43
B8	98.92	49.78	67.65
B9	78.01	76.87	77.31
B10	42.88	23.46	31.69
B11	187.68	90.57	124.81

Table 5.12. Ratios of keyness to all words in Theme and Rheme of Group B Sample Texts

The ratios of keyness over all words in Theme and Rheme of Group B Sample Texts are graphically presented in Figure 5.2 below.

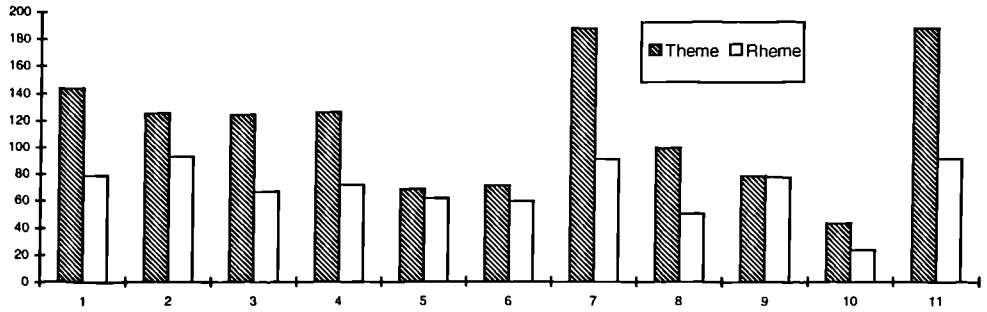


Figure 5.2. Ratios of keyness to all words in Theme and Rheme of Group B Sample Texts

The ratios of keyness over all words in Theme are invariably higher than those in Rheme. The *t*-test (two-sample assuming equal variances) shows that the p-value is 0.0094 (P(T<=t) two-tail), less than 0.01. That is, the differences between Ratios of keyness to all words in Theme and Rheme areas are statistically significant.

By now it may be concluded that Hypothesis 5.2 is supported, that there is a higher overall 'keyness' in Theme than in Rheme.

5.6.3 Testing Hypothesis 5.3

Hypothesis 5.3 was that 'on average, a key word in Theme will be reiterated more than a key word in Rheme, thus forming more links in Theme than in Rheme'. In order to test this hypothesis, it was necessary to find out how many times a key word is repeated in the context in which it appears. This was obtained by comparing the token and type ratios of key words in Theme and the token and type ratios of key words in Rheme.

Take Sample Text B1 for example. There are 27 types of key words, with 156 actual occurrences (tokens). The type/token ratio is 156 divided by 27, which equals 5.78. However, 70 of these tokens occur in Theme, which contains 23 types. The type/token ratio in Theme is 70 divided by 23, resulting in 3.04. In Rheme there are 86 tokens, which are of 26 types. The type/token ratio in Rheme is 86 divided by 26, which equals 3.31. This shows that the type/token

ratio is highest for the whole text and is at its lowest in Theme. Table 5.13 below shows the key word type/token ratios of the eleven Group B Sample Texts.

Texts	*(1) Key-word token*			*(2) Key-word type*			*(3) Type/token ratio*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
B1	70	86	156	23	26	27	3.04	3.31	5.78
B2	74	82	156	22	22	25	3.36	3.73	6.24
B3	63	64	127	21	19	21	3.00	3.37	6.05
B4	43	54	97	18	20	22	2.39	2.70	4.41
B5	73	109	182	17	18	18	4.29	6.06	10.11
B6	49	59	108	18	20	20	2.72	2.95	5.40
B7	57	48	105	19	17	19	3.00	2.82	5.53
B8	42	37	79	14	14	15	3.00	2.64	5.27
B9	46	70	116	14	18	19	3.29	3.89	6.11
B10	43	32	75	11	9	11	3.91	3.56	6.82
B11	114	101	215	25	26	30	4.56	3.88	7.17

Table 5.13. Key-word type/token ratios in Group B Sample Texts

In Table 5.13, we see that the key word type/token ratios in Theme are generally lower than those in Rheme. This seems to mean that the key word is repeated more in the Rheme than in the Theme. However, this simple comparison produces a distorted picture. The results cannot be interpreted in this way.

The reason why a key word type in Theme appears less repeated than it is in Rheme is that the overall types of words in English, or rather in any language, are limited when compared with the tokens. The longer the text, the greater the repetition of word types is likely to be, and the higher the type/token ratio. In all the eleven Group B Sample Texts, without exception, Rhemes are roughly twice as long as Themes (See Table 5.5).

A fairer picture may be obtained by comparing the key word type/token ratios in a fixed length of text. One way of doing this would be to calculate the key word type/token ratios in every 1,000 words of the Theme area and then every 1,000 words of the Rheme area of a sample text. But a more practicable and reliable solution to this problem is to divide the key word type/token ratios in Table 5.13 by the number of words in different text segments, so that the results become compatible. The latter approach was adopted in the analysis. The results of this calculation are presented in Table 5.14, which is an extension of Table 5.13. Again, for legibility, the numbers are multiplied by 1,000.

Texts	*(4) All words*			*(5) T-t ratio/all words (×1,000)*		
	Theme	*Rheme*	*Overall*	*Theme*	*Rheme*	*Overall*
B1	554	1,005	1,559	5.49	3.29	3.71
B2	552	830	1,382	6.09	4.49	4.52
B3	507	963	1,470	5.92	3.50	4.11
B4	375	828	1,203	6.37	3.26	3.67
B5	796	1,326	2,122	5.39	4.57	4.76
B6	639	924	1,563	4.26	3.19	3.45
B7	436	762	1,198	6.88	3.71	4.61
B8	509	891	1,400	5.89	2.97	3.76
B9	474	732	1,206	6.93	5.31	5.06
B10	750	1,020	1,770	5.21	3.49	3.85
B11	640	1,175	1,815	7.13	3.31	3.95

Table 5.14. Key-word type/token ratios in Group B Sample Texts (based on all words)

Table 5.14 shows that the key word type/token ratios calculated on the basis of all words are invariably higher in the Theme areas of the eleven Group B Sample Texts. That is, in a context of a fixed number of words, a key word generally forms more links in the Theme area than in the Rheme area.

Figure 5.3 below presents a schematic demonstration of this result.

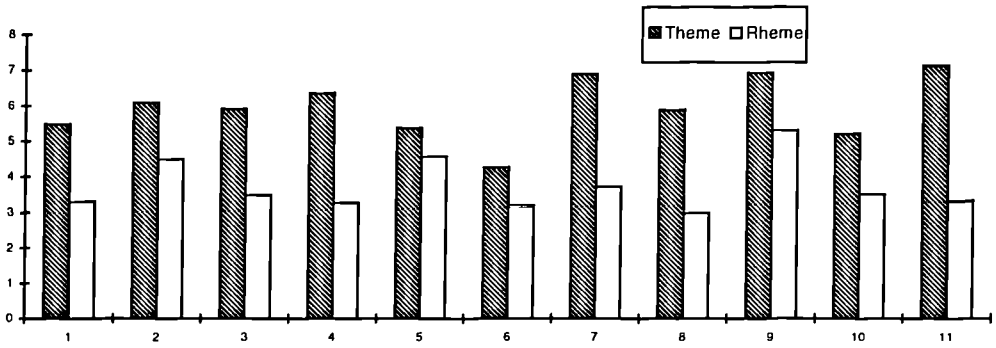


Figure 5.3. Key-word type/token ratios of Group B Sample Texts

Figure 5.3 shows that the key word type/token ratios based on all words are invariably higher in Theme than in Rheme in all the eleven Group B Sample Texts. A *t*-test (two-sample assuming equal variances) was carried out to test the results. The p-value is 0.000002 (P(T<=t) two-tail), which means that the differences in the number of links formed by key words in Theme and those in Rheme are statistically highly significant.

It may be concluded that in a fixed length of text key words in Theme tend to form more links than those in Rheme. That is, Hypothesis 5.3 of the current chapter is supported by the analysis.

5.7 Conclusion

In the light of the above findings, we may conclude that hypotheses 1, 2 and 3 are all supported by the analysis. Comparing Figure 5.1, 5.2 and 5.3, we find that the patterns are strikingly similar. This may suggest that the proportions of key words, keyness, and links formed by key words are all closely correlated, although 'keyness' seems to be a more objective index of the role of lexical items in the Theme and Rheme areas with regard to the organisation of the text, and an index of the 'aboutness' of a text. The results of the *t*-test may imply that, with increased samples, the results become more reliable. In order for such a claim to be confirmed, more research is certainly needed on the basis of a much larger corpus.

This chapter only reveals some features of the Theme-Rheme system in a specific genre in the English language. However, it is hoped that this method of study may also apply to different genres and to several other European languages, and even oriental languages such as Chinese, which have similar features in the Theme-Rheme system.

The results of this chapter have implications for various areas. In automatic text summarisation, especially in summarisation of scientific reports, computers may be programmed to add weight to key words in the Theme area, in order to better retain the central information of the source text. In English language teaching, e.g. in teaching reading strategies to non-native English language

learners, students may be trained to pay attention to the clause Themes of a text, to establish quickly a framework of what the text is about, so that they may become more efficient in reading. In translation, the retention of key words in Theme may reflect the conservation of the pattern of information distribution within the text or its stylistic flavour in the translation of its source text, depending on the languages involved.

In response to the second general research question raised in Section 1.5 of Chapter 1, this chapter has demonstrated that evidently there is a close relationship between key words and the Theme-Rheme system in the text. The relationship is manifested in the form of a higher proportion of key words, a higher rate of keyness and a likelihood of reiteration of key words in the Theme area of the text. The next chapter will move on to answer the third general research question, examining whether there is a three-way relationship amongst lexical patterning, key words and the Theme-Rheme system, and if there is, how it is manifested in the text.

Chapter 6. Key-word Links in Theme and Rheme

6.1 Introduction to this chapter

In Chapter 4 we answered the first general research question by exploring the distribution of cohesive links formed by lexical items in the Theme and Rheme areas of the text. We found that there are proportionally more links between lexical items in Theme than there are between lexical items in Rheme. We also found that links formed by lexical items between the Theme of the succeeding sentences and the Rheme of the preceding sentences outnumber links formed by lexical items between the Rheme of the succeeding sentences and the Theme of the preceding sentences. Since lexical links are regarded as a device to retain information focus in the text, this finding suggests that, collectively, viewed at the level of the text, Theme is more important than Rheme in creating cohesive links and retaining topical continuity in the text.

In Chapter 5 we answered the second general research question by exploring the distribution of key words in the Theme and Rheme areas of the text. We found that there is a very strong tendency for key words to concentrate in the

Theme area of the text. Since key words may be regarded as a good indicator of the 'aboutness' of the text, the concentration of key words in Theme partly supports Berry's hypothesis that 'discourse $Theme_M$' is realised by 'clause $Theme_Fs$' of the text. It also suggests that Halliday's definition of Theme as 'what the clause is about' should be expanded onto the text level; that is, cumulatively, the clause Themes of the text reflect the aboutness of the text.

This chapter attempts to carry the exploration of cohesive links in Theme and Rheme of the text a step further. It intends to explore the distribution of cohesive links formed by key words in the Theme and Rheme areas, and in so doing tackles the third general research question, to see whether there is a three-way relationship amongst lexical patterning, key words and the Theme-Rheme system.

6.2 Assumptions

The findings of Chapter 4 reveal the relationship between lexical links and the Theme-Rheme system, which may be represented in Figure 6.1 below.

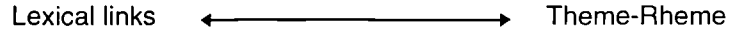


Figure 6.1. Relationship between lexical links and the Theme-Rheme system

The findings of Chapter 5 reveal the relationship between key words and the Theme-Rheme system, which may be represented in Figure 6.2 below.

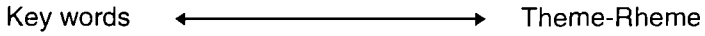


Figure 6.2. Relationship between key words and the Theme-Rheme system

Figure 6.1 and Figure 6.2 together reveal that the key words and lexical links are connected by the Theme-Rheme system. This may be shown in Figure 6.3 below.

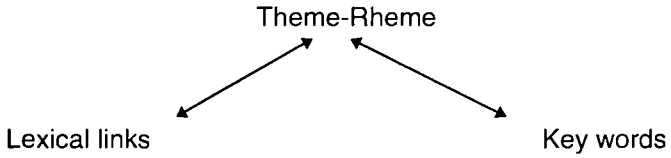


Figure 6.3. Relationship between the findings of Chapter 4 and Chapter 5 of this thesis

From Figure 6.3 it may be assumed that the findings of Chapters 4 and 5 are meaningfully related, i.e. there will be a relationship between lexical links and key words in the text. Thus Figure 6.3 may be redrawn as Figure 6.4 below.

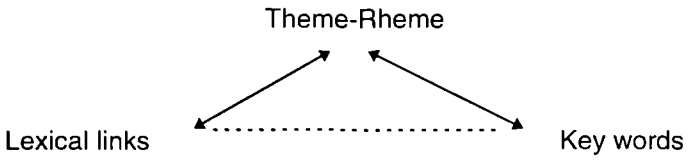


Figure 6.4. Relationship between lexical links and key words in the text

In order to explore the relationship between lexical links and key words, we need a new notion, the *key-word links*, which are defined in this thesis as cohesive links formed by key words in the text. This notion is of central importance to the study in this chapter, so it will be discussed in more detail in

the next section. The notion of key-word links connects lexical links and key words. This leads to Figure 6.5 below.

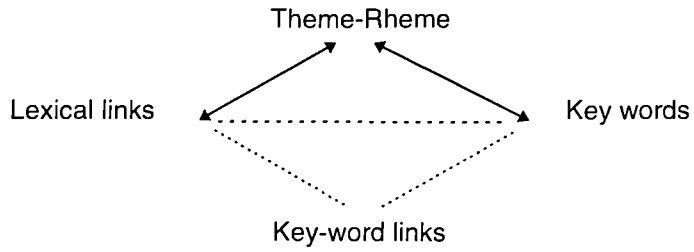


Figure 6.5. Lexical links, key words and key-word links in the text

However, Figure 6.5 does not complete the picture. Just as the test of lexical links in Chapter 4 and the test of key words in Chapter 5 are both related to the Theme-Rheme system, the test of key-word links will also be related to the Theme-Rheme system. Thus we arrive at Figure 6.6.

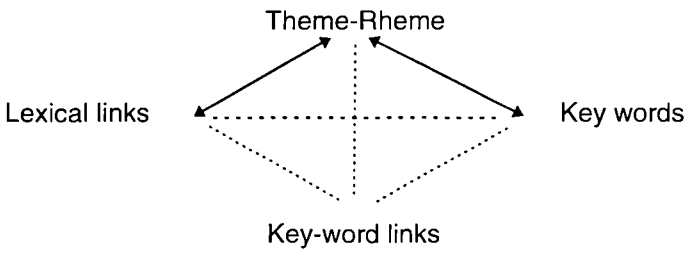


Figure 6.6. Relationship between the four categories

As explained in Chapter 5, most key words are simultaneously lexical items. This suggests that a correlation may exist between the number of lexical items and the number of key words in the text. Therefore, it is assumed that the occurrence of lexical items in the text will be in proportion to the occurrence of key words in the same text. In other words, other things being equal, a text with

more lexical items will have more key words, and conversely, a text with more key words will have more lexical items.

Consequently, there may be a correlation between the occurrences of lexical links and key-word links in the text; it might be hypothesised that a text with more lexical links will have more key-word links and a text with more key-word links will have more lexical links.

Further, there may be a correlation between the occurrences of key words and key-word links in the text. It might reasonably be assumed that a text with more key words will have more key-word links and a text with more key-word links will have more key words.

However, what interests us most in this chapter is the distribution of the words and links in the Theme and Rheme areas of the text. Based on the above assumptions, a logical further step would be to investigate the properties of key-word links, so that the relationship between key words and lexical links in the Theme and Rheme areas of the text may be better revealed.

This chapter builds on the basic assumption that key-word links can indicate information distribution in the text, as a special case of lexical links indicating information distribution in the text. Likewise, just as key words indicate information distribution in the text, key-word links will also indicate information distribution in the text. Therefore, a statistical analysis of key-word links should reveal information distribution in the text as clearly as the

statistical analysis of other lexical links and/or the statistical analysis of key words.

6.3 Hypotheses

The assumptions stated in Section 6.2 suggest that, in order to investigate the relationships between lexical links and key words in the Theme and Rheme areas of the text, it is necessary to explore the correlation between the ratios of occurrences amongst lexical links, key words and key-word links in the Theme and Rheme areas of the text. For this purpose, it is necessary to set up some testable hypotheses.

In this chapter, the term 'key words' will generally be used to mean key-word tokens as opposed to key-word types. For reasons of brevity, and wherever the context is clear enough not to cause confusion, the short-hand term 'key words' will be used instead of the fuller term 'key-word tokens'. On the few occasions when reference to key-word types is made, it will be explicitly stated.

By definition, key words as proposed by Scott (1996) are words that have an unusual frequency in the text under study comparative to some norm in the language. Most key words are lexical items. Moreover, all key words identified in the study are positive key words, which have a minimum frequency of three occurrences in the text, and which tend to have a higher frequency of

occurrence than other lexical items in the text comparative to some norm in the language (cf. Scott 1996; also Chapter 5 of this thesis). This suggests that key words are likely to form a higher proportion of links than average lexical items, since the number of links is a direct consequence of the degree of repetition of word tokens. Therefore, Hypothesis 6.1 of this chapter is that:

Hypothesis 6.1. **In a text, the ratio of key-word links to key words will be higher than the ratio of lexical links to lexical items.**

Further, the distribution of key words may be in proportion to the distribution of lexical items both in the Theme and Rheme areas of the text. Therefore we can further formulate a hypothesis that

Hypothesis 6.2. **In the Themes of the text, the ratio of key-word links to key words will be higher than the ratio of lexical links to lexical items.**

And similarly,

Hypothesis 6.3. **In the Rhemes of the text, the ratio of key-word links to key words will be higher than the ratio of lexical links to lexical items.**

Hypothesis 6.1, Hypothesis 6.2 and Hypothesis 6.3 together suggest that the link/word ratios will be higher with the key words than with average lexical items either in the whole text or in the Theme and Rheme areas. Testing these

hypotheses will reveal some aspects of the property of key-word links in relation to the test of lexical links in Chapter 4.

On the other hand, it may be assumed from experience that the total number of links is generally greater than the total number of word tokens. Therefore, Hypothesis 6.4 of this chapter is set up as follows:

Hypothesis 6.4. **In a text, the ratio of key-word links to all words will be higher than the ratio of key words to all words.**

The differences of ratios may also be in proportion in Theme and Rheme areas of the text, thus it may be further hypothesised that,

Hypothesis 6.5. **In the Themes of the text, the ratio of key-word links to all words will be higher than the ratio of key words to all words.**

Likewise,

Hypothesis 6.6. **In the Rhemes of the text, the ratio of key-word links to all words will be higher than the ratio of key words to all words.**

Testing Hypotheses 6.4 to 6.6 is intended to reveal some aspects of the property of key-word links in relation to key words.

Hypotheses 6.1 to 6.6 together imply that key words as a special kind of lexical item may be a clearer indicator of the cohesion of a text than average lexical items. Hasan's (1984) notion of cohesive chains may be better thought of as chains formed basically of key words rather than of average lexical items, and the degree of coherence in a text may be measured more clearly by counting cohesive chains formed by key-word links in the text than by counting cohesive chains formed by average lexical items in the text. Further, a text is better viewed as an instance of texts rather than an isolated phenomenon. The degree of coherence of a text should thus be better measured in comparison to some norm of the language rather than measured in isolation. Since a key word cannot be identified without comparison of its performance in the text to its performance in the language in general, the comparison with a norm corpus may allow one to identify significant cohesive chains more effectively than cohesive chains formed by average lexical items. Thus key-word links may signal the patterns of information distribution in text more clearly than either lexical links or key words only. Therefore, further exploring the distribution of key-word links in the Theme and Rheme areas of the text may help to reveal the text organisation in terms of information distribution better than simply exploring the distribution of either lexical items or key words.

As shown by the analysis in Chapter 4, there is a higher ratio of lexical links to all words in the Theme area than in the Rheme area of the text; and as shown by the analysis in Chapter 5, there is a higher ratio of key words to all words in the Theme area than in the Rheme area of the text. If a correlation exists

between the occurrences of lexical links and key words in the text, this correlation may also be reflected by key-word links.

If Hypotheses 6.1, 6.2 and 6.3 are supported, it could be assumed that the density of key-word links is generally higher than the density of other lexical links. If Hypotheses 6.4, 6.5 and 6.6 are supported, it could be assumed that there are proportionally more key-word links than key words in the text. Then, based on the findings in Chapters 4 and 5, we would be in a position to hypothesise that Themes contain proportionally more key-word links than Rhemes. Thus Hypothesis 6.7 of this chapter is that:

Hypothesis 6.7. The ratio of key-word links to all words will be higher in the Themes than in the Rhemes of the text.

In order to test the above hypotheses, it is necessary to obtain data concerning lexical links, key words, and key-word links, in both the Themes and Rhemes of the sample texts. Since the investigation in this chapter is a continuation of the studies reported in the previous chapters, it is consistent as well as economical to re-use some of the data already obtained in the previous chapters.

6.4 An initial analysis

In this section, Sample Text A, the 'Planet X' text which was analysed in Chapter 4, was re-used for an initial examination of the hypotheses of this chapter. Since in Chapter 4 we already counted the number of sentences in the text and made lists of both lexical items and lexical links of Sample Text A, all we needed to do was to make a key-word list and count the links formed by key words in the Theme and Rheme areas. The computer program 'KeyWords' was used again for making the key-word list on the basis of the word list of Sample Text A (See Appendix 2), with minimum frequency of the key words set at three occurrences and the P value at 0.000001. The key-word list of Sample Text A is presented in Table 6.1 below, in the decreasing order of keyness of the key words.

N	*Key words*	*A1freq*	*A1%*	*Ref freq*	*Ref %*	*Keyness*	*P*
1.	Planet	37	17.29	2,158		392.9	0.000000
2.	Harrington	18	8.41	169		351.2	0.000000
3.	Tombaugh	12	5.61	2		301.3	0.000000
4.	Uranus	12	5.61	80		241.5	0.000000
5.	Pluto	12	5.61	96		237.4	0.000000
6.	Neptune	8	3.74	121		144.0	0.000000
7.	Astronomers	7	3.27	263		117.3	0.000000
8.	Observatory	6	2.80	202		101.8	0.000000
9.	Orbit	7	3.27	511		92.8	0.000000
10.	Solar	6	2.80	746		86.3	0.000000
11.	Astronomer	5	2.34	196		83.3	0.000000
12.	Search	6	2.80	5,692		62.0	0.000000
13.	Lowell	3	1.40	94		51.3	0.000000
14.	Mathematicians	3	1.40	137		49.1	0.000000
15.	Sun	5	2.34	9,602	0.01	44.6	0.000000
16.	Computer	5	2.34	11,531	0.01	42.8	0.000000
17.	Small	6	2.80	32,027	0.03	41.4	0.000000
18.	Faint	3	1.40	690		39.4	0.000000
19.	Sky	4	1.87	4,872		39.3	0.000000
20.	Naval	3	1.40	1,579		34.5	0.000000
21.	Searching	3	1.40	1,665		34.2	0.000000
22.	Models	3	1.40	2,748		31.2	0.000000
23.	Predicted	3	1.40	3,218		30.2	0.000000
24.	Positions	3	1.40	3,430		29.8	0.000000
25.	Stars	3	1.40	4,773		27.9	0.000000

Table 6.1. Key-word list of Sample Text A

Table 6.1 shows that 25 key-word types were selected from Sample Text A. On the top of the list are mostly names of planets in our solar system and scientists (astronomers) who are reported as having carried out the search for the new planet. Not surprisingly, the most frequent key word is 'planet', which occurs 37 times in the text. This fits well with common intuition about the topic of this text. But the list does not tell us about the distribution of the key words in the Theme and Rheme areas. The number of occurrences of each key word in the Theme and Rheme areas of Sample Text A is presented in Table 6.2 below, in decreasing order of occurrence in the whole text.

N	*Word*	*Theme*	*Rheme*	*Overall*
1.	Planet	13	24	37
2.	Harrington	17	1	18
3.	Tombaugh	10	2	12
4.	Pluto	8	4	12
5.	Uranus	2	10	12
6.	Neptune	4	4	8
7.	Astronomers	6	1	7
8.	Observatory	4	2	6
9.	Orbit	2	4	6
10.	Search	2	4	6
11.	Small	1	5	6
12.	Solar	1	5	6
13.	Astronomer	5	0	5
14.	Computer	3	2	5
15.	Sun	0	5	5
16.	Sky	1	3	4
17.	Mathematicians	3	0	3
18.	Models	3	0	3
19.	Naval	3	0	3
20.	Lowell	2	1	3
21.	Positions	1	2	3
22.	Predicted	1	2	3
23.	Faint	0	3	3
24.	Searching	0	3	3
25.	Stars	0	3	3
	Total	*92*	*90*	*182*

Table 6.2. Key words in Theme and Rheme of Sample Text A

Table 6.2 shows that the overall occurrences of key-word tokens are divided fairly equally between the Themes and Rhemes of Sample Text A, although variations exist between individual key-word types. The key word 'planet', for example, occurs 13 times in Theme and 24 times in Rheme, which may be regarded as a fairly even distribution, considering the respective lengths of the Themes and Rhemes of the text. On the other hand, the key words 'Harrington', 'Tombaugh' and 'Astronomers' mostly occur in Theme rather than in Rheme of the text, while the key words 'Uranus', 'solar' and 'sun' concentrate in the Rheme of the text. This suggests that key words may be divided into different categories: those that mainly concern the actors of the processes and those that mainly relate to the recipients of the processes. Interesting though this hypothesis is, it is not the focus of the present study, and will be left for further investigation.

The percentages of all words, of lexical items and of key words in the Theme and Rheme areas of the text were all worked out by comparing Table 6.2 with Table 4.1 of Chapter 4. The percentages are presented in Table 6.3 below.

	Theme	*Rheme*	*Overall*	*Theme %*	*Rheme %*	*Overall %*
All words	426	745	1,171	36.38	63.62	100.00
Lexical items	236	404	640	36.88	63.13	100.00
Key words	92	90	182	50.55	49.45	100.00

Table 6.3. Word distribution in Sample Text A

Table 6.3 shows that whereas the proportions of all word tokens and all lexical items in Theme are around one third of the whole text, half of the key-word tokens of Sample Text A are in Theme. This confirms the examination of

Hypothesis 5.1 of Chapter 5, that there is a higher ratio of key words to all words in Theme than in Rheme.

In order to test the hypotheses of this chapter, it was also necessary to count the number of key-word links in Sample Text A. This count was carried out in three steps. Firstly, the location of each of the key-words in the sentences of Sample Text A was identified. Then, key-word links were counted sentence by sentence. For example, Sentence 1 has no key words in its Theme, thus it has no Theme to Theme (T-T) links or Theme to Rheme (T-R) links with other sentences of this text. However, Sentence 1 has two key words in Rheme, so it may have Rheme to Rheme (R-R) links and Rheme to Theme (R-T) links with other sentences. One of the key words in the Rheme of Sentence 1 is 'solar', which also occurs in the Theme of Sentence 32 and Rheme of Sentences 9, 42, 48 and 56, thus forming one R-T link and four R-R links with those sentences . The other key word in Sentence 1 is 'planet', which also occurs in the Theme of Sentences 2, 7, 14, 37, 39, 45, 48, 55 and 57, and in the Rheme of Sentences 3, 4, 9, 11, 12, 13, 15, 16, 18, 23, 26, 27, 32, 34, 42, 43, 44 and 58, thus forming nine R-T links and nineteen R-R links with those sentences. In the analysis intra-sentential links were not counted: when a key word occurs more than once in a sentence, it was counted only once and was regarded as forming only one link. In summary, Sentence 1 has no T-T or T-R links, but it has ten R-T links and twenty-three R-R links with other sentences in the text.

Finally, in the third step, all the key-word links counted in this way, sentence by sentence, were added up. The total number of key-word links in Sample Text A is that of 340 T-T links, 267 R-T links, 392 R-R links, and 105 T-R links, as is shown in Table 6.4 below. For convenience of comparison, data regarding average lexical items and lexical links are also included.

	Tokens			*Links*				
	Theme	*Rheme*	*Overall*	*T-T*	*R-T*	*R-R*	*T-R*	*Overall*
Lexical items	154	269	423	362	314	494	139	1,309
Key words	92	90	182	340	267	392	105	1,104

Table 6.4. Frequencies of the four basic types of links in Sample Text A

Apparently, in Table 6.4, there are more R-R links than T-T links formed by key words. However, as stated in Chapter 4, this is a direct consequence of the fact that Rheme is longer than Theme in this sample text. So we need to consider the context factor and calculate the ratios of links to all words in the Theme and Rheme areas. As for the links that run across the Theme and Rheme areas, since they are based on the same text areas, a simple comparison would be enough. Table 6.4 shows that, with regard to these, there are more than twice as many R-T links as T-R links formed by key words in this text. The results of the analysis of key-word links are in good agreement with the results of the analysis of lexical links as reported in Chapter 4.

It should be remembered that this analysis is concerned with comparing the links within the Theme and Rheme areas. Since R-T links and T-R links run across these two areas, it was decided to leave these two types of links aside and focus on comparing the ratios of occurrences of the other two types of links

to all words in the Theme and Rheme areas of the text. The results of the comparison are shown in Table 6.5 below.

	Tokens			*Links*			*Link/Token Ratios*		
	Th	*Rh*	*Overall*	*T-T*	*R-R*	*Overall*	*Th*	*Rh*	*Overall*
Lexical items	154	269	423	362	494	856	2.3506	1.8364	2.0236
Key words	92	90	182	340	392	732	3.6957	4.3556	4.0220

Table 6.5. Link/token ratios in Sample Text A

Table 6.5 shows that in Sample Text A the ratios of key-word links to key words are constantly higher than the ratios of lexical links to lexical items, whether in Theme, Rheme or the whole text. This implies that Hypotheses 6.1 to 6.3 of this chapter may be supported.

However, while the ratio of lexical links to lexical items is higher in Theme than in Rheme, the ratio of key-word links to key words is lower in Theme than in Rheme. One reason for this imbalance is that the ratios of key-word links were not calculated on the basis of the length in words of the Theme and Rheme areas of the text. As explained in Chapter 5, in the examination of the link/token ratios, it is necessary to consider the length of the segment of text in which the words or links occur. As shown in Table 6.3 above, there are 1,171 words in Sample Text A, including 426 words in Theme and 745 words in Rheme. A recalculation of the ratios on the basis of the original text length reveals that both the ratio of lexical links to lexical items and the ratio of key-word links to key words are higher in Theme than in Rheme, as shown in Table 6.6 below. For legibility, the resulting ratios are multiplied by 1,000.

	Theme	*Rheme*	*Overall*		*Theme*	*Rheme*	*Overall*
LL/LW	2.3506	1.8364	2.0236	LL/LW//TL (x1,000)	5.5178	2.4650	1.7281
KL/KW	3.6957	4.3556	4.0220	KL/KW//TL (x1,000)	8.6754	5.8464	3.4347

Table 6.6. Link/token ratios corrected for length of text areas in Sample Text A

In Table 6.6, LL means lexical links, LW means lexical items, and LL/LW means link/token ratios for lexical items. KL means key-word links, KW means key words, and KL/KW means link/token ratios for key words. TL means the length of text areas in words. So, LL/LW//TL means link/token ratios for lexical items corrected for the length of text areas in words, and KL/LW//TL means link/token ratios for key words corrected for the length of text areas in words. As shown in Table 6.6, comparing the link/word ratios on the basis of the context in which the links and words occur, the ratios are higher in Theme than in Rheme.

To test Hypotheses 6.4 to 6.6 of this chapter, it was necessary to compute the ratios of key-word links on the basis of all words. Table 6.5 above contains the number of key words and key-word links in the Theme and Rheme areas of Sample Text A, and Table 6.3 above shows that there are altogether 1,171 words in Sample Text A, with 426 words in Theme and 745 words in Rheme. The ratios of key words to all words and key-word links to all words are presented in Table 6.7 below.

	Theme	*Rheme*	*Overall*		*Theme*	*Rheme*	*Overall*
KW	92	90	182	KW/TL	0.2160	0.1208	0.1554
KL	340	392	732	KL/TL	0.7981	0.5262	0.6251

Table 6.7. Ratios of key words and key-word links to the length of text areas in words in Sample Text A

Table 6.7 shows that both the ratios of key words to the length of text areas in words (KW/TL) and the ratios of key-word links to the length of text areas in words (KL/TL) are higher in Theme than in Rheme in Sample Text A. Moreover, the ratios of key-word links to the length of text areas in words in Theme and Rheme are both about four times as great as the ratios of key words to the length of text areas in words. Figure 6.7 below clearly represents these differences.

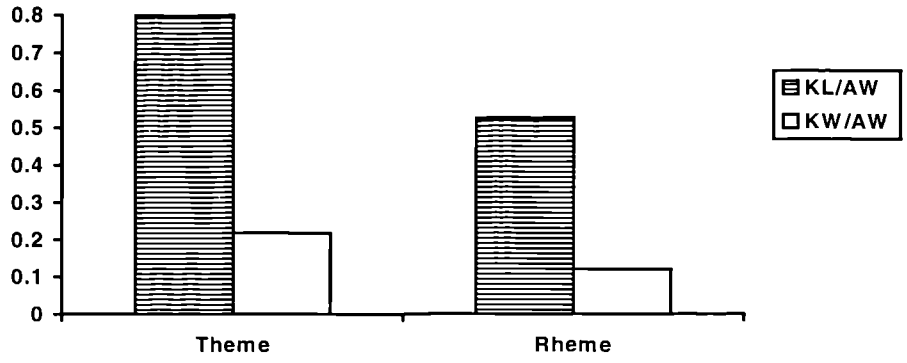


Figure 6.7. Comparison of the ratios of key words to all words and key-word links to all words in Sample Text A

Figure 6.7 shows that the ratios of key-word links to all words are higher than the ratios of key words to all words either in Theme, Rheme or the whole text. This finding supports Hypothesis 6.4, Hypothesis 6.5 and Hypothesis 6.6 of this chapter.

In addition, if we read Table 6.7 horizontally, we may find that there is a higher ratio of key-word links to all words in Theme than in Rheme of Sample Text A, just as there is a higher ratio of key words to all words in Theme than in Rheme. This is shown in Figure 6.8 below. This suggests that Hypothesis 6.7 may also be supported by the analysis.

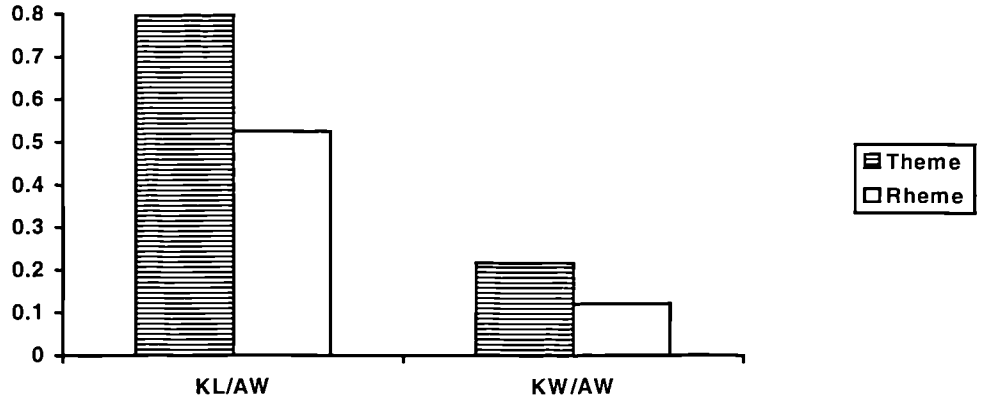


Figure 6.8. Comparison of the ratios of key words and key-word links to all words in the Theme and Rheme areas of Sample Text A

So far, an experiment has been carried out on Sample Text A in testing the seven hypotheses of this chapter. Apparently all the seven hypotheses have been supported. However, although the results of the initial analysis are quite encouraging, analysis of only one sample text is obviously not enough, especially when the nature of the study is to draw on statistical evidence to describe text organisation. Therefore, more texts will need to be analysed before any confident claims can be made.

6.5 A large-scale analysis

This section will report on a large-scale analysis along the lines described above. For this analysis, the eleven sample texts called 'Group B Texts', which were used in Chapter 5, were re-used. In addition, a group of twenty more texts, which will be called 'Group C Sample Texts', were used. The two groups add up to thirty-one sample texts in total.

6.5.1 Re-using Group B Sample Texts

This analysis required data concerning the number of all words, lexical items, key words, as well as the number of lexical links and key-word links in the Theme and Rheme areas of the texts.

Since key word lists had already been made out of the Group B Sample Texts for the work reported in Chapter 5, what needed to be done with these texts was to count the links formed by lexical items and key words in the Theme and Rheme areas. The relevant data of Group B Sample Texts as reported in Chapter 5 will be represented in the following sections in connection with Group C Sample Texts.

6.5.2 Collecting Group C Sample Texts

Group C Sample Texts consist of twenty texts collected from the same source as the other sample texts, i.e. the British newspaper 'The Independent on Sunday'. Like the texts in Groups A and B, these texts are also science reports. They are mostly between 1,000 and 3,000 words in length, with only one text shorter than 1,000 words and one longer than 3,000 words. The main consideration in collecting these sample texts from the same source was the consistency of the study, which focuses on the text organisation of a text type accessible to the general public. As for the characteristics of this text type, see Chapter 4 of this thesis. The date of publication and the length counted in words and sentences of the twenty texts are listed in Table 6.8 below.

Texts	*Date*	*Words*	*Sentences*
C1	07/07/1996	3,141	144
C2	14/07/1996	1,131	58
C3	21/07/1996	1,410	67
C4	28/07/1996	1,601	77
C5	04/08/1996	2,224	120
C6	11/08/1996	2,690	109
C7	18/08/1996	1,362	59
C8	25/08/1996	761	32
C9	01/09/1996	1,233	57
C10	08/09/1996	1,089	45
C11	15/09/1996	1,859	64
C12	22/09/1996	1,452	75
C13	29/09/1996	1,974	86
C14	06/10/1996	1,964	96
C15	13/10/1996	2,202	91
C16	20/10/1996	1,580	75
C17	27/10/1996	1,839	80
C18	03/11/1996	1,030	48
C19	10/11/1996	2,146	95
C20	17/11/1996	1,177	69

Table 6.8. Group C Sample Texts

6.5.3 Preparing Group C Sample Texts for analysis

As with the previous sample texts, the first step of the analysis was to delimit the Themes of each clause. Theme delimitation followed the same principle as before, which may be briefly summarised as being that Theme generally stops at the main verb of the main clause (cf. Chapters 4 and 5 of this thesis). The distribution of words in the Theme and Rheme areas in the twenty sample texts is shown in Table 6.9 below. For convenience of reference, the relevant data for the Group B texts are also included.

Texts	*All Words*			*%*		
	Theme	*Rheme*	*Overall*	*Theme*	*Rheme*	*Overall*
B1	554	1,005	1,559	35.54	64.46	100.00
B2	552	830	1,382	39.94	60.06	100.00
B3	507	963	1,470	34.49	65.51	100.00
B4	375	828	1,203	31.17	68.83	100.00
B5	796	1,326	2,122	37.51	62.49	100.00
B6	639	924	1,563	40.88	59.12	100.00
B7	436	762	1,198	36.39	63.61	100.00
B8	509	891	1,400	36.36	63.64	100.00
B9	474	732	1,206	39.30	60.70	100.00
B10	750	1,020	1,770	42.37	57.63	100.00
B11	640	1,175	1,815	35.26	64.74	100.00
C1	1,084	2,057	3,141	34.51	65.49	100.00
C2	420	711	1,131	37.14	62.86	100.00
C3	466	944	1,410	33.05	66.95	100.00
C4	589	1,012	1,601	36.79	63.21	100.00
C5	699	1,525	2,224	31.43	68.57	100.00
C6	972	1,718	2,690	36.13	63.87	100.00
C7	455	907	1,362	33.41	66.59	100.00
C8	310	451	761	40.74	59.26	100.00
C9	467	766	1,233	37.88	62.12	100.00
C10	335	754	1,089	30.76	69.24	100.00
C11	694	1,165	1,859	37.33	62.67	100.00
C12	634	818	1,452	43.66	56.34	100.00
C13	718	1,256	1,974	36.37	63.63	100.00
C14	661	1,303	1,964	33.66	66.34	100.00
C15	944	1,258	2,202	42.87	57.13	100.00
C16	499	1,081	1,580	31.58	68.42	100.00
C17	699	1,140	1,839	38.01	61.99	100.00
C18	356	674	1,030	34.56	65.44	100.00
C19	677	1,469	2,146	31.55	68.45	100.00
C20	394	783	1,177	33.47	66.53	100.00
Total	*18,305*	*32,248*	*50,553*	*36.21*	*63.79*	*100.00*

Table 6.9. The distribution of Theme and Rheme of Groups B and C Sample Texts

Table 6.9 shows that Theme and Rheme in each text roughly distribute in the proportion of 1:2, although with slightly more variation than the previous sample texts. For example, Texts C12 and C15 have a larger proportion of Theme than the other texts so far analysed. However, as will be shown later, the variation in length of Theme has little effect on the results of the analysis.

The next step was to make word lists out of the twenty Group C Sample Texts, using the 'WordList' program of the 'WordSmith' tools. Again three word lists were made for each sample text, namely the list of the whole text, the list of Theme, and the list of Rheme. Unlike in the analysis in the previous chapters, this time no separate files were made for each text, because it was found that the 'tag' function of the 'WordList' program allowed one to make the three word lists of each text without making separate files. The program could recognise which part of the text was to be analysed if a certain part was appropriately tagged. For example, the Theme area was tagged between <Th> and </Th>, and the Rheme area between <Rh> and </Rh>. Then the program was run on each text file three times, once only processing the part between <Th> and </Th>, once only the part between <Rh> and </Rh>, and once inclusively both parts of the text.

In addition, as with the other sample texts, a stopword list was used to filter out grammatical items from the word lists. Actually, as already noted in Chapter 5, the stopword list was unnecessary for making key-word lists, because most of the grammatical items would be automatically excluded from the key-word lists, owing to their high frequency both in the sample texts and the reference corpus. However, for this analysis, the stopword list was useful for the purpose of obtaining statistics of lexical items and lexical links of the sample texts. So the three word lists obtained for each sample text were composed of lexical items only.

When the three lists were made, they were compared, using the 'Consistency' function of the 'WordList' program, to form a list under three headings of 'Theme', 'Rheme' and 'Overall'. The word lists of Group C Sample Texts are in Appendix 10.

6.5.3.1 Lexical links

When the word lists were ready, I was in a position to start counting the lexical links of all the thirty-one sample texts both in Groups B and C. Unfortunately, manual counting as conducted previously proved to be rather unreliable, since mistakes tended to crop up, especially when I was processing a large number of texts. Therefore, an alternative approach was needed.

Adopting the alternative approach required more selectivity with the information in the analysis. In other words, some information which might be less essential to the analysis needed to be excluded. For example, as explained in Section 6.4, since T-R and R-T links run across the Theme and Rheme areas, they are strictly speaking not confined within the Theme or Rheme area. Because this chapter is more concerned with the distribution of links within the Theme and Rheme areas, it was decided not to consider these two types of links, since it was obvious in the previous analysis that R-T links comfortably outnumber T-R links. Had we included these two types of links, the links ending in Theme would have greatly outnumbered the links ending in Rheme.

So the exclusion of R-T links and T-R links would only make the job of supporting our hypotheses more demanding.

In addition, since intra-sentential links occupied only a very small proportion of the total links in the data, it was decided that these links should not be differentiated. Finally, since the purpose of this analysis was to obtain the number of links in the whole texts, there was no need for information about the sentence location of these links.

With the above considerations in mind, all that was needed at this point was to calculate the number of T-T links and R-R links formed by individual word tokens in the sample texts. For example, the word type 'apes' occurs in the Theme of Sentences 11, 38, 67, 68 and 74 of Sample Text B1 and consequently forms links between Sentences 11 and 38, 11 and 67, 11 and 68, etc. For this analysis, however, the sentence location of these links could be ignored. All it was needed to know was that the word occurred five times in the Theme area of the text and formed ten T-T links. A simple way to find the number of T-T links formed by this word in the Theme area of this text was using the formula:

$$L = \frac{n(n-1)}{2}$$

where L is the number of links we want to find, and n is the number of occurrences of the word in the text. Therefore, the number of links formed by the word 'apes' within the Theme of this text was calculated as '5×(5-1)÷2',

which makes 10, meaning ten T-T links. To sum up the number of links in the Theme and Rheme areas of the text, it was necessary to use another formula:

$$L = \sum_{i=1}^{m} \left\{ \frac{n_i (n_i - 1)}{2} \right\}$$

where L is the total number of links we are looking for, i is the series of word types, m is the total number of word types, and n is the number of tokens of each word type. All the links were counted in the same way to add up to the total number of links in the Theme and Rheme areas respectively. Thus, in effect, the procedure was to generate an approximation to the sentential link procedure described in the thesis so far.

Because the two groups of sample texts were analysed together, the number of lexical links in the Theme and Rheme areas of Groups B and C Sample Texts is presented together in Table 6.10 below.

Texts	*Lexical items*			*Lexical links*		
	Theme	*Rheme*	*Overall*	*T-T*	*R-R*	*Overall*
B1	242	503	745	165	221	386
B2	235	436	671	166	246	412
B3	212	558	770	154	326	480
B4	177	445	622	69	120	189
B5	342	627	969	305	865	1,170
B6	249	479	728	114	145	259
B7	223	397	620	133	172	305
B8	230	410	640	75	123	198
B9	227	381	608	109	204	313
B10	384	553	937	302	170	472
B11	276	577	853	353	423	776
C1	462	938	1,400	541	994	1,535
C2	211	351	562	95	80	175
C3	203	459	662	116	245	361
C4	243	488	731	331	330	661
C5	321	784	1,105	157	425	582
C6	429	813	1,242	369	761	1,130
C7	200	406	606	113	66	179
C8	164	231	395	34	59	93
C9	204	364	568	104	176	280
C10	146	345	491	53	108	161
C11	339	600	939	188	660	848
C12	308	427	735	414	148	562
C13	331	628	959	143	278	421
C14	327	654	981	335	324	659
C15	481	629	1,110	545	421	966
C16	230	548	778	118	460	578
C17	316	572	888	241	272	513
C18	197	344	541	135	229	364
C19	303	670	973	399	892	1,291
C20	179	381	560	53	101	154
Total	*8,391*	*15,998*	*24,389*	*6,429*	*10,044*	*16,473*

Table 6.10. Number of lexical items and lexical links in the Theme and Rheme areas

Table 6.10 does not include data of T-R links or R-T links, since we are here only concerned with the lexical links within the Theme and Rheme areas, not the lexical links across the two areas. Therefore the 'Overall' column under 'Lexical links' only represents the sum of T-T and R-R links in the sample texts, not all the four types of basic links.

6.5.3.2 Key words

After obtaining data concerning the lexical items and lexical links of the texts in Groups B and C, the next step of the analysis was to make key-word lists. Since data concerning the key words of Group B Sample Texts was already obtained in Chapter 5, it was now only necessary to make key-word lists of Group C Sample Texts on the basis of the word lists of the whole texts, using the KeyWords program with the same setting as before.

After the key-word lists were made, the distribution of each key word in Theme and Rheme was identified by comparing the key-word list with the word list of each of the twenty Group C Sample Texts, a procedure similar to that described in Section 6.4, where Sample Text A was processed. The numbers of key words in the Theme and Rheme areas of Group C Sample Texts are shown in Table 6.11 below. For the purpose of the analysis, data with regard to the key-words of Group B Sample Texts are also included.

Texts	*Key words*			*%*		
	Theme	*Rheme*	*Overall*	*Theme*	*Rheme*	*Overall*
B1	70	86	156	44.87	55.13	100.00
B2	74	82	156	47.44	52.56	100.00
B3	63	64	127	49.61	50.39	100.00
B4	43	54	97	44.33	55.67	100.00
B5	73	109	182	40.11	59.89	100.00
B6	49	59	108	45.37	54.63	100.00
B7	57	48	105	54.29	45.71	100.00
B8	42	37	79	53.16	46.84	100.00
B9	46	70	116	39.66	60.34	100.00
B10	43	32	75	57.33	42.67	100.00
B11	114	101	215	53.02	46.98	100.00
C1	168	207	375	44.80	55.20	100.00
C2	38	24	62	61.29	38.71	100.00
C3	46	54	100	46.00	54.00	100.00
C4	97	104	201	48.26	51.74	100.00
C5	86	131	217	39.63	60.37	100.00
C6	115	165	280	41.07	58.93	100.00
C7	46	23	69	66.67	33.33	100.00
C8	29	32	61	47.54	52.46	100.00
C9	54	58	112	48.21	51.79	100.00
C10	27	35	62	43.55	56.45	100.00
C11	72	159	231	31.17	68.83	100.00
C12	78	56	134	58.21	41.79	100.00
C13	53	72	125	42.40	57.60	100.00
C14	97	106	203	47.78	52.22	100.00
C15	147	120	267	55.06	44.94	100.00
C16	76	132	208	36.54	63.46	100.00
C17	89	96	185	48.11	51.89	100.00
C18	29	39	68	42.65	57.35	100.00
C19	105	151	256	41.02	58.98	100.00
C20	11	15	26	42.31	57.69	100.00
Total	*2,137*	*2,521*	*4,658*	*45.88*	*54.12*	*100.00*

Table 6.11. Number of key words in the Theme and Rheme areas of Groups B and C Sample Texts

6.5.3.3 Key-word links

The next step was to calculate the key-word links of all the thirty-one sample texts. Each text was processed individually. To calculate the links formed by all the key words in the Theme and Rheme areas of each text, the formula $L = \sum_{i=1}^{m} \left\{ \frac{n_i(n_i - 1)}{2} \right\}$ was used, where L is the number of key-word links being

sought. In this way, the total number of T-T and R-R links formed by key words in the text was achieved. To present an example, Table 6.12 below shows the key words and number of T-T and R-R links formed by the key words in Sample Text B1.

		Key words			*Key-word links*		
N	*Words*	*Theme*	*Rheme*	*Overall*	*T-T*	*R-R*	*Overall*
1	Intelligence	7	9	16	21	36	57
2	Apes	5	5	10	10	10	20
3	Byrne	9	0	9	36	0	36
4	Social	3	6	9	3	15	18
5	Gorillas	4	3	7	6	3	9
6	Chimps	3	4	7	3	6	9
7	Ape	4	2	6	6	1	7
8	Human	4	2	6	6	1	7
9	Orangutans	4	2	6	6	1	7
10	Groups	2	4	6	1	6	7
11	Species	2	4	6	1	6	7
12	Complex	0	6	6	0	15	15
13	Ancestors	3	2	5	3	1	4
14	Primates	3	2	5	3	1	4
15	Humans	2	3	5	1	3	4
16	Primate	2	3	5	1	3	4
17	Behaviour	1	4	5	0	6	6
18	Understanding	0	5	5	0	10	10
19	Orangutan	3	1	4	3	0	3
20	Monkeys	2	2	4	1	1	2
21	Sheep	2	2	4	1	1	2
22	Animals	1	3	4	0	3	3
23	Requires	0	4	4	0	6	6
24	Borneo	2	1	3	1	0	1
25	Evolution	1	2	3	0	1	1
26	Imitation	1	2	3	0	1	1
27	Evolved	0	3	3	0	3	3
	Total	*70*	*86*	*156*	*113*	*140*	*253*

Table 6.12. Key words and key-word links in the Theme and Rheme areas of Sample Text B1

Note that T-R and R-T links are excluded from Table 6.12, for the reasons given above. Therefore the 'Overall' column under 'Key-word links' does not contain all the four types of links in the sample texts, but only contains data of T-T and R-R links. The frequencies of key words and their links in the Theme

and Rheme areas of all the thirty-one Groups B and C Sample Texts are shown in Table 6.13 below.

Texts	*Key words*			*Key-word links*		
	Theme	*Rheme*	*Overall*	*T-T*	*R-R*	*Overall*
B1	70	86	156	113	140	253
B2	74	82	156	133	179	312
B3	63	64	127	112	127	239
B4	43	54	97	46	57	103
B5	73	109	182	229	656	885
B6	49	59	108	76	83	159
B7	57	48	105	97	95	192
B8	42	37	79	59	47	106
B9	46	70	116	86	148	234
B10	43	32	75	155	52	207
B11	114	101	215	330	264	594
C1	168	207	375	490	732	1,222
C2	38	24	62	75	38	113
C3	46	54	100	101	148	249
C4	97	104	201	311	266	577
C5	86	131	217	116	231	347
C6	115	165	280	304	492	796
C7	46	23	69	97	33	130
C8	29	32	61	18	43	61
C9	54	58	112	87	121	208
C10	27	35	62	46	52	98
C11	72	159	231	114	565	679
C12	78	56	134	369	71	440
C13	53	72	125	82	148	230
C14	97	106	203	275	185	460
C15	147	120	267	462	289	751
C16	76	132	208	97	298	395
C17	89	96	185	204	168	372
C18	29	39	68	105	158	263
C19	105	151	256	361	733	1,094
C20	11	15	26	9	13	22
Total	*2,137*	*2,521*	*4,658*	*5,159*	*6,632*	*11,791*

Table 6.13. Key words and key-word links in the Theme and Rheme areas of Groups B and C Sample Texts

Having obtained data of key words and key-word links in the Theme and Rheme areas of the thirty-one texts in Groups B and C, I was now in a position to test the seven hypotheses outlined in Section 6.3.

6.5.4 Testing Hypotheses 6.1 to 6.3

To test Hypotheses 6.1 to 6.3, I needed first to work out the ratios of key-word links to key words and lexical links to lexical items in the Theme and Rheme areas and in the whole texts, and then to compare the ratios. The ratios of key-word links to key words was obtained by calculating the data from Table 6.13 above, and the ratios of lexical links to lexical items was obtained by calculating the data from Table 6.10 above. These ratios are presented in Table 6.14 below.

Texts	*Theme*		*Rheme*		*Overall*	
	KL/KW	*LL/LW*	*KL/KW*	*LL/LW*	*KL/KW*	*LL/LW*
B1	1.6143	0.6818	1.6279	0.4394	1.6218	0.5181
B2	1.7973	0.7064	2.1829	0.5642	2.0000	0.6140
B3	1.7778	0.7264	1.9844	0.5842	1.8819	0.6234
B4	1.0698	0.3898	1.0556	0.2697	1.0619	0.3039
B5	3.1370	0.8918	6.0183	1.3796	4.8626	1.2074
B6	1.5510	0.4578	1.4068	0.3027	1.4722	0.3558
B7	1.7018	0.5964	1.9792	0.4332	1.8286	0.4919
B8	1.4048	0.3261	1.2703	0.3000	1.3418	0.3094
B9	1.8696	0.4802	2.1143	0.5354	2.0172	0.5148
B10	3.6047	0.7865	1.6250	0.3074	2.7600	0.5037
B11	2.8947	1.2790	2.6139	0.7331	2.7628	0.9097
C1	2.9167	1.1710	3.5362	1.0597	3.2587	1.0964
C2	1.9737	0.4502	1.5833	0.2279	1.8226	0.3114
C3	2.1957	0.5714	2.7407	0.5338	2.4900	0.5453
C4	3.2062	1.3621	2.5577	0.6762	2.8706	0.9042
C5	1.3488	0.4891	1.7634	0.5421	1.5991	0.5267
C6	2.6435	0.8601	2.9818	0.9360	2.8429	0.9098
C7	2.1087	0.5650	1.4348	0.1626	1.8841	0.2954
C8	0.6207	0.2073	1.3438	0.2554	1.0000	0.2354
C9	1.6111	0.5098	2.0862	0.4835	1.8571	0.4930
C10	1.7037	0.3630	1.4857	0.3130	1.5806	0.3279
C11	1.5833	0.5546	3.5535	1.1000	2.9394	0.9031
C12	4.7308	1.3442	1.2679	0.3466	3.2836	0.7646
C13	1.5472	0.4320	2.0556	0.4427	1.8400	0.4390
C14	2.8351	1.0245	1.7453	0.4954	2.2660	0.6718
C15	3.1429	1.1331	2.4083	0.6693	2.8127	0.8703
C16	1.2763	0.5130	2.2576	0.8394	1.8990	0.7429
C17	2.2921	0.7627	1.7500	0.4755	2.0108	0.5777
C18	3.6207	0.6853	4.0513	0.6657	3.8676	0.6728
C19	3.4381	1.3168	4.8543	1.3313	4.2734	1.3268
C20	0.8182	0.2961	0.8667	0.2651	0.8462	0.2750
Total	*2.1947*	*0.7075*	*2.2646*	*0.5700*	*2.2857*	*0.6207*

Table 6.14. Comparison of the ratios of key-word links to key words and the ratios of lexical links to lexical items in Groups B and C Sample Texts

In Table 6.14, KL means key-word links, KW means key words, LL means lexical links, and LW means lexical items. Table 6.14 shows that, without exception, the overall ratios of key-word links to key words in all the thirty-one Groups B and C Sample Texts are higher than the ratios of lexical links to lexical items; this is true for the ratios in both the Theme and Rheme areas. This can be seen clearly in Figure 6.9, Figure 6.10 and Figure 6.11 below.

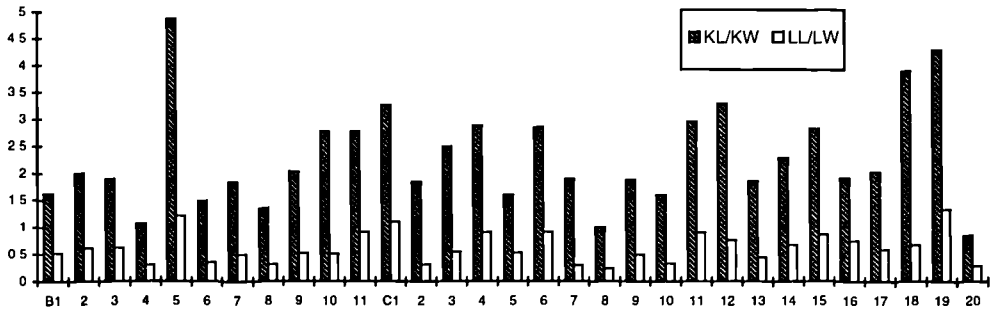


Figure 6.9. Overall key-word link/key word ratios compared with lexical link/lexical item ratios in Groups B and C Sample Texts

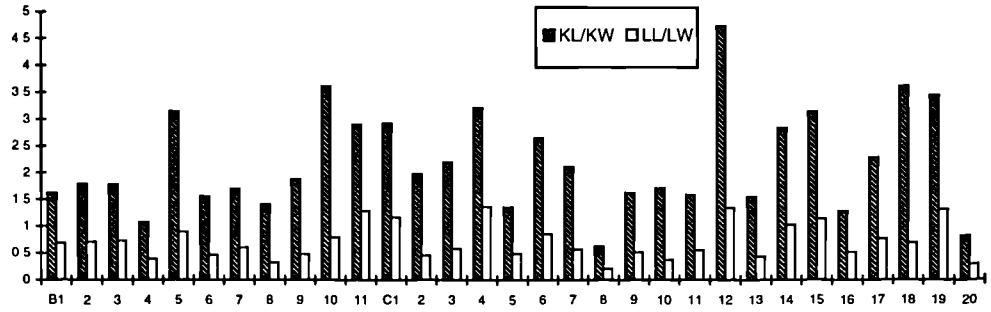


Figure 6.10. Key-word link/key word ratios compared with lexical link/lexical item ratios in the Themes of Groups B and C Sample Texts

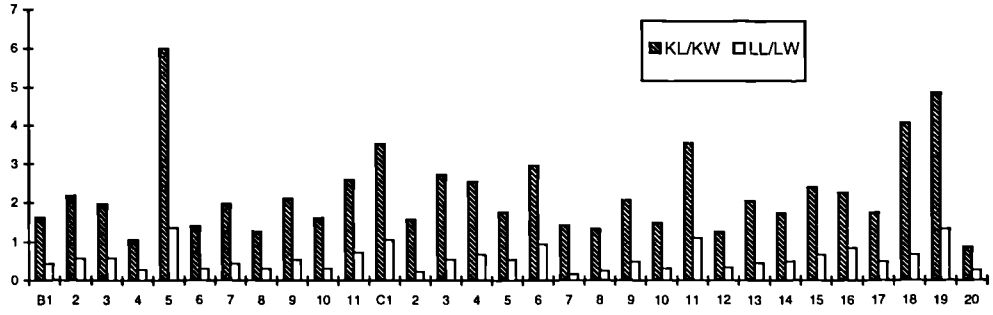


Figure 6.11. Key-word link/key word ratios compared with lexical link/lexical item ratios in the Rhemes of Groups B and C Sample Texts

Figure 6.9, Figure 6.10 and Figure 6.11 clearly show that the overall ratios of key-word links to key words are higher than the ratios of lexical links to lexical items in the whole text, in the Theme area, and in the Rheme area of all the thirty-one Groups B and C Sample Texts, respectively. To test the statistical significance of the differences in ratios, a *t*-test was carried out on the 3 groups of ratios. The *t*-test results gave P-values of 0.0000000000002 for the whole texts, 0.0000000000209 for Theme, and 0.0000000000502 for Rheme, suggesting that the differences in ratios are statistically highly significant. This confirms the initial analysis of Sample Text A earlier in this chapter, and firmly supports Hypotheses 6.1 to 6.3.

6.5.5 Testing Hypotheses 6.4 to 6.6

In order to test Hypotheses 6.4 to 6.6, I first needed to calculate the ratios of key words to all words and the ratios of key-word links to all words in the thirty-one Groups B and C Sample Texts. On the basis of Table 6.9 and Table 6.11 above, it was possible to calculate the ratios of key words to all words in the texts, and on the basis of Table 6.9 and Table 6.14 above, it was possible to calculate the ratios of key-word links to all words in the texts. Table 6.15 and Table 6.16 below present the results of these calculations.

Texts	*KW*			*TL*			*KW/TL*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
B1	70	86	156	554	1,005	1,559	0.1264	0.0856	0.1001
B2	74	82	156	552	830	1,382	0.1341	0.0988	0.1129
B3	63	64	127	507	963	1,470	0.1243	0.0665	0.0864
B4	43	54	97	375	828	1,203	0.1147	0.0652	0.0806
B5	73	109	182	796	1,326	2,122	0.0917	0.0822	0.0858
B6	49	59	108	639	924	1,563	0.0767	0.0639	0.0691
B7	57	48	105	436	762	1,198	0.1307	0.0630	0.0876
B8	42	37	79	509	891	1,400	0.0825	0.0415	0.0564
B9	46	70	116	474	732	1,206	0.0970	0.0956	0.0962
B10	43	32	75	750	1,020	1,770	0.0573	0.0314	0.0424
B11	114	101	215	640	1,175	1,815	0.1781	0.0860	0.1185
C1	168	207	375	1,084	2,057	3,141	0.1550	0.1006	0.1194
C2	38	24	62	420	711	1,131	0.0905	0.0338	0.0548
C3	46	54	100	466	944	1,410	0.0987	0.0572	0.0709
C4	97	104	201	589	1,012	1,601	0.1647	0.1028	0.1255
C5	86	131	217	699	1,525	2,224	0.1230	0.0859	0.0976
C6	115	165	280	972	1,718	2,690	0.1183	0.0960	0.1041
C7	46	23	69	455	907	1,362	0.1011	0.0254	0.0507
C8	29	32	61	310	451	761	0.0935	0.0710	0.0802
C9	54	58	112	467	766	1,233	0.1156	0.0757	0.0908
C10	27	35	62	335	754	1,089	0.0806	0.0464	0.0569
C11	72	159	231	694	1,165	1,859	0.1037	0.1365	0.1243
C12	78	56	134	634	818	1,452	0.1230	0.0685	0.0923
C13	53	72	125	718	1,256	1,974	0.0738	0.0573	0.0633
C14	97	106	203	661	1,303	1,964	0.1467	0.0814	0.1034
C15	147	120	267	944	1,258	2,202	0.1557	0.0954	0.1213
C16	76	132	208	499	1,081	1,580	0.1523	0.1221	0.1316
C17	89	96	185	699	1,140	1,839	0.1273	0.0842	0.1006
C18	29	39	68	356	674	1,030	0.0815	0.0579	0.0660
C19	105	151	256	677	1,469	2,146	0.1551	0.1028	0.1193
C20	11	15	26	394	783	1,177	0.0279	0.0192	0.0221
Total	*2,137*	*2,521*	*4,658*	*18,305*	*32,248*	*50,553*	*0.1130*	*0.0742*	*0.0881*

Table 6.15. Ratios of key words to all words in Groups B and C Sample Texts

Texts	*KL*			*TL*			*KL/TL*		
	T-T	*R-R*	*Overall*	*Th*	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
B1	113	140	253	554	1,005	1,559	0.2040	0.1393	0.1623
B2	133	179	312	552	830	1,382	0.2409	0.2157	0.2258
B3	112	127	239	507	963	1,470	0.2209	0.1319	0.1626
B4	46	57	103	375	828	1,203	0.1227	0.0688	0.0856
B5	229	656	885	796	1,326	2,122	0.2877	0.4947	0.4171
B6	76	83	159	639	924	1,563	0.1189	0.0898	0.1017
B7	97	95	192	436	762	1,198	0.2225	0.1247	0.1603
B8	59	47	106	509	891	1,400	0.1159	0.0527	0.0757
B9	86	148	234	474	732	1,206	0.1814	0.2022	0.1940
B10	155	52	207	750	1,020	1,770	0.2067	0.0510	0.1169
B11	330	264	594	640	1,175	1,815	0.5156	0.2247	0.3273
C1	490	732	1,222	1,084	2,057	3,141	0.4520	0.3559	0.3890
C2	75	38	113	420	711	1,131	0.1786	0.0534	0.0999
C3	101	148	249	466	944	1,410	0.2167	0.1568	0.1766
C4	311	266	577	589	1,012	1,601	0.5280	0.2628	0.3604
C5	116	231	347	699	1,525	2,224	0.1660	0.1515	0.1560
C6	304	492	796	972	1,718	2,690	0.3128	0.2864	0.2959
C7	97	33	130	455	907	1,362	0.2132	0.0364	0.0954
C8	18	43	61	310	451	761	0.0581	0.0953	0.0802
C9	87	121	208	467	766	1,233	0.1863	0.1580	0.1687
C10	46	52	98	335	754	1,089	0.1373	0.0690	0.0900
C11	114	565	679	694	1,165	1,859	0.1643	0.4850	0.3653
C12	369	71	440	634	818	1,452	0.5820	0.0868	0.3030
C13	82	148	230	718	1,256	1,974	0.1142	0.1178	0.1165
C14	275	185	460	661	1,303	1,964	0.4160	0.1420	0.2342
C15	462	289	751	944	1,258	2,202	0.4894	0.2297	0.3411
C16	97	298	395	499	1,081	1,580	0.1944	0.2757	0.2500
C17	204	168	372	699	1,140	1,839	0.2918	0.1474	0.2023
C18	105	158	263	356	674	1,030	0.2949	0.2344	0.2553
C19	361	733	1,094	677	1,469	2,146	0.5332	0.4990	0.5098
C20	9	13	22	394	783	1,177	0.0228	0.0166	0.0187
Total	*5,159*	*6,632*	*11,791*	*18,305*	*32,248*	*50,553*	*0.2577*	*0.1824*	*0.2109*

Table 6.16. Ratios of key-word links to all words in Groups B and C Sample Texts

In Table 6.15 and Table 6.16, KW means key words, KL means key-word links, and TL means the length of text areas in words. A comparison was then made of the ratios of key-word links to all words and the ratios of key words to all words in the Theme and Rheme areas and in the whole texts as presented in Table 6.15 and Table 6.16 above. The results of this comparison are shown in Table 6.17 below.

Texts	*Theme*		*Rheme*		*Overall*	
	KL/TL	*KW/TL*	*KL/TL*	*KW/TL*	*KL/TL*	*KW/TL*
B1	0.2040	0.1264	0.1393	0.0856	0.1623	0.1001
B2	0.2409	0.1341	0.2157	0.0988	0.2258	0.1129
B3	0.2209	0.1243	0.1319	0.0665	0.1626	0.0864
B4	0.1227	0.1147	0.0688	0.0652	0.0856	0.0806
B5	0.2877	0.0917	0.4947	0.0822	0.4171	0.0858
B6	0.1189	0.0767	0.0898	0.0639	0.1017	0.0691
B7	0.2225	0.1307	0.1247	0.0630	0.1603	0.0876
B8	0.1159	0.0825	0.0527	0.0415	0.0757	0.0564
B9	0.1814	0.0970	0.2022	0.0956	0.1940	0.0962
B10	0.2067	0.0573	0.0510	0.0314	0.1169	0.0424
B11	0.5156	0.1781	0.2247	0.0860	0.3273	0.1185
C1	0.4520	0.1550	0.3559	0.1006	0.3890	0.1194
C2	0.1786	0.0905	0.0534	0.0338	0.0999	0.0548
C3	0.2167	0.0987	0.1568	0.0572	0.1766	0.0709
C4	0.5280	0.1647	0.2628	0.1028	0.3604	0.1255
C5	0.1660	0.1230	0.1515	0.0859	0.1560	0.0976
C6	0.3128	0.1183	0.2864	0.0960	0.2959	0.1041
C7	0.2132	0.1011	0.0364	0.0254	0.0954	0.0507
C8	0.0581	0.0935	0.0953	0.0710	0.0802	0.0802
C9	0.1863	0.1156	0.1580	0.0757	0.1687	0.0908
C10	0.1373	0.0806	0.0690	0.0464	0.0900	0.0569
C11	0.1643	0.1037	0.4850	0.1365	0.3653	0.1243
C12	0.5820	0.1230	0.0868	0.0685	0.3030	0.0923
C13	0.1142	0.0738	0.1178	0.0573	0.1165	0.0633
C14	0.4160	0.1467	0.1420	0.0814	0.2342	0.1034
C15	0.4894	0.1557	0.2297	0.0954	0.3411	0.1213
C16	0.1944	0.1523	0.2757	0.1221	0.2500	0.1316
C17	0.2918	0.1273	0.1474	0.0842	0.2023	0.1006
C18	0.2949	0.0815	0.2344	0.0579	0.2553	0.0660
C19	0.5332	0.1551	0.4990	0.1028	0.5098	0.1193
C20	0.0228	0.0279	0.0166	0.0192	0.0187	0.0221
Total	*0.2577*	*0.1130*	*0.1824*	*0.0742*	*0.2109*	*0.0881*

Table 6.17. Comparison of ratios of key-word links to all words and ratios of key words to all words in Groups B and C Sample Texts

Table 6.17 shows a general trend of higher ratios in the Themes than in the Rhemes of the texts. However, there are a few exceptional cases. For convenience of identification of those cases, in Table 6.17 and henceforth, the exceptional cases are marked by underlining in the table. For example, in Table 6.17, wherever the ratio of key-word links to key words is lower than the ratio of key words to all words, the number is underlined. The results in Table 6.17 are clearly shown in Figure 6.12, Figure 6.13 and Figure 6.14 below.

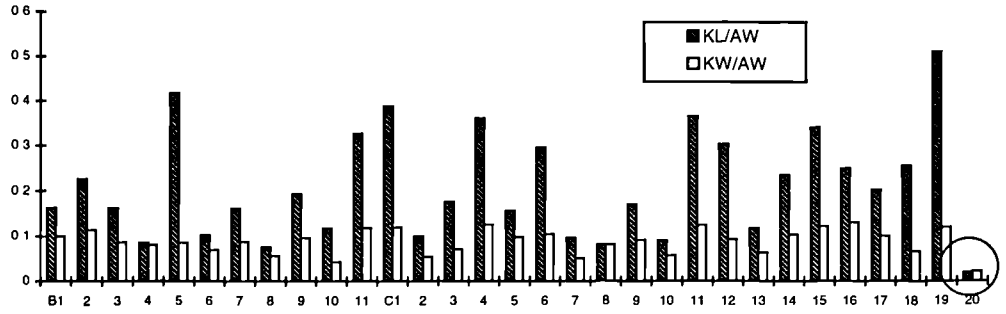


Figure 6.12. Comparison of ratios of key-word links to all words and ratios of key words to all words in Groups B and C Sample Texts

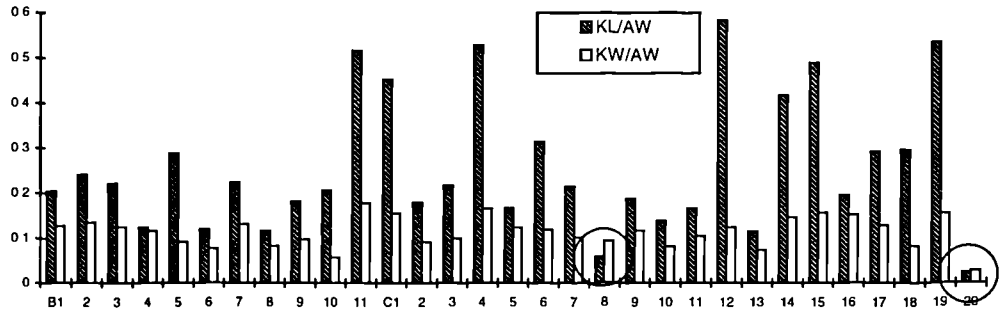


Figure 6.13. Comparison of ratios of key-word links to all words and ratios of key words to all words in the Theme area of Groups B and C Sample Texts

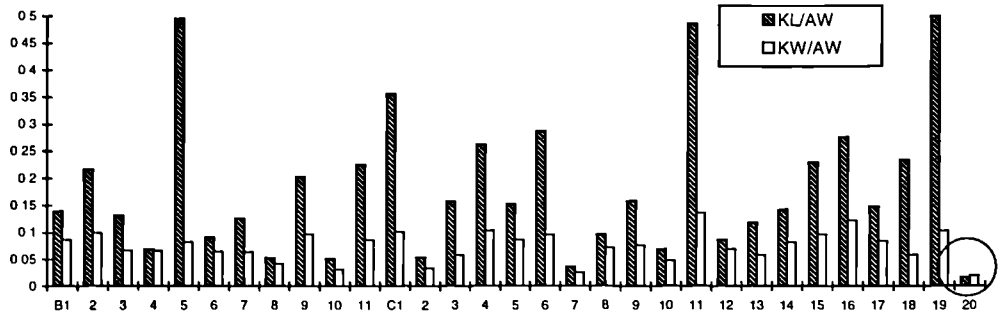


Figure 6.14. Comparison of ratios of key-word links to all words and ratios of key words to all words in the Rheme area of Groups B and C Sample Texts

Figure 6.12, Figure 6.13 and Figure 6.14 show that the ratios of key-word links to all words are higher than the ratios of key words to all words in the whole texts as well as in the Theme and Rheme areas of almost all the thirty-one sample texts, with only a few exceptional cases. For ease of identification, the exceptional cases are marked by circling the representing columns. The exceptions involve only two texts. In the Theme area of Sample Text C8, the ratio of key-words to all words is slightly higher than the ratios of key-word links to key words. In the Theme and Rheme areas, as well as in the whole text of Sample Text C20, the ratios of key words to all words are slightly higher than the ratios of key-word links to key words. It should be remembered that only links within the Theme and Rheme areas, not links across the two areas, were counted. So it is possible in these few cases that the number of key-word links was smaller than the number of key words. This could happen when most of the key word types in the Theme area or in the Rheme area were repeated fewer than three times, since one item forms no link, two items form only one link, and three items form three links. However, in spite of these exceptions, the *t*-test results showed that the differences of the ratios are again statistically highly significant, with P-values of 0.00000242 in Theme, 0.00000001 in Rheme, and 0.00000056 in the whole text. This means that Hypothesis 6.4, Hypothesis 6.5 and Hypothesis 6.6 of this chapter are all well supported.

6.5.6 Testing Hypothesis 6.7

Hypothesis 6.7 of this chapter is that 'the ratio of key-word links to all words will be higher in the Themes than in the Rhemes of the text.' In order to test this hypothesis, it was necessary to compare the ratios of key-word links to all words in the Theme and Rheme areas of the sample texts. These ratios are presented in Table 6.17 above, but the overall picture is presented more clearly in Figure 6.15 below.

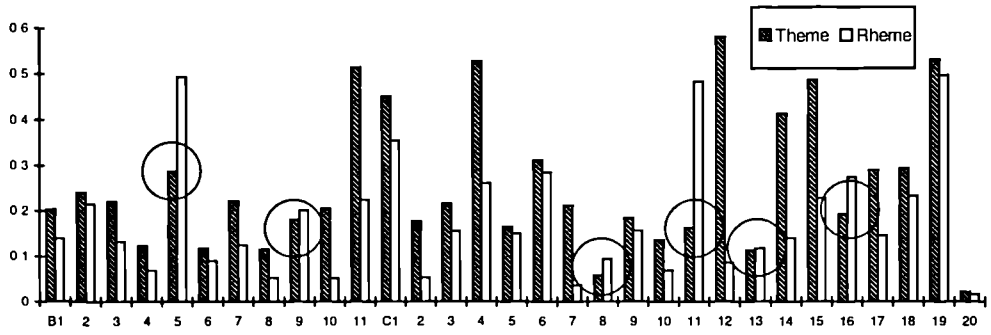


Figure 6.15. Ratios of key-word links to all words in the Theme and Rheme areas of Groups B and C Sample Texts

Figure 6.15 shows that there were six counter examples, where in Texts B5, B9, C8, C11, C13 and C16 the ratios of key-word links to all words are higher in Rheme than in Theme. In addition, the *t*-test returned a P-value of 0.041415, which is only marginally acceptable. There is apparently something wrong, either with the hypothesis itself or with the way it has been tested. This problem will be discussed in the next section.

6.6 Further analysis

6.6.1 Revising Hypothesis 6.7

Analogous to Hasan's (1984) notion of 'cohesive chains' in examining cohesion in the text, it may be assumed that one measurement of the centrality of a word to the 'aboutness' of a text may be the degree of concentration of cohesive links in the vicinity of which this word occurs. In a fixed length of text, if there are more cohesive links, these links will be more concentrated; if there are fewer cohesive links, these links will be less concentrated. This assumption is partly supported by the analysis in Section 6.5.5, where it was found that key words generally form more links than other lexical items. Then, if Theme is a more important area than Rheme in terms of clues to the information organisation of the text, and if key words are indicative of what the text is about, it may be hypothesised that the key words of a text will form proportionally more links in Theme rather than in Rheme. This leads to the notion of *key-word link density,* which is the ratio of the sum of key-word links to the sum of key words in the Theme or Rheme of a text. So, instead of comparing the ratios of key-word links to all words, it is better to compare key-word link density to all words in the Theme and Rheme areas of the text. Then, Hypothesis 6.7 may be revised as

Hypothesis 6.7a. The ratio of key-word link density to all words will be higher in the Themes than in the Rhemes of the text.

To test Hypothesis 6.7a, it was necessary firstly to obtain the data concerning key-word link density of the sample texts; that is, ratios of key-word links to key words in the Theme and Rheme areas of the sample texts. These ratios were calculated for the tests of Hypotheses 6.1 to 6.3 and were presented in Table 6.13 and Table 6.14. For ease of reference, relevant data from Table 6.13 and Table 6.14 are represented in Table 6.18 below.

Texts	*KL*			*KW*			*KLD*		
	T-T	*R-R*	*Overall*	*Th*	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
B1	113	140	253	70	86	156	1.6143	1.6279	1.6218
B2	133	179	312	74	82	156	1.7973	2.1829	2.0000
B3	112	127	239	63	64	127	1.7778	1.9844	1.8819
B4	46	57	103	43	54	97	1.0698	1.0556	1.0619
B5	229	656	885	73	109	182	3.1370	6.0183	4.8626
B6	76	83	159	49	59	108	1.5510	1.4068	1.4722
B7	97	95	192	57	48	105	1.7018	1.9792	1.8286
B8	59	47	106	42	37	79	1.4048	1.2703	1.3418
B9	86	148	234	46	70	116	1.8696	2.1143	2.0172
B10	155	52	207	43	32	75	3.6047	1.6250	2.7600
B11	330	264	594	114	101	215	2.8947	2.6139	2.7628
C1	490	732	1,222	168	207	375	2.9167	3.5362	3.2587
C2	75	38	113	38	24	62	1.9737	1.5833	1.8226
C3	101	148	249	46	54	100	2.1957	2.7407	2.4900
C4	311	266	577	97	104	201	3.2062	2.5577	2.8706
C5	116	231	347	86	131	217	1.3488	1.7634	1.5991
C6	304	492	796	115	165	280	2.6435	2.9818	2.8429
C7	97	33	130	46	23	69	2.1087	1.4348	1.8841
C8	18	43	61	29	32	61	0.6207	1.3438	1.0000
C9	87	121	208	54	58	112	1.6111	2.0862	1.8571
C10	46	52	98	27	35	62	1.7037	1.4857	1.5806
C11	114	565	679	72	159	231	1.5833	3.5535	2.9394
C12	369	71	440	78	56	134	4.7308	1.2679	3.2836
C13	82	148	230	53	72	125	1.5472	2.0556	1.8400
C14	275	185	460	97	106	203	2.8351	1.7453	2.2660
C15	462	289	751	147	120	267	3.1429	2.4083	2.8127
C16	97	298	395	76	132	208	1.2763	2.2576	1.8990
C17	204	168	372	89	96	185	2.2921	1.7500	2.0108
C18	105	158	263	29	39	68	3.6207	4.0513	3.8676
C19	361	733	1,094	105	151	256	3.4381	4.8543	4.2734
C20	9	13	22	11	15	26	0.8182	0.8667	0.8462
Total	*5159*	*6632*	*11,791*	*2137*	*2521*	*4658*	*2.1947*	*2.2646*	*2.2857*

Table 6.18. Key-word link density in Groups B and C Sample Texts

In Table 6.18, KLD means key-word link density, or the ratio of key-word links to key words. It may be noticed in Table 6.18 that many of the ratios are higher in the Rheme area than in the Theme area, or the key-word density is higher in the Rheme area than in the Theme area. This is because the Rhemes tend to account for a larger proportion of a text than do the Themes in the sample texts. To test Hypothesis 6.7a, these ratios need therefore to be further calculated on the basis of the Theme and Rheme context. See Table 6.19 below for the results of the calculation.

Texts	*KLD*			*TL*			*KLD/TL*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
B1	1.6143	1.6279	1.6218	554	1,005	1,559	2.9139	1.6198	1.0403
B2	1.7973	2.1829	2.0000	552	830	1,382	3.2560	2.6300	1.4472
B3	1.7778	1.9844	1.8819	507	963	1,470	3.5065	2.0606	1.2802
B4	1.0698	1.0556	1.0619	375	828	1,203	2.8528	1.2749	0.8827
B5	3.1370	6.0183	4.8626	796	1,326	2,122	3.9410	4.5387	2.2915
B6	1.5510	1.4068	1.4722	639	924	1,563	2.4272	1.5225	0.9419
B7	1.7018	1.9792	1.8286	436	762	1,198	3.9032	2.5974	1.5264
B8	1.4048	1.2703	1.3418	509	891	1,400	2.7599	1.4257	0.9584
B9	1.8696	2.1143	2.0172	474	732	1,206	3.9443	2.8884	1.6726
B10	3.6047	1.6250	2.7600	750	1,020	1,770	4.8063	1.5931	1.5593
B11	2.8947	2.6139	2.7628	640	1,175	1,815	4.5230	2.2246	1.5222
C1	2.9167	3.5362	3.2587	1,084	2,057	3,141	2.6907	1.7191	1.0375
C2	1.9737	1.5833	1.8226	420	711	1,131	4.6993	2.2269	1.6115
C3	2.1957	2.7407	2.4900	466	944	1,410	4.7118	2.9033	1.7660
C4	3.2062	2.5577	2.8706	589	1,012	1,601	5.4435	2.5274	1.7930
C5	1.3488	1.7634	1.5991	699	1,525	2,224	1.9296	1.1563	0.7190
C6	2.6435	2.9818	2.8429	972	1,718	2,690	2.7197	1.7356	1.0568
C7	2.1087	1.4348	1.8841	455	907	1,362	4.6345	1.5819	1.3833
C8	0.6207	1.3438	1.0000	310	451	761	2.0023	2.9796	1.3141
C9	1.6111	2.0862	1.8571	467	766	1,233	3.4499	2.7235	1.5062
C10	1.7037	1.4857	1.5806	335	754	1,089	5.0857	1.9704	1.4514
C11	1.5833	3.5535	2.9394	694	1,165	1,859	2.2814	3.0502	1.5812
C12	4.7308	1.2679	3.2836	634	818	1,452	7.4618	1.5500	2.2614
C13	1.5472	2.0556	1.8400	718	1,256	1,974	2.1549	1.6366	0.9321
C14	2.8351	1.7453	2.2660	661	1,303	1,964	4.2891	1.3394	1.1538
C15	3.1429	2.4083	2.8127	944	1,258	2,202	3.3293	1.9144	1.2773
C16	1.2763	2.2576	1.8990	499	1,081	1,580	2.5577	2.0884	1.2019
C17	2.2921	1.7500	2.0108	699	1,140	1,839	3.2791	1.5351	1.0934
C18	3.6207	4.0513	3.8676	356	674	1,030	10.1705	6.0108	3.7550
C19	3.4381	4.8543	4.2734	677	1,469	2,146	5.0784	3.3045	1.9913
C20	0.8182	0.8667	0.8462	394	783	1,177	2.0766	1.1069	0.7189
Total	*2.1947*	*2.2646*	*2.2857*	*18,305*	*32,248*	*50,553*	*3.8348*	*2.2399*	*1.4428*

Table 6.19. Ratios of key-word link density to all words in Groups B and C Sample Texts

The results are schematically shown in Figure 6.16 below.

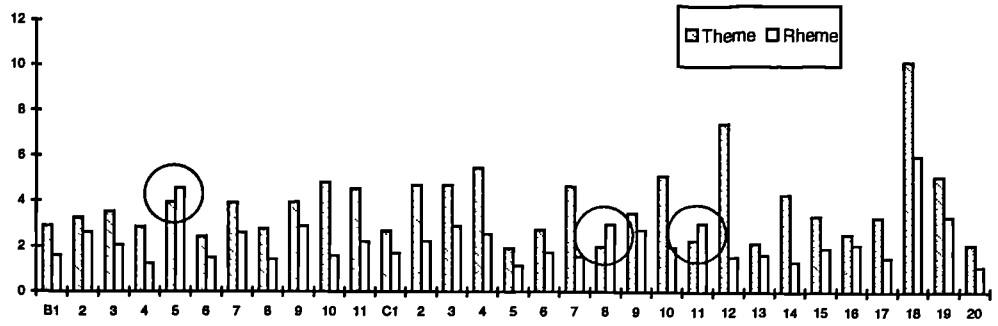


Figure 6.16. Ratios of key-word link density to all words in the Theme and Rheme areas of Groups B and C Sample Texts

Figure 6.16 shows that there is a very strong tendency for the key-word link density to be higher proportionally in Theme than in Rheme. The *t*-test results show that the differences between the ratios of key-word link density to all words in the Theme and Rheme areas are statistically highly significant, with a P-value of 0.00003998. This suggests that Hypothesis 6.7a is supported by the analysis. However, there are still three exceptional cases: in Texts B5, C8 and C11, the ratios are higher in Rheme than Theme. This phenomenon will be discussed in the next sub-section.

6.6.1.1 Realised links versus potential links

The three exceptional cases where the ratios of key-word link density to all words were higher in Rheme than Theme may suggest that despite the fact that the results of the test have shown Hypothesis 6.7a to be supported, the way it was tested may still not be the most suitable for our purpose. Since it is links

that are being compared, the calculation of the ratios should also have been based on links instead of on word tokens.

Here it is necessary to introduce another notion, that of *potential links*, which is the maximum number of links possibly formed by all words in the Theme area or in the Rheme area of the text. This number could be obtained by using the formula $L = \frac{n(n-1)}{2}$, where L is the number of links possibly formed by all words in Theme or Rheme of the text, and n is the total number of word tokens in the relevant areas of the text.

To justify the adoption of this notion, it should be noted that in a fixed size of context, if the number of key words increased, the number of key-word links increased at a much higher rate, and consequently the ratio between key-word links and all words in the context would change drastically. However, the ratio between the number of key-word links and the number of potential links formed by all words would remain fairly close to the ratio between the number of key words and the number of all words in the context. Therefore, when computing the ratio of key-word links to the context in which the key-word links occur, the results will be more reliable if we make a link-to-link comparison instead of a link-to-word comparison.

This would be better illustrated with an example. Suppose a key word, say 'ape', occurs 5 times in Theme and 10 times in Rheme. The difference between its occurrence in Theme and Rheme is 1:2. If the Theme contains 50 words and

Rheme contains 100 words, then the difference between the lengths of Theme and Rheme is also 1:2; the ratio is the same. So, the key word 'ape' may be regarded as evenly distributed in the Theme and Rheme areas. However, comparing the two areas in terms of links, this key word forms 10 links in Theme and 45 links in Rheme, with the difference increased to 4.5 times in favour of Rheme. If the calculation was still based on all words in Theme and Rheme, it would be concluded that this key word made proportionally more links in Rheme than in Theme. This is certainly a distorted picture. For a fair comparison, it would be necessary to compare the key-word links on the basis of all potential links, instead of all words, in the Theme and Rheme areas. In this instance, it would be necessary to calculate how many links the 50 words in Theme and 00 words in Rheme could possibly form. With the formula $L = \frac{n(n-1)}{2}$ where L is the total number of potential links and n is the total number of words in Theme or Rheme, a calculation would obtain the results that 1,225 potential links could be formed by the 50 words in Theme but about four times as many potential links (4,950) could be formed by the 100 words in Rheme. The proportion of potential links in the Theme area and potential links in the Rheme area would therefore be 1:4, slightly lower than the proportion of key-word links in the Theme area and key-word links in the Rheme area.

Although this does not fully compensate for the increase in difference from words to links in the case of 'ape', the potential link is a more reasonable basis for comparison than simply the word tokens. In the light of this calculation, the

links involving the key word 'ape' in Theme (10) and Rheme (45) may be regarded as fairly evenly distributed. Therefore Hypothesis 6.7a may be further revised as Hypothesis 6.7b, as follows.

Hypothesis 6.7b. The ratio of the density of key-word links to the number of potential links by all words will be higher in the Theme area than in the Rheme area of the text.

Now, Hypothesis 6.7b could be tested on the basis of potential links by all words in the Theme and Rheme areas of the sample texts. The results of the calculation of the ratios are presented in Table 6.20 below. For legibility, the results are multiplied by 100,000.

Texts	*KLD*		*PAL*		*KLD/PAL (x100,000)*	
	Theme	*Rheme*	*T-T*	*R-R*	*Theme*	*Rheme*
B1	1.6143	1.6279	153,181	504,510	1.0539	0.3227
B2	1.7973	2.1829	152,076	344,035	1.1818	0.6345
B3	1.7778	1.9844	128,271	463,203	1.3860	0.4284
B4	1.0698	1.0556	70,125	342,378	1.5256	0.3083
B5	3.1370	6.0183	316,410	878,475	0.9914	0.6851
B6	1.5510	1.4068	203,841	426,426	0.7609	0.3299
B7	1.7018	1.9792	94,830	289,941	1.7946	0.6826
B8	1.4048	1.2703	129,286	396,495	1.0866	0.3204
B9	1.8696	2.1143	112,101	267,546	1.6678	0.7903
B10	3.6047	1.6250	280,875	519,690	1.2834	0.3127
B11	2.8947	2.6139	204,480	689,725	1.4156	0.3790
C1	2.9167	3.5362	586,986	2,114,596	0.4969	0.1672
C2	1.9737	1.5833	87,990	252,405	2.2431	0.6273
C3	2.1957	2.7407	108,345	445,096	2.0266	0.6158
C4	3.2062	2.5577	173,166	511,566	1.8515	0.5000
C5	1.3488	1.7634	243,951	1,162,050	0.5529	0.1517
C6	2.6435	2.9818	471,906	1,474,903	0.5602	0.2022
C7	2.1087	1.4348	103,285	410,871	2.0416	0.3492
C8	0.6207	1.3438	47,895	101,475	1.2960	1.3243
C9	1.6111	2.0862	108,811	292,995	1.4806	0.7120
C10	1.7037	1.4857	55,945	283,881	3.0453	0.5234
C11	1.5833	3.5535	240,471	678,030	0.6584	0.5241
C12	4.7308	1.2679	200,661	334,153	2.3576	0.3794
C13	1.5472	2.0556	257,403	788,140	0.6011	0.2608
C14	2.8351	1.7453	218,130	848,253	1.2997	0.2058
C15	3.1429	2.4083	445,096	790,653	0.7061	0.3046
C16	1.2763	2.2576	124,251	583,740	1.0272	0.3867
C17	2.2921	1.7500	243,951	649,230	0.9396	0.2696
C18	3.6207	4.0513	63,190	226,801	5.7299	1.7863
C19	3.4381	4.8543	228,826	1,078,246	1.5025	0.4502
C20	0.8182	0.8667	77,421	306,153	1.0568	0.2831
Total	*2.1947*	*2.2646*	*5,933,156*	*18,455,661*	*1.4717*	*0.4909*

Table 6.20. Ratios of key-word link density to potential all-word links in the Theme and Rheme areas of Groups B and C Sample Texts

In Table 6.20, PAL means potential all-word links. The differences between the ratios are clearly visible in Figure 6.17 below.

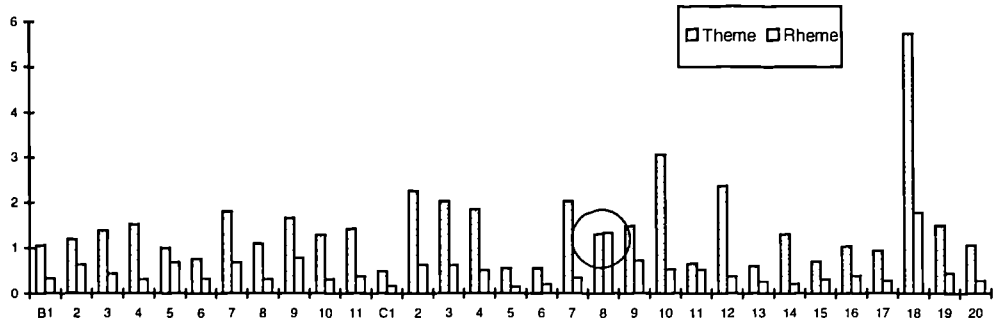


Figure 6.17. Ratios of key-word link density to potential all-word links in the Theme and Rheme areas of Groups B and C Sample Texts

Figure 6.17 shows a very strong tendency for the ratios to be higher in Theme than in Rheme. There is only one counter-example, where Text C8 has the ratio slightly higher in Rheme than in Theme.

However, there remains a problem. When the potential all-word links were calculated using the formula $L = \frac{n(n-1)}{2}$, all possible links the word tokens in a text might form were counted; thus links within the same sentence were also included. But this was not justified, since intra-sentential links could not be regarded as having a textual status. Therefore, the above formula needed to be modified so that those intra-sentential links could be excluded from the total number of potential links. To exclude intra-sentential links, the number of such links in a text needed to be counted. Based on the consideration that manual counting would be unreliable as well as time-consuming, it was decided that a new formula was needed to achieve an approximate number of potential inter-

sentential links in the Theme and Rheme areas of the sample texts. The new formula is in fact a refinement of the above formula, as follows:

$$L = \frac{n(n-1)}{2} - \frac{n/s\,(n/s - 1)}{2} \times s$$

where s is the number of sentences in the text. The second half of this formula returns an approximate number of potential intra-sentential links in the text areas concerned. When this number is subtracted from the total number of all the potential links returned by the first half of the formula, we get the number of potential inter-sentential links. To illustrate how this formula works, let us again take Sample Text B1 as an example. As shown in Table 5.4 of Chapter 5, this text has 75 sentences, and as shown in Table 6.9 of this chapter, it has 554 word tokens in Theme and 1,005 word tokens in Rheme, so there are on average 7.39 words in the Theme of each sentence (554/75) and 13.40 words in the Rheme of each sentence (1,005/75). When the resulting numbers were rounded up to integers, there would be 1,769 potential intra-sentential links in Theme ([7.39×(7.39-1)÷2]×75) and 6,231 potential intra-sentential links in Rheme ([13.40×(13.40-1)÷2]×75). In the original calculation, the 554 words in Theme formed 153,181 potential links and the 1,005 words in Rheme formed 504,510 potential links. After subtraction of the potential intra-sentential links, there were 151,412 potential inter-sentential links in Theme (153,181-1,769) and 498,279 potential inter-sentential links in Rheme (504,510-6,231).

The recalculated potential inter-sentential links in all the thirty-one Groups B and C Sample Texts are represented in Table 6.21 below.

Texts	*S*	*IntraSPL*		*OrigPL*		*InterSPL*	
		Th	*Rh*	*T-T*	*R-R*	*T-T*	*R-R*
B1	75	1,769	6,231	153,181	504,510	151,412	498,279
B2	68	1,964	4,650	152,076	344,035	150,112	339,385
B3	74	1,483	5,785	128,271	463,203	126,788	457,418
B4	61	965	5,206	70,125	342,378	69,160	337,172
B5	78	3,664	10,608	316,410	878,475	312,746	867,867
B6	85	2,082	4,560	203,841	426,426	201,759	421,866
B7	64	1,267	4,155	94,830	289,941	93,563	285,786
B8	69	1,623	5,307	129,286	396,495	127,663	391,188
B9	63	1,546	3,887	112,101	267,546	110,555	263,659
B10	84	2,973	5,683	280,875	519,690	277,902	514,007
B11	90	1,956	7,083	204,480	689,725	202,524	682,642
C1	144	3,538	13,663	586,986	2,114,596	583,448	2,100,933
C2	58	1,311	4,002	87,990	252,405	86,679	248,403
C3	67	1,388	6,178	108,345	445,096	106,957	438,918
C4	77	1,958	6,144	173,166	511,566	171,208	505,422
C5	120	1,686	8,928	243,951	1,162,050	242,265	1,153,122
C6	109	3,848	12,680	471,906	1,474,903	468,058	1,462,223
C7	59	1,527	6,518	103,285	410,871	101,758	404,353
C8	32	1,347	2,953	47,895	101,475	46,548	98,522
C9	57	1,680	4,764	108,811	292,995	107,131	288,231
C10	45	1,079	5,940	55,945	283,881	54,866	277,941
C11	64	3,416	10,021	240,471	678,030	237,055	668,009
C12	75	2,363	4,052	200,661	334,153	198,298	330,101
C13	86	2,638	8,544	257,403	788,140	254,765	779,596
C14	96	1,945	8,191	218,130	848,253	216,185	840,062
C15	91	4,424	8,066	445,096	790,653	440,672	782,587
C16	75	1,411	7,250	124,251	583,740	122,840	576,490
C17	80	2,704	7,553	243,951	649,230	241,247	641,678
C18	48	1,142	4,395	63,190	226,801	62,048	222,406
C19	95	2,074	10,623	228,826	1,078,246	226,752	1,067,623
C20	69	928	4,051	77,421	306,153	76,493	302,102
Total	*2,358*	*61,898*	*204,388*	*5,933,156*	*18,455,661*	*5,871,258*	*18,251,273*

Table 6.21. Potential inter-sentential links in the Theme and Rheme areas of Groups B and C Sample Texts

In Table 6.21, 'IntraSPL' means potential intra-sentential links, 'OrigPL' means originally calculated all-word potential links and 'InterSPL' means potential inter-sentential links, which were achieved by subtracting intra-

sentential links from originally calculated all-word potential links. Having achieved the potential inter-sentential links, the ratios of key-word link density to potential all-word links in the Theme and Rheme areas of Groups B and C Sample Texts were also recalculated to exclude intra-sentential links. The results of this recalculation are represented in Table 6.22 below.

Texts	*KLD*		*PIL*		*KLD/ PIL(x100,000)*	
	Theme	*Rheme*	*Theme*	*Rheme*	*Theme*	*Rheme*
B1	1.6143	1.6279	151,412	498,279	1.0662	0.3267
B2	1.7973	2.1829	150,112	339,385	1.1973	0.6432
B3	1.7778	1.9844	126,788	457,418	1.4022	0.4338
B4	1.0698	1.0556	69,160	337,172	1.5468	0.3131
B5	3.137	6.0183	312,746	867,867	1.0031	0.6935
B6	1.551	1.4068	201,759	421,866	0.7687	0.3335
B7	1.7018	1.9792	93,563	285,786	1.8189	0.6925
B8	1.4048	1.2703	127,663	391,188	1.1004	0.3247
B9	1.8696	2.1143	110,555	263,659	1.6911	0.8019
B10	3.6047	1.625	277,902	514,007	1.2971	0.3161
B11	2.8947	2.6139	202,524	682,642	1.4293	0.3829
C1	2.9167	3.5362	583,448	2,100,933	0.4999	0.1683
C2	1.9737	1.5833	86,679	248,403	2.2770	0.6374
C3	2.1957	2.7407	106,957	438,918	2.0529	0.6244
C4	3.2062	2.5577	171,208	505,422	1.8727	0.5061
C5	1.3488	1.7634	242,265	1,153,122	0.5567	0.1529
C6	2.6435	2.9818	468,058	1,462,223	0.5648	0.2039
C7	2.1087	1.4348	101,758	404,353	2.0723	0.3548
C8	0.6207	1.3438	46,548	98,522	1.3335	1.3640
C9	1.6111	2.0862	107,131	288,231	1.5039	0.7238
C10	1.7037	1.4857	54,866	277,941	3.1052	0.5345
C11	1.5833	3.5535	237,055	668,009	0.6679	0.5320
C12	4.7308	1.2679	198,298	330,101	2.3857	0.3841
C13	1.5472	2.0556	254,765	779,596	0.6073	0.2637
C14	2.8351	1.7453	216,185	840,062	1.3114	0.2078
C15	3.1429	2.4083	440,672	782,587	0.7132	0.3077
C16	1.2763	2.2576	122,840	576,490	1.0390	0.3916
C17	2.2921	1.75	241,247	641,678	0.9501	0.2727
C18	3.6207	4.0513	62,048	222,406	5.8353	1.8216
C19	3.4381	4.8543	226,752	1,067,623	1.5162	0.4547
C20	0.8182	0.8667	76,493	302,102	1.0696	0.2869
Total	*2.1947*	*2.2646*	*5,871,258*	*18,251,273*	*0.0374*	*0.0124*

Table 6.22. Ratios of key-word link density to potential inter-sentential links in the Theme and Rheme areas of Groups B and C Sample Texts

In Table 6.22, KLD means key-word link density, PIL means potential inter-sentential links. The differences between the ratios of key-word link density to potential inter-sentential links in the Theme and Rheme areas of Groups B and C Sample Texts are clearly visible in Figure 6.18 below.

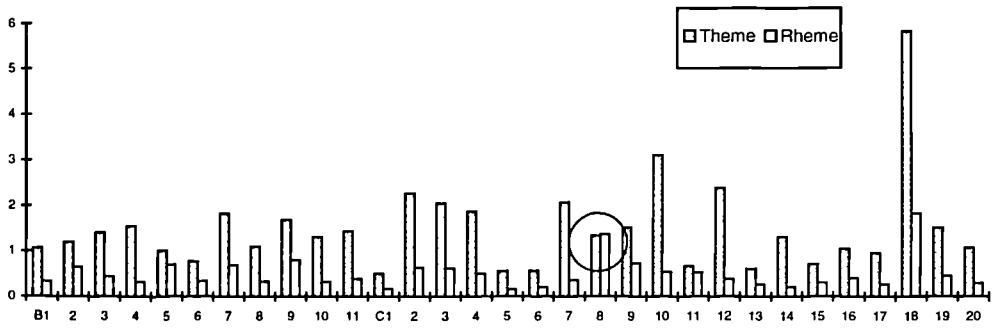


Figure 6.18. Differences between the ratios of key-word link density to potential inter-sentential links in the Theme and Rheme areas of Groups B and C Sample Texts

In spite of the subtraction of intra-sentential links, Figure 6.18 looks almost identical to Figure 6.17. It also shows a very strong tendency for the ratios to be higher in Theme than in Rheme. Similar to Figure 6.17, there also remains one exception: Sample Text C8. A close look at Sample Text C8 reveals that this text is the shortest in all the sample texts, with only 761 word tokens in total. It has only sixteen key word types in sixty-one key word tokens, with twenty-nine tokens in Theme and thirty-two in Rheme. One key word 'body' occurs three times in Theme and eight times in Rheme, a difference of only a little more than 1:2. But this means that this key word forms three links in Theme and twenty-eight links in Rheme, which makes the difference increase to over 1:9.

In other sample texts, such a difference would be counter balanced by the concentration of other key words and links they form in Theme. However, this is not possible with Text C8, owing to the very small total number of key words in the text.

Even with this one exception, Hypothesis 6.7b was still strongly supported. The *t*-test resulted in a P-value of 0.00000269, showing that the differences between the ratios in Theme and Rheme are statistically highly significant. Comparing the P-value of 0.00003998 in testing Hypothesis 6.7a, which is nearly 15 times greater than the P-value of this test, Hypothesis 6.7b may be regarded as more firmly supported than Hypothesis 6.7a.

6.6.2 Hypotheses 6.8 and 6.9

Up to this point, Hypotheses 6.1 to 6.7 have been tested. The analysis has supported all the hypotheses, with Hypothesis 6.7 reformulated. On the basis of these findings, it is possible to propose two additional hypotheses:

Hypothesis 6.8. **The differences between the ratios of key-word link density to potential inter-sentential links in the Theme area and to those in the Rheme area will be greater than the differences between the ratios of lexical links to all words in the Theme area and to those in the Rheme area.**

Similarly,

Hypothesis 6.9. **The differences between the ratios of key-word link density to potential inter-sentential links in the Theme area and to those in the Rheme area will be greater than the differences between the ratios of key words to all words in the Theme area and to those in the Rheme area.**

If these two hypotheses could be supported, it would imply that the ratios of key-word links to all potential inter-sentential links may be more useful in identifying the textual status of key words in the Theme and Rheme areas than either the ratios of lexical links to all words or the ratios of key words to all words.

6.6.3 Testing Hypotheses 6.8 and 6.9

To test Hypothesis 6.8 and Hypothesis 6.9, it was necessary to compare the differences between the ratios in Theme and Rheme of lexical links to all words, of key words to all words, and of key-word link density to potential inter-sentential links formed by all words. Drawing on Table 6.15 in Section 6.5.5, we get the ratios of key words to all words of the thirty-one texts in Groups B and C; and drawing on Table 6.22 in Section 6.6.1.1, we get the ratios of key-word link density to potential inter-sentential links by all words. Then the ratios of lexical links to all words and their differences in Theme and Rheme of the thirty-one sample texts were calculated. All these ratios and their differences in the Theme and Rheme areas are presented in Table 6.23 below.

Texts	*LL/TL*			*KW/TL*			*KLD/PIL*		
	Th	*Rh*	*Th/Rh*	*Th*	*Rh*	*Th/Rh*	*Th*	*Rh*	*Th/Rh*
B1	0.2978	0.2199	1.3543	0.1264	0.0856	1.4766	1.0662	0.3267	3.2635
B2	0.3007	0.2964	1.0145	0.1341	0.0988	1.3573	1.1973	0.6432	1.8615
B3	0.3037	0.3385	0.8972	0.1243	0.0665	1.8692	1.4022	0.4338	3.2324
B4	0.1840	0.1449	1.2698	0.1147	0.0652	1.7592	1.5468	0.3131	4.9403
B5	0.3832	0.6523	0.5875	0.0917	0.0822	1.1156	1.0031	0.6935	1.4464
B6	0.1784	0.1569	1.1370	0.0767	0.0639	1.2003	0.7687	0.3335	2.3049
B7	0.3050	0.2257	1.3514	0.1307	0.0630	2.0746	1.8189	0.6925	2.6266
B8	0.1473	0.1380	1.0674	0.0825	0.0415	1.9880	1.1004	0.3247	3.3890
B9	0.2300	0.2787	0.8253	0.0970	0.0956	1.0146	1.6911	0.8019	2.1089
B10	0.4027	0.1667	2.4157	0.0573	0.0314	1.8248	1.2971	0.3161	4.1034
B11	0.5516	0.3600	1.5322	0.1781	0.0860	2.0709	1.4293	0.3829	3.7328
C1	0.4991	0.4832	1.0329	0.1550	0.1006	1.5408	0.4999	0.1683	2.9703
C2	0.2262	0.1125	2.0107	0.0905	0.0338	2.6775	2.2770	0.6374	3.5723
C3	0.2489	0.2595	0.9592	0.0987	0.0572	1.7255	2.0529	0.6244	3.2878
C4	0.5620	0.3261	1.7234	0.1647	0.1028	1.6021	1.8727	0.5061	3.7003
C5	0.2246	0.2787	0.8059	0.1230	0.0859	1.4319	0.5567	0.1529	3.6409
C6	0.3796	0.4430	0.8569	0.1183	0.0960	1.2323	0.5648	0.2039	2.7700
C7	0.2484	0.0728	3.4121	0.1011	0.0254	3.9803	2.0723	0.3548	5.8408
C8	0.1097	0.1308	0.8387	0.0935	0.0710	1.3169	1.3335	1.3640	0.9776
C9	0.2227	0.2298	0.9691	0.1156	0.0757	1.5271	1.5039	0.7238	2.0778
C10	0.1582	0.1432	1.1047	0.0806	0.0464	1.7371	3.1052	0.5345	5.8095
C11	0.2709	0.5665	0.4782	0.1037	0.1365	0.7597	0.6679	0.5320	1.2555
C12	0.6530	0.1809	3.6097	0.1230	0.0685	1.7956	2.3857	0.3841	6.2111
C13	0.1992	0.2213	0.9001	0.0738	0.0573	1.2880	0.6073	0.2637	2.3030
C14	0.5068	0.2487	2.0378	0.1467	0.0814	1.8022	1.3114	0.2078	6.3109
C15	0.5773	0.3347	1.7248	0.1557	0.0954	1.6321	0.7132	0.3077	2.3178
C16	0.2365	0.4255	0.5558	0.1523	0.1221	1.2473	1.0390	0.3916	2.6532
C17	0.3448	0.2386	1.4451	0.1273	0.0842	1.5119	0.9501	0.2727	3.4840
C18	0.3792	0.3398	1.1160	0.0815	0.0579	1.4076	5.8353	1.8216	3.2034
C19	0.5894	0.6072	0.9707	0.1551	0.1028	1.5088	1.5162	0.4547	3.3345
C20	0.1345	0.1290	1.0426	0.0279	0.0192	1.4531	1.0696	0.2869	3.7281
Total	*0.3244*	*0.2823*	*1.3241*	*0.1130*	*0.0742*	*1.6429*	*0.0374*	*0.0124*	*3.0161*

Table 6.23. Comparison of the three kinds of Theme/Rheme ratio differences

In Table 6.23, the columns under 'Th' represent the ratios in the Theme area, the columns under 'Rh' represent the ratios in the Rheme area, and the columns under 'Th/Rh' represent the differences between the ratios in Theme and Rheme. For example, in Sample Text B1, the ratio of lexical links to all words in the Theme area is 0.2978 and that in the Rheme area is 0.2199. The

difference between the ratios is 1.3543, which was obtained by dividing 0.2978 by 0.2199. In the same text, the ratio of key words to all words in the Theme area is 0.1264 and that in the Rheme area is 0.0856. The difference between the ratios is 1.4766, which was obtained by dividing 0.1264 by 0.0856. However, in Sample Text B1 the greatest difference lies between the ratios of key-word link density and potential inter-sentential links formed by all words. The ratio of key-word link density to potential inter-sentential links in the Theme area is 1.0662 and that in the Rheme area is 0.3267. The difference between these ratios is 3.2635, more than twice that of the Theme/Rheme differences either between the ratios of lexical links to all words (1.3543) or between the ratios of key words to all words (1.4766).

This comparison shows that, in all the thirty-one Groups B and C Sample Texts, the differences between the ratios of key-word link density to all potential links in Theme and the ratios of key-word link density to all potential links in Rheme are greater than the differences between the ratios of lexical links to all words in Theme and the ratios of lexical links to all words in Rheme. The results of this comparison are clearly represented in Figure 6.19 below.

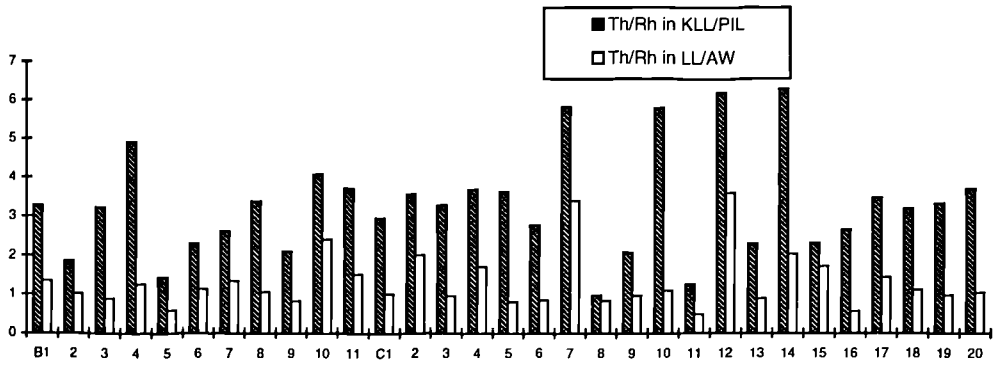


Figure 6.19. Comparison of Theme/Rheme ratio differences between key-word link density to potential inter-sentential links and lexical links to all words in Groups B and C Sample Texts

Figure 6.19 shows that the differences between the ratios of key-word link density to potential inter-sentential links in Theme and the ratios of key-word link density to potential inter-sentential links in Rheme are greater than the differences between the ratios of lexical links to all words in Theme and the ratios of lexical links to all words in Rheme in all the thirty-one Groups B and C Sample Texts. The *t*-test returned a P-value of 0.000000002. This shows that Hypothesis 6.8 of this chapter was strongly supported by the analysis.

Turning then to Hypothesis 6.9, the results of the comparison of the differences between the ratios of key-word link density to potential inter-sentential links in Theme and the ratios of key-word link density to potential inter-sentential links in Rheme and the differences between the ratios of key words to all words in Theme and the ratios of key words to all words in Rheme of Groups B and C Sample Texts are clearly shown in Figure 6.20 below.

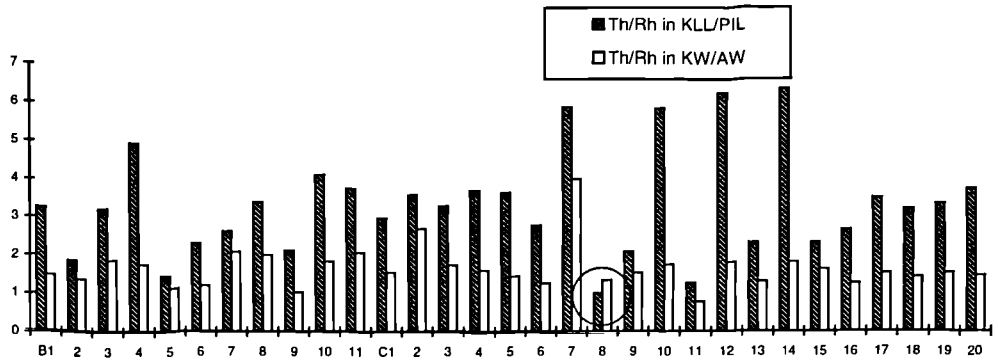


Figure 6.20. Comparison of Theme/Rheme ratio differences between key-word link density to potential inter-sentential links and key words to all words in Groups B and C Sample Texts

Figure 6.20 shows that the differences between the ratios of key-word link density to potential inter-sentential links in Theme and the ratios of key-word link density to potential inter-sentential links in Rheme are greater than the differences between the ratios of key words to all words in Theme and the ratios of key words to all words in Rheme. Admittedly there was still one exception, where in Text C8 the difference between the ratios of key words to all words in Theme and Rheme was greater than the difference between the ratios of key-word link density to all potential links in Theme and Rheme. Nevertheless, the *t*-test returned a P-value of 0.000000051. This shows that Hypothesis 6.9 of this chapter was strongly supported by the analysis.

6.7 Conclusion

In this chapter, it was assumed that key-word links are a better indicator of text organisation than either lexical links or key words alone. So, some hypotheses were set up to explore the characteristics of key-word links in comparison with lexical links and key words. Hypotheses 6.1 to 6.6 were strongly supported. The distribution of key-word links in the Theme and Rheme areas of the sample texts was displayed in a much clearer picture than either the distribution of lexical links or the distribution of key words in the Theme and Rheme areas of the sample texts.

However, six counter examples came into view in the test of Hypothesis 6.7. So some modifications were made to the formulation of this hypothesis. When Hypothesis 6.7 was modified to Hypothesis 6.7a, test results were improved to a certain extent, with three exceptional cases. Then Hypothesis 6.7a was revised to Hypothesis 6.7b, and more significant results were achieved, with only two exceptions.

Based on the fact that the first seven hypotheses were all supported, two further hypotheses were put forward: Hypothesis 6.8 and Hypothesis 6.9, which proposed that key-word links could be used to signal patterns of information organisation in the text more clearly. Hypothesis 6.8 proposed that between the Theme and Rheme areas there would be a greater difference as regards the ratios of key-word link density to potential inter-sentential links than the ratios

of lexical links to all words. Hypothesis 6.9 proposed that between the Theme and Rheme areas there would be a greater difference as regards the ratios of key-word link density to potential inter-sentential links than the ratios of key words to all words. Greater differences between the ratios may be regarded as a clearer manifestation of the patterning of information distribution and text organisation. Both Hypothesis 6.8 and Hypothesis 6.9 were supported, though in testing Hypothesis 6.9 there was still one exception with Sample Text C8. Sample Text C8 is unusual in that all the exceptional cases involved this text. It is the shortest text in the corpus of this thesis, and moreover, it is the text with the fewest key words according to our procedure.

Since the study reported in this chapter is corpus-based quantitative study, and the nature of the results is probabilistic, it only reveals a tendency in text organisation as reflected in a very limited text type. For more general conclusions it would necessarily use a much larger corpus, composed of texts of a reasonable length and a variety of text types. In the corpus of this thesis, the texts are all between 1,000 to 3,000 words in length. Sample Text C8 was purposely included as the shortest text with fewer than 800 words. Thus it seems that the length of sample texts may affect our results.

The implications of the exceptions is that text is organised rather than structured. An author is able to choose to present information idiosyncratically or in a way which does not conform to linguistic norm. There may be different methods of text organisation, as Fries (1983, 1995) proposes. In our data, an

overwhelming majority of texts develop around Theme, yet some texts may develop around Rheme.

Nevertheless, the study reveals a very strong tendency for the text type analysed so far to have key words concentrated in the Theme rather than in the Rheme of the text, and to have proportionally more links in Theme than in Rheme.

The most important implication of the findings of this chapter is that key-word links can be claimed to be a better indicator of text organisation. The results in testing the hypotheses strongly support this claim. If proportionally more lexical links tend to occur in Theme than in Rheme, the proportion would increase with key-word links. Likewise, if there is a high probability for key words to concentrate in the Theme area of the text, this may be more clearly demonstrated by the key-word links. In the cases where there are proportionally fewer key-word links than key words in Theme, the reason may possibly be that these key word types are not highly concentrated in the Theme area of the text and thus not highly representative of the main concerns of the text. This thesis will not explore this point further, but it is one that may lead to more research work in the future.

Chapter 7. Conclusions

7.1 Introduction to this chapter

Johnson (1996) recommended that 'the conclusion of an essay should draw together all the previous points of your argument into one general statement which is then directly related to the essay topic or the question you have been answering.' R. Berry (1986: 99) made a similar point and continued that 'a research paper should be circular in argument. That is, the formal aim of the paper should be stated in the opening paragraph; the conclusion should return to the opening, and examine the original purpose in the light of the data assembled.' Following their recommendation, this chapter will provide a summary of the research work from the opening questions to the major findings, followed by a comment on the features, attainment of aims, and limitations of this study. Then, by discussing implications of the findings in this study, this chapter will answer the fourth general research question raised in Chapter 1 of this thesis. Finally, it will propose possible directions for further research.

7.2 A summary of the study

Chapter 1 of this thesis provided the background for the present study. The original motivation for the study was a need to process text more efficiently. Facing the explosive increase of textual data to process, one possible solution was to make use of automatic summarisation of long texts. For this purpose, it is necessary to have a better understanding of the nature of text, specifically the organisation of information in the text.

After discussing major approaches to the study of information distribution in text, it was decided to approach this issue through the triple interface of lexical patterning, key words and the Theme-Rheme system in text. The premise for the present study was that the patterning of lexis, the distribution of key words, and the Theme-Rheme system in text are all important devices in creating text organisation. Viewed from another perspective, they help reveal the text's organisation. So a study of the relationships amongst these categories of devices might result in a better understanding of the properties of text.

As stated in Chapter 1 of this thesis, the present study is exploratory in nature. So, analogously to Theme-Rheme development, the underlying principle of the research method in this study has been to progress from the known to the unknown, or from the familiar ground to a new field. The study has been basically 'data-driven' or 'corpus-driven'. A large quantity of text data was used. Statistical results from a preliminary study of the texts were used as the

spring-board for an on-going exploration. Because of the exploratory nature of the study, hypotheses were set up a chapter at a time rather than stated entirely at the beginning of the thesis. Hypotheses by definition should be open to confirmation or nullification; if a hypothesis failed, it had to be given up or modified.

Chapter 2 reviewed influential works on the Theme-Rheme system, which was one of the two major bases of the present study. In particular, approaches to the definition, function, and scope of Theme were reviewed. In this chapter, Theme was reviewed from two perspectives: the first concerned its function at the clause level, the other its function at the text level. Within the two perspectives, views presented by the so-called 'combiners' and 'separators' were examined in detail. For the purpose of the present study, the focus was laid on the function of Theme at the text level. Within the 'combiners', Daneš's (1974) framework of Thematic Progression (TP) and his later modifications were reviewed and felt to be a most useful starting point for text analysis in the present study. Within the 'separators', or the systemic linguists, Berry's (1996) study on the function of Theme in prioritising communicative meaning in text, especially her notion of 'discourse $Theme_M$', was felt to be central to the concerns of the present study.

Chapter 3 reviewed works on the function of lexis in text organisation. One of the forerunners in this field was Ure (1971), who first used a corpus of a considerable size at the time to discover the relationship between lexical

density and text types. Halliday and Hasan (1976) represented a milestone in the study of text in that they were the first to introduce the concept of cohesion that directly contributes to the coherence of text. However, in spite of the importance of their work, they underestimated the value of lexical cohesion in text organisation. It was in Hasan (1984) that an attempt was made to remedy this neglect. In this later work, Hasan gave much more weight to the role lexis plays in creating a coherent text. Central to the present study are the notions of 'lexical chains' and 'cohesive harmony', which stress the correlation between the distribution of lexical items and people's perception of coherence in the text.

Following their work, various linguists made attempts to retrieve focal information in text by exploiting the distribution of lexical items in text. However, it is in Hoey (1991a) that the first systematic study of patterns of lexis in text was made to reveal the relationship between the distribution of lexical items and the distribution of information in text. Hoey's notion of lexical 'links' by repetition of lexical items was the other major basis of the study in this thesis. Chapter 3 also reviewed some attempts to apply Hoey's system to the retrieval of central information in text with the use of computer technology.

Chapter 4 marks the beginning of text analysis in this study. It reported on a preliminary analysis of the relationship between different types of thematic progression and patterns of lexis in text. The analysis was based on the

assumption that there should be a correlation between the two categories. Drawing on the relevant literature reviewed in the previous two chapters, it was argued that a clause Theme is what the clause is about, and that all the clause Themes in a text realise discourse $Theme_M$, which represents the prioritised meanings of the text. In addition, patterns of thematic progression manifest the foundation of topic development; and above all, thematic progression is basically realised by lexical repetition, which is also important in producing a framework for interpreting what is changed.

Based on these general assumptions, four hypotheses were set up in Chapter 4.

Hypothesis 4.1. Proportionally, there are more T-T links than R-R links between sentences in a text.

Hypothesis 4.2. Proportionally, there are more R-T links than T-R links between sentences in a text.

Hypothesis 4.3. T-T links or R-T links will create more coherence than R-R links or T-R links, where coherence is measured in terms of readers' judgements about the text

Hypothesis 4.4. In the same text a series of sentences with T-T links tend to be closely related in meaning.

These hypotheses were tested on a sample text, a newspaper science report taken from the 'Independent on Sunday'. To 'clean the text' in preparation for the analyse, pronouns and ellipses were lexicalised. Theme was delimited mainly following Berry's (1989) approach, and word lists were made using a computer program.

In testing Hypothesis 4.1, all the T-T links and R-R links in the sample text were counted, and their relative ratios to the total number of words in the Theme and Rheme areas were calculated. The results revealed that there were proportionally more T-T links than R-R links between sentences in the text, and a chi-square test confirmed that the results were statistically significant. Hypothesis 4.2 was tested along the same lines, and similarly a chi-square test confirmed that the results were statistically significant.

Hypothesis 4.3 was different from the previous two hypotheses in that it required the judgement of human informants. Two informant tests were conducted, in which two groups of native English speakers were requested to make comparative evaluations of the coherence of pairs of sentences with T-T links, presumably the most cohesive, and R-T links, presumably the least cohesive. The majority of the informants judged that sentence pairs with T-T links were more coherent than sentence pairs with T-R links. Again a chi-square test confirmed that Hypothesis 4.3 was supported by the analysis.

Hypothesis 4.4 was more qualitative in nature and thus less testable by quantitative means. For testing this hypothesis, some series of sentences with T-T links were selected from the sample text to form sub-texts, and the semantic properties of the sub-texts were discussed. It was evident that although the selected sentences were not adjacent in the original text, and the number of links among them was below the threshold proposed by Hoey's system, the sub-texts nevertheless read reasonably coherently and showed a

clear focus on a certain topic represented by the lexical items threading the T-T links.

In summary, Chapter 4 showed that there was indeed a close relationship between the Theme-Rheme system and patterns of lexis in text. This was manifested not only in the variation of different types of links between the Theme and Rheme areas of the text, but also by the perceived degrees of coherence of sentence pairs and groups with the different types of links.

Chapter 5 moved on from the findings of Chapter 4. It attempted to explore the relationship between the Theme-Rheme system and the distribution of computer generated key words. Key words were defined as words that had an unusual frequency of occurrence in a text as compared with that in the language in general, and keyness as the degree of unusualness of the key words in question. In this thesis, key words were selected from lexical items that were unusually frequent in the sample texts in comparison with a 95-million-word corpus of the same genre.

Just as clause $Theme_{FS}$ of a text realise discourse $Theme_{M}$ which in turn prioritises the meaning of the text (Berry 1996), the key words of a text together were assumed to be able to reflect the 'aboutness' of the text (Scott 1997a). On this basis, three hypotheses were set up in Chapter 5:

Hypothesis 5.1. A higher proportion of key words will occur in Theme than in Rheme.

Hypothesis 5.2. There will be a higher overall 'keyness' in Theme than in Rheme.

Hypothesis 5.3. On average, a key word in Theme will be reiterated more than a key word in Rheme, thus forming more links in Theme than in Rheme.

For testing the hypotheses, eleven more sample texts were collected from the same source as the text used in Chapter 4, so that the generic structure of all the texts used was consistent. The preparation of these texts for analysis was similar to that in Chapter 4, except that pronouns and ellipses were not lexicalised. The reason for this omission was as follows. Based on the analysis in Chapter 4, it was observed that the pronouns and ellipses that had inter-sentential references mostly occurred in the Theme area; therefore not lexicalising these items would reduce the number of key words in the Theme area. This would make testing the hypotheses more demanding but would simplify the analytical procedure.

In testing Hypothesis 5.1, the occurrence of the key words in Theme and Rheme was counted, and the ratios of the occurrence of key words to the total number of words in the Theme and Rheme areas were calculated and compared, by an analytical method similar to that used in Chapter 4. The results revealed that there were indeed proportionally more key words in Theme than in Rheme. A *t*-test was carried out to confirm that the differences were statistically significant; Hypothesis 5.1 was supported.

In testing Hypothesis 5.2, the value of keyness of each key word, which was provided by the computer program, was divided between the Theme and Rheme areas according to the actual occurrence of the key word in these two areas of a text. Then the overall keyness of all the key words in the Theme and Rheme areas of the text was summed up, and the ratios of keyness to all words in the two areas were calculated. The results showed that there was a higher ratio of keyness to all words in the Theme area than in the Rheme area of each of the sample texts, and the *t*-test showed that the differences were significant with all the eleven sample texts. Therefore, Hypothesis 5.2 was supported.

In testing Hypothesis 5.3, the reiteration of key words was measured by calculating the type/token ratios of the key words in the context in which the key words occurred. Then the type/token ratios of the key words in the Theme and Rheme areas of the sample text were compared, taking the number of all words in the two areas as a basis for the comparison. In all the eleven sample texts, the ratios were invariably higher in Theme than in Rheme. This suggests that a key word would generally be reiterated more in the Theme area than in the Rheme area, and consequently would form more links in the Theme area than in the Rheme area. The *t*-test showed that the differences were highly significant, and Hypothesis 5.3 was firmly supported by the analysis.

The results of analysis in Chapter 5 were encouraging. The relationship between the Theme-Rheme system and the distribution of key words in text was regarded as even closer than the relationship between the Theme-Rheme

system and the distribution of lexical links in text. So, for a reliable discovery procedure of information distribution in the text, it was believed that it might be fruitful to explore the relationship between key words and lexical links.

Chapter 6 attempted to carry on the exploration on the basis of the findings of the previous two chapters. It was assumed that if there was a correlation between the Theme-Rheme system and the occurrence of lexical links, as manifested by the analysis of Chapter 4, and if there was a correlation between the Theme-Rheme system and the occurrence of key words, as manifested by the analysis of Chapter 5, then there might also be a relationship between lexical links and key words. Moreover, the relationship between lexical links and key words might be mediated through their relationships with the Theme-Rheme system, realised by key-word links, which meant links formed by key words.

To reveal this relationship, Chapter 6 set out to explore the distribution of key-word links in the Theme and Rheme areas of the sample texts. It was believed that a study of the distribution of key-word links in Theme and Rheme would more clearly reveal the patterns of information distribution in text. For this purpose, initially seven hypotheses were set up, the first six of which were intended to test whether key-word links were a more sensitive indicator of text organisation than either lexical links alone or key words alone. It was assumed that if the ratios of key-word links to key words were higher than those of lexical links to lexical items, and if the ratios of key-word links to all words

were higher than those of key words to all words, then it might be regarded as an indication that key-word links contribute more, more powerfully and more reliably to the creation of the patterns of text organisation than either lexical links or key words. The first seven hypotheses in Chapter 6 were as follows.

Hypothesis 6.1. In a text, the ratio of key-word links to key words will be higher than the ratio of lexical links to lexical items.

Hypothesis 6.2. In the Themes of the text, the ratio of key-word links to key words will be higher than the ratio of lexical links to lexical items.

Hypothesis 6.3. In the Rhemes of the text, the ratio of key-word links to key words will be higher than the ratio of lexical links to lexical items.

Hypothesis 6.4. In a text, the ratio of key-word links to all words will be higher than the ratio of key words to all words.

Hypothesis 6.5. In the Themes of the text, the ratio of key-word links to all words will be higher than the ratio of key words to all words.

Hypothesis 6.6. In the Rhemes of the text, the ratio of key-word links to all words will be higher than the ratio of key words to all words.

Hypothesis 6.7. The ratio of key-word links to all words will be higher in the Themes than in the Rhemes of the text.

Because Chapter 6 was intended to be a continuation of the exploration carried out in Chapters 4 and 5, it was decided that the sample texts used in the previous two chapters, as well as some of the calculated statistical data, should be reused. For testing the hypotheses of this chapter, an initial analysis was carried out on the sample text used in Chapter 4. The results were all positive, so that it was likely that *all* the hypotheses could be supported. Therefore, a full-scale analysis was carried out. For this analysis, the eleven sample texts used in Chapter 5 were reused, and in addition, twenty more sample texts were collected from the same source, for the sake of consistency.

In testing Hypotheses 6.1 to 6.3, the ratios of key-word links to key words and the ratios of lexical links to lexical items in the Theme and Rheme areas and the overall texts were calculated and compared. The results showed that the ratios of key-word links to key words were invariably higher than the ratios of lexical links to lexical items in all the thirty-one sample texts. The *t*-test carried out on the results indicated that all the three hypotheses were firmly supported.

In testing Hypotheses 6.4 to 6.6, the ratios of key-word links to all words and the ratios of key words to all words in all the thirty-one sample texts were calculated and compared. With a few exceptional cases, the results showed a general trend for the ratios of key-word links to all words to be higher than the ratios of key words to all words, either in the Theme area, Rheme area, or in the overall text. Again a *t*-test was carried out to check each of the results and the tests returned very significant P-values, showing that Hypotheses 6.4 to 6.6 were firmly supported.

However, in testing Hypothesis 6.7, six exceptional cases were encountered out of the thirty-one sample texts. Moreover, the *t*-test returned a P-value that was only marginally acceptable. Therefore, Hypothesis 6.7 was rejected and modified as Hypothesis 6.7a, where the notion of 'key-word density' was introduced, referring to the number of links formed by a certain number of key words, as opposed to the distance traversed by the links. Hypothesis 6.7a is as follows:

Hypothesis 6.7a. The ratio of key-word link density to all words will be higher in the Themes than in the Rhemes of the text.

With this modification, the results showed a strong tendency for the ratios to be higher in the Theme area than in the Rheme area, although there were still three exceptional cases. However, the *t*-test returned a P-value of less than 0.0004, which implied that Hypothesis 6.7a was strongly supported by the analysis.

A further attempt was made to discover a clearer indicator of information distribution in the Theme and Rheme areas of the text. So a new notion was introduced, that of the 'potential links'. Potential links, as the name implies, are not realised; in fact the majority of potential links are never realised in any normal text. It was noted that, in the extremely unlikely situation of all the words in a text being of the same word type, every word would be linked with every other by repetition in the same text. In such a case, all potential links would be realised. The notion of potential links was useful for serving as a basis for calculating the distribution of key-word links in the Theme and Rheme areas of the text. Hypothesis 6.7a was therefore reformulated as Hypothesis 6.7b:

Hypothesis 6.7b. The ratio of the density of key-word links to the number of potential links by all words will be higher in the Theme area than in the Rheme area of the text.

Because the Rheme area is always larger than the Theme area in all the thirty-one sample texts, there are more potential links in Rheme than in Theme. In testing Hypothesis 6.7b, the ratios of realised key-word links to potential links

by all words in the Theme and Rheme areas were calculated. It was decided that intra-sentential links should be excluded from the calculation, because they should be regarded as having little textual status. Therefore, the calculation only took account of potential inter-sentential links. The calculation revealed that the ratios are higher in the Theme area than in the Rheme area of almost all the thirty-one sample texts, with only one exception. The *t*-test returned a P-value of an even higher significance than the P-value in testing Hypothesis 6.7a. Therefore it was concluded that Hypothesis 6.7b better suited our purpose. Developing from this, Hypotheses 6.8 and 6.9 were then set up,

Hypothesis 6.8. The differences between the ratios of key-word link density to potential inter-sentential links in the Theme area and to those in the Rheme area will be greater than the differences between the ratios of lexical links to all words in the Theme area and to those in the Rheme area.

Hypothesis 6.9. The differences between the ratios of key-word link density to potential inter-sentential links in the Theme area and to those in the Rheme area will be greater than the differences between the ratios of key words to all words in the Theme area and to those in the Rheme area.

These two hypotheses together imply that key-word link density in the Theme and Rheme areas, measured against potential inter-sentential links, may be a more useful notion than either lexical links or key words for identifying information patterning and text organisation.

In testing Hypothesis 6.8, the differences between the ratios of key-word link density to all potential inter-sentential links in the Theme area and the ratios of key-word link density to all potential inter-sentential links in the Rheme area

were compared with the differences between the ratios of lexical links to all words in the Theme area and the ratios of lexical links to all words in the Rheme area. In testing Hypothesis 6.9, the differences between the ratios of key-word link density to all potential inter-sentential links in the Theme area and the ratios of key-word link density to all potential inter-sentential links in the Rheme area were compared with the differences between the ratios of key words to all words in the Theme area and the ratios of key words to all words in the Rheme area. The results showed a very strong tendency for the Theme/Rheme differences of key-word link density to all potential inter-sentential links to be greater than either the Theme/Rheme differences of lexical links to all words or the Theme/Rheme differences of key words to all words. The *t*-tests returned P-values of 0.000000002 and 0.000000051 respectively. This means that Hypotheses 6.8 and 6.9 were very firmly supported. Therefore, Chapter 6 showed key-word link density to be a better indicator than either lexical links or key words alone to represent the picture of information distribution in text.

7.3 Features of the study

There are three prominent features of this study. Firstly, it is a quantitative study on the basis of a large corpus of written texts. Secondly, it takes text as its unit of analysis. Thirdly, it explores text organisation in terms of the triple

interface of lexical patterning, key words and the Theme-Rheme system in the text.

For a quantitative study, computer technology is necessarily employed, as the computer is efficient in processing large quantities of textual data, as well as numerical data. In this study, all the textual data were collected in electronic form, and except for the Theme delimitation, which at this stage of computer technology required a human operation, all the categories of this study were computer recognisable. This made it possible to process a larger quantity of textual data with reasonable accuracy than could possibly be processed manually. This study not only used computer programs to generate word lists and key-word lists with relevant statistics of the sample texts, but also used computer programs to calculate the comparative ratios of the lexical items and key words, and to carry out statistical tests on the results. All of this could not have been done without recent advances in computer technology.

Taking text as the basic unit of analysis in this study meant that the study of lexis was based on the performance of the lexical items in text, that the study of key words was based on the performance of the key words in text, and that the study of the Theme-Rheme system was based on the performance of the thematic elements in text. This is different from many other studies of lexis or the Theme-Rheme system. In many research projects, lexis has either been regarded as peripheral (e.g. Halliday and Hasan 1976), or regarded as an isolated category for its own sake (e.g. West 1960). In like manner, as reviewed

in Chapter 2, the Theme-Rheme system has been regarded by many linguists only as a category on the clause level, and those linguists who have noticed its significance on the text level have focused on the impact that thematic choices have on the text and have neglected how text constrains thematic choices.

The mapping of patterns of lexical repetition onto the patterns of thematic progression presupposes that text is a complex phenomenon. Therefore, it was believed that these factors of text should be studied together to present a clearer picture of text organisation than if they were studied separately. Through exploration of the inter-relationships amongst lexical patterning, key words and the Theme-Rheme system, this study provided evidence that the distribution of information in a text has some recognisable patterns, and that these patterns may reveal some aspects of the property of the language in general.

7.4 Attainment of aims

This study was motivated by the practical need for the processing of large quantities of textual data especially in the context of information retrieval. The primary aim of this study was to explore the nature of text organisation, with the specific objective of developing a discovery procedure for recognising patterns of information distribution in the text. A secondary aim was possible application of the discovery procedure to the eventual production of computer programs for automatic summarisation.

Two questions were raised at the beginning of the thesis, concerning the properties in the text itself that enable the reader to judge a text as a coherent and meaningful piece of writing, and the features of the source text that an abridgement should retain (See Section 1.3 of Chapter 1). To answer these questions, this thesis embarked on the exploration of text organisation from the point of view of the triple interface of lexical patterning, key words and the Theme-Rheme system in the written text. It endeavoured to explore the three-way relationships amongst these facets of text, and the implications of the properties of these relationships for a better understanding of text organisation.

The present study has attained the primary aim by developing a statistical analytical methodology to explore text organisation. Using this methodology, this study has discovered a close relationship between lexical patterning and the Theme-Rheme system in text. It has also discovered a close relationship between the distribution of key words and the Theme-Rheme system, key words being a special type of lexical item. Moreover, it has discovered that key-word link density is a more sensitive and more reliable indicator of the three-way relationships of lexical patterning, key word distribution and the Theme-Rheme system in text organisation. Importantly, these relationships *were quantitatively* described on the basis of a corpus of naturally written texts.

The secondary aim has also been attained to an extent. This study implies that a good summary should retain the characteristics of lexical patterning, key word distribution and thematic progression of the source text. On the basis of Hoey's

(1991a) proposal to select sentences containing an above average number of lexical links to form an abridgement of the source text, this study suggests that it is also possible to generate an automatic abridgement by programming the computer to select sentences with an above average number of lexical links, or even better, with an above average number of key-word links, in their Themes. This point will be returned to in Section 7.7. As the Theme area has been shown to be closely associated with the main concerns of the text, such an abridgement would retain the information content of the source text better.

Notwithstanding the achievements, this study still has limitations, which will be discussed in the next section.

7.5 Limitations of the study

Sinclair holds that 'the beginning of any corpus study is the creation of the corpus itself... The results are only as good as the corpus' (1991: 13). The corpus used in this study mainly has two forms of limitation: the limitation of text type in the corpus and the limitation of the size of the corpus. Firstly, all the sample texts used in the study were of only one text-type, that of the newspaper science report. Although this text type is in a sense 'domain-free', and justification was given in Chapter 1 for using this text type, it cannot be guaranteed that the results of this study would be replicated in similar analyses of other text types.

Secondly, although the corpus used in this study was already much larger than a researcher could easily handle without the aid of computer technology, it is still desirable to use an even larger corpus given the nature of this study, the capacity of the computer and the availability of electronic textual data. This limitation is a direct consequence of another limitation, that the analysis was not completely automatic and free from human intervention. Specifically, Theme delimitation in this study was carried out manually, which was rather time-consuming and made it difficult to expand the analysis to a corpus of a larger size.

Further, this study revealed an overall picture of lexical patterning and key word distribution within the Theme and Rheme areas respectively. It only briefly tackled cohesive links across Theme and Rheme areas, but did not take up issues such as the distance of the links and the effect of intervening sentences between the links. As reviewed in Section 2.4.1.3 of Chapter 2, Daneš modified his model of thematic progression to cover the factor of distance in the context, which was distinguished as wider context and immediate context. As for the notion of immediate context, or here more specifically 'co-text', Svoboda treated it as within a span of seven clauses (1981: 154, 178), Firbas (1992b) regarded it as a span of three sentences, and Cummings proposed to set it for 30 words (1995b: 451, 1998: 258). On the other hand, Hoey (1991a) noticed that lexical linking may be operative over hundreds of intervening pages within the same text. This study might have revealed more interesting properties of text organisation if it had analysed the

patterning in terms of distances between the lexical links or key-word links in the Theme and Rheme areas in the texts.

Finally, it should be mentioned that only simple repetition and some simple forms of complex repetition were recognised for the identification of lexical links and key-word links in this study. More complex forms of lexical repetition, such as paraphrase and co-reference as described in Hoey (1991a), were not covered. However, this may not be regarded as a serious limitation. Benbrahim (1996) reported that about three quarters (78%) of lexical links in his study were formed by simple lexical repetition. The addition of complex lexical repetition only increased the coverage by 16%, and the remaining 6% of links were realised by mutual paraphrase. This implies that it is generally unnecessary to count all kinds of links for retrieval of all possible bonds between the sentences. So, in this study, counting simple lexical repetition may not have distorted the overall picture of text organisation.

In spite of the limitations, this study did provide some insights into the nature of text. The next section will answer the fourth general research question raised in Chapter 1 concerning the implications of the three-way relationships *amongst lexical patterning*, key words and the Theme-Rheme system for our understanding of text organisation.

7.6 Implications of the findings

7.6.1 Implications for the Theme-Rheme system

An important implication of this study for the Theme-Rheme system is that the study of the Theme-Rheme system is more fruitful on the textual level than only on the sentential or clausal level. Only viewed from the text perspective can the potential of Theme be fully exposed.

After reviewing different approaches to the notion of Theme, this study adopted Halliday's definition that Theme is 'the point of departure of the clause' and 'what the clause is about', currently the most commonly accepted definition (see Chapter 2 of this thesis). Viewed on the clause level, there is little doubt about the first part of the definition that Theme is the 'point of departure' of the clause, which reflects the linear arrangement of information in the clause. But this statement has little significance by itself, since Theme according to Halliday is the initial component of the clause, and the initial component is necessarily the point of departure. However, viewed on the level of text, the aggregate of 'point of departure' of the constituent clauses plays a significant textual function. This study showed that the Theme area of a text is where lexical patterning is prominent and key words are concentrated; lexical items or key words in the Theme of a sentence tend to reiterate lexical items or

key words in the preceding sentences, and thus relate their host sentence to its preceding sentences in the text.

This finding has two further implications. Firstly, the thematic choices in the constituent sentences of a text reflect the overall organisation of the text. As shown in the results of the statistical analysis, the Theme area tends to contain lexical items that are reiterated and thus significant in the overall continuity of the text. Secondly, the thematic choices in the constituent sentences of a text are constrained by the overall organisation of the text. The choice of what to mean in the Themes of the constituent sentences of a text is therefore a textual phenomenon rather than a sentential phenomenon. Therefore the 'point of departure' of each sentence should be regarded as textually the cohesive point of the sentence, and accordingly the first part of the Hallidayan definition may be modified as the 'point of departure of the clause as viewed from the perspective of the clause, and a cohesive area of the text as viewed from the perspective of the text'. It should be noted here that Hasan's (1984) 'cohesive chains' may be more significant if they pass through the Theme area. This implies that thematisation itself may be a means to provide a framework for significant cohesion to take place. In other words, thematisation in each sentence may be a means to allow cohesion to contribute significantly to the coherence of the text. Choosing what to mean in the Theme of a constituent sentence may contribute to as well as be constrained by the overall organisation of the text.

As for the other part of the definition of Theme in the Hallidayan tradition, that Theme is 'what the clause is about', this study provided evidence for a new interpretation. Firstly, the 'aboutness' could not be definite unless the sentence is placed in a context. As pointed out by Winter, 'out of context, any sentence can represent answers to as many questions as it has parts' (1982: 8). In other words, a sentence out of context may be about so many things that its aboutness is indefinite. However, if the sentence is placed in context, 'the adjoining sentences, especially the preceding sentence(s), would narrow down to a specific question to which the sentence under consideration would represent a reply' (1982: 8). In written text, where features of intonation are comparatively insignificant, information flow depends heavily on the linear arrangement of words. In this case, it is typical for the writer to express his or her main concerns in the Themes of constituent sentences, and it is sensible to assume on this ground that the Theme represents what the sentence is about. This assumption was supported by the statistical analysis, where the aboutness of every sentence is indicated by the connection of its Theme to the preceding context.

Secondly, the 'aboutness' is not limited to the sentence, but may be extended to the text when the Themes of a text are viewed cumulatively. From the perspective of text, as was found in this study, the Themes of the constituent sentences together reveal the main concerns of the text; thus the aboutness extends to the whole text as opposed to individual sentences.

Thirdly, the aboutness represents an integration of all the three meta-functions: textual, interpersonal and ideational, rather than any one of the three meta-functions in isolation. In addition to the common belief that Theme mainly plays a textual function and (secondarily) an interpersonal function, this study shows that Theme is also an important area for the realisation of the ideational function. The high value of key-word link density in Theme, which reflects the high rates of repetition of lexical items and concentration of key words in Theme, eloquently suggests that Theme is indeed an area where the main content information of the text is conveyed. Admittedly, viewed from the perspective of the clause, Theme seems to be less 'newsworthy' than Rheme. But viewed from the perspective of the text, it is the Theme area that is prominent in expressing what the text is about in ideational terms.

Halliday states that 'the relation between the semantics and the grammar is one of realisation: the wording "realises", or encodes, the meaning' (1985a: xx). Meaning is encoded in the wording as an integrated whole. The choice of a particular word may mean one thing, its position in the sentence another, and its combination with other words yet another. This study supported Halliday's statement and showed that Theme as a grammatical category mainly functions as a realisation of the semantics. Thematisation in the sentences is indeed an important way to convey the main messages of the text, as Theme is the area in the text where most of the lexical repetition takes place whereby the main concerns of the writer are expressed.

When Halliday divided Theme into textual, interpersonal and topical components, he seemed to tell us that the textual and interpersonal functions were realised by grammatical items and the ideational function was realised by lexical items (1985a: 54). But this study showed that lexical items could also fulfil the textual and interpersonal functions. The fact that certain lexical items and key words tend to occur and reiterate in the Theme area of the text implies that, on the one hand, Theme is used to convey the writer's point of view; on the other hand, these items in the Theme area play an important role in guiding the readers to follow the information flow and in linking up the sentences to form a coherent text.

This study confirmed that the Theme-Rheme system is a very useful notion for the understanding of text organisation. In particular, the various types of thematic progression as proposed by Daneš (1974, 1989), Theme as a method of development of the text as proposed by Fries (1983) and the notion of discourse Theme_{M} proposed by Berry (1996) are all very useful for understanding the nature of text. But their proposals would have been more useful if they had taken account of lexical patterning and key word distribution in the text.

This study also made a contribution to the delimitation of Theme. Firstly, this study departed from Halliday's approach of limiting the scope of Theme to the first ideational element in the clause. It was found that if Theme is 'what the clause is about' in ideational terms, it should at least cover the grammatical

Subject in the affirmative sentence. When the grammatical Subject in each affirmative sentence of the text is counted as thematic, Theme will not only be 'what the clause is about' in individual clauses, the Theme area in a text will also be indicative of 'what the text is about'.

Secondly, if Theme functions as a way of linking its host sentence to the preceding sentences, it is better to exclude lexical verbs from the Theme area, because it was found in this study that verbs are not reiterated as commonly as other lexical items in the Theme area. Even if they are reiterated in form, they frequently refer to different entities in the text. Therefore, instead of adopting Berry's (1996) proposal to include lexical verbs in the Theme, this study followed Berry's (1989) position that Theme extends from the beginning of the sentence up to but does not include the lexical verb of the main clause.

Ravelli's (1991, 1995) integrated path analysis of Theme is insightful in that it views language categories as integrated and as inquiring analysis in an integrated manner. Her analysis implies that to have a better understanding of a language phenomenon, it is necessary to view the phenomenon in relation to other phenomena in complementary grammatical categories of the same rank. For example, by analysing Theme simultaneously with Mood and Transitivity, she showed that in producing a text, sentence by sentence, the decision to choose what elements to be in the Theme has priority over the decision to choose what elements to be in the Mood and Transitivity systems. Theme decision is closely related to Mood decision. This explains why Theme usually

contains Subject in the Mood system. On the other hand, from the reader's perspective, one cannot decide the boundary of Theme until the Process in the Transitivity system is reached, because the occurrence of the verb marks the ending of the nominal group which can function as Theme. This explains why Theme comes up to but does not include the lexical verb of the main clause.

Building on Ravelli's approach to Theme delimitation, this study further implied that to have a better understanding of a language phenomenon, it is necessary to view the phenomenon in relation to other phenomena in lexical as well as in grammatical categories. It predicted that the characteristics of the Theme-Rheme system will be understood better when the factors of lexical patterning and key word distribution in the text are also considered.

Ravelli's study stresses the linear characteristics of the Theme-Rheme system. This study however emphasised the non-linear characteristics of the Theme-Rheme system. Whereas clause Themes are realised by linear arrangement of information in the sentences, the thematic choices of constituent sentences are constrained by the overall organisation of the text. If viewed from a global perspective, the Theme area of a text represents a concentration of certain lexical items which convey the main concerns of the text and a network of lexical links which helps to create the coherence of the text. These characteristics are basically non-linear.

One of the foci of the present study was on the cumulative force of the clause Themes, or, to use Berry's term, clause $Theme_F$s. In doing so, the present study revealed how clause $Theme_F$s realise the discourse $Theme_M$s in the text. On the other hand, the present study also focused on the aggregate of lexical items and their distribution in the complete text, rather than on individual lexical items in isolated sentences or sentence pairs. The next sub-section will discuss this point in more detail.

7.6.2 Implications for lexical patterning

The study of lexical patterning as a textual phenomenon has undergone three stages. In the first stage, it was only marginally noticed, as is shown in Halliday and Hasan (1976). In the second stage, its importance began to be recognised, as shown in Hasan (1984) and Halliday and Hasan (1985). In the third stage, it was regarded as central to the study of language, as proposed in Hoey (1991a). This study followed Hoey's argument that lexical patterning is an important measure to create coherence in text. In real-life communication, the writer usually tells something new on the basis of something (assumed to be) already known to the reader. The known (or given) information is commonly provided by means of lexical patterning as well as thematic progression in the text.

Winter (1982) argues that a word cannot be regarded as having meaning unless it is placed in a context which enables it to have a definite meaning. For example, the words 'rats' and 'bats' do not mean anything except their

dictionary definitions, until they are put into a sentence which relates them significantly to each other. Winter holds that 'the sentence is the minimum grammatical context for the word to have meaning as a word' (1982: 177). Further, in an attempt to define the notion of sentence, Winter claims that a sentence cannot be fully understood out of context. The information in an isolated sentence is usually insufficient and a reader would necessarily refer to the context in order to understand the sentence. Therefore, ultimately, only when a word is placed in the context of a text can it be regarded as having a definite meaning.

In general agreement with Winter's observation, Scott (1998) claims that a sentence out of context has only minimum aboutness. It is not fully comprehensible and not easy to use in real-life communication. The reader can guess but cannot be sure what the sentence is about. Consequently, Scott claims, text is the only language unit which has the properties of completeness and aboutness (1998: 155). A naturally produced written text, without the support of intonation signals, normally needs to be self-sufficient in order to be understood. Lexical patterning is an important means to build up the context. This study showed that the cohesive strength of lexical patterning may vary in different areas of the text. The lexical links may be more significant to the aboutness if they run through the Theme area of the text. Therefore, if coherence is the reader's perception of the text property, then the writer can simultaneously make use of lexical patterning and thematical patterning to enable the reader to perceive the text as coherent.

Hoey (1991a) took the sentence as the basic unit of lexical analysis. But he did not distinguish the positions of the lexical items within the sentence. This study mainly followed Hoey in regarding lexical patterning as a central driver to text organisation, but it went further by distinguishing the intra-sentential positions of the lexical items, or specifically, by distinguishing lexical patterning in Theme from lexical patterning in Rheme. This implies that the componential view of cohesion may be justifiable in that while cohesion may be viewed as linking up of whole messages, some components of the messages play a more significant role than others in terms of cohesion. The findings of this study imply that while the notion of patterns of lexis in text is very useful for identifying text organisation, it is especially so when patterns of lexis are studied together with patterns of thematic progression.

Further, this study suggested that the positioning of lexical items in the sentences is more of a textual phenomenon than a sentential phenomenon. In an isolated sentence, the positioning of a lexical item is mainly constrained by syntactic rules. But once the sentence is part of a text, the positioning of the lexical item would be constrained in addition by the context of preceding sentences. This tendency cannot be observed in an isolated sentence, but it becomes obvious when the patterning is viewed on the text level, and even more obvious when a number of texts are analysed, as in this study. Although this study did not tackle the issue of what particular lexical items tend to occur in what particular part of the sentence, it was assumed that they would occur in recognisable patterns.

7.6.3 Implications for key words

The notion of key words was initially proposed to reveal the aboutness of a text in terms of its ideational content. This study found that the aboutness may not only be ideational, but also be interpersonal and textual. The concentration of key words, the high value of keyness, and high density of key-word links in the Theme area imply that, in addition to conveying content information, key words may also reveal the writer's intention and attitudes towards the message in the sentence, and have a strong cohesive power to link the sentences together to form a coherent text.

In this study, key words were lexical items selected from the text by comparison with a much larger reference corpus of texts. In this sense, they are a special kind of lexical item which mainly highlight the information content of the text. While lexical patterning conflates with thematic progression, key words emphasise the conflation, and key-word link density represents an even more striking picture of the conflation. This finding implies that patterns of information distribution are describable in quantitative terms and may be observed by the conflation of key word distribution and key-word link density with thematic progression.

This study shows that as far as text organisation is concerned, just as lexical patterning in the Theme area is more important than lexical patterning in the Rheme area, key-word link density may be an even better indicator of the

relationships between sentences in the text. In the corpus of sample texts used in this study, the comparative value of key-word link density tends to be significantly higher in Theme than in Rheme. This has at least two further implications. On the one hand, key-word link density may clearly indicate the interaction between lexical patterning or key word distribution and thematic progression, thus providing convincing evidence of the aboutness of the text. On the other hand, key-word link density may possibly replace cohesion as currently defined, in that it represents significant cohesion as opposed to insignificant cohesion. The higher the key-word link density, the more focused the text might be. The higher the key-word link density, the more cohesive the text might be.

It is worth mentioning at this point that the notion of key words as a linguistic category is still in its infancy, and there is still much room for refinement and modification concerning their definition and identification. This study found a concentration of key words and consequently a high density of key-word links in the Theme area of the text, but it did not explore in depth what impact each set of key-word links has on the text and why this is so.

Having discussed the implications of this study for the Theme-Rheme system, lexical patterning and key word distribution respectively, it is now possible to address issues more directly related to the implications for text organisation. This leads us to the next sub-section.

7.6.4 Implications for text organisation

This study shed some new light on the understanding of text organisation. Firstly, in order to understand how language works, as already noted, text is a more appropriate unit of study than sentence or clause. This is because that text is not a single phenomenon. It is an integration of many factors. The patterning of lexical items and distribution of key words could all be viewed as a manifestation of text organisation.

In previous studies, text was seldom used as an independent unit of analysis. Most studies either focus on the sentence or clause, or focus on a short stretch of text such as a paragraph. Even within corpus linguistics, the focus has often been placed on incomplete text fragments, individual lexical items or concordance lines for the study of collocation. In contrast, this study viewed text as an independent level of the language, a unit to complete a communicative function, and thus a legitimate unit for linguistic analysis. The statistical data were calculated on the basis of the text as the unit of analysis.

As noted in Chapter 1, text is a language event which fulfils some communicative task. This means that a text must be about something, or it has the quality of 'aboutness'. Therefore, a pre-requirement to appreciate a text is to know what the text is about. Scott (1998) discussed two positions to textual aboutness. One is what he called the individual, subjective position, which holds that aboutness is reader-dependent. In this position, a text can be about

anything; it solely relies upon the reader to decide what it is about. Another position is the collective, objective position, which suggests that what the text is about is signalled in the text by the writer and that the reader perceives what the text is about on the basis of the signals. The former refers to the implications the reader draws from reading the text; the latter refers to the explicit choice and arrangement of components of the text. This study supported the claim that the reader plays an active role in the understanding of the text, but the reader relies on the signals in the text supplied by the writer. In short, text has the quality of aboutness, and the aboutness is signalled in the text simultaneously by many devices. The integration of lexical patterning, key word distribution and the Theme-Rheme system in text clearly signals the intention and content information of the text.

This study strongly supported Hoey's claim that 'texts are not organised linearly and that motivated selections from a text may make sense' (1991a: 187). The findings of this study suggest that text may be viewed as networks of related meanings. Among them, some meanings may be prioritised over other meanings. This study argued that lexical items in a text are repeated differently in different areas of the text, both in terms of quantity and quality. Information is distributed unevenly but in recognisable patterns in the text. It should therefore be possible to make use of the patterns as described in this study to construct computer programs for making summaries.

The concentration of counter-examples involving Sample Text C8 implies that text size may be an important factor in affecting the interplay of patterns of lexis and key words with thematic progression. Possibly the effects may be owing to the style of individual authors. Nevertheless, it may be assumed from this study that the net of lexical patterning would be more comprehensive with longer texts, and the aboutness would be clearer with more focused texts. Although this in part supports the proposal that text is organised rather than structured, how much freedom the author has in organising the text remains an area of investigation for further research.

7.7 Further research

In answering the general research questions, this study left many questions unanswered. There are two major directions for further exploration. One is mainly theoretical, the other mainly practical.

On the theoretical line, there are three areas worth exploring. Firstly, more text types may be analysed than did in this study. As mentioned in Section 7.5, only one text type, namely the newspaper science report, was analysed in this study. Hoey (1991a) found that his system of lexical repetition did not work well with narrative texts, though it has limited applicability (Hoey 1994). Phillips (1985) likewise found that his findings concerning the organising properties of collocation in science text did not apply to narrative text. So it might be fruitful

to analyse narrative texts in further research to see if there is also a correlation amongst lexical patterning, key word distribution and the Theme-Rheme system in that text type.

Secondly, further research may carry out analysis on language units lower than complete texts and explore the interrelationships between lexis and text segments. Renouf points out that 'a text is not always uniformly about one topic', and 'the flow from one topic to another is characteristic of many text types' (1993: 88). So when one talks of 'what the text is about', one often means a very broad or rather indefinite semantic area of the text. Therefore the knowledge of what each of the constituent segments of the text is about would help one to know more accurately what the complete text is about. Phillips (1985) carried out research into the relationship between lexis and chapters. Berber Sardinha (1997) likewise applied the notions of lexical links to the automatic identification of text segments. Further research may explore the relationship between lexis and text segments, such as sections of longer texts or paragraphs of shorter texts, and trace the shift of information focus in different segments as manifested by the correlation of lexical patterning, key word distribution and the Theme-Rheme system. This may present a more accurate picture of the aboutness of the text.

Further research may also expand the analysis to the interrelationship of a large group of texts. Scott (1997, 1998) reported on studies of inter-text relationships reflected by 'key-key words', i.e. key words shared by a group of texts. If the

correlation of lexical patterning, key word distribution and the Theme-Rheme system were analysed on a large corpus of texts of a certain text type, it might reveal the inter-textual relationship amongst these texts and possibly present a popular organisation pattern of the text type in question. Moreover, in shedding light on the property of key words, this study also raised questions that await further exploration. For example, does the category of key words itself represent a cohesive device? Does key-word link density conflate with cohesive chains? Are there other ways of selecting key words which may better reflect the aboutness of the text? All these may be worth exploring in further studies.

Thirdly, this study discovered a general trend for key-word link density to be higher in the Theme area than in the Rheme area of the texts, but it did not explore the implications of the different ratios of key-word link density for the aboutness of different texts. It may be hypothesised that a text with a higher key-word link density in the Theme area will be more focused than a text with a lower key-word link density in the Theme area. The former will be more closely associated with the aboutness of the text than the latter. Further research may pursue this issue and find the relationship between different ratios of lexical patterning and key word distribution in the Theme and Rheme areas and the perception of the aboutness of the text. Moreover, this study did not compare key-word link density among different sample texts, nor did it tackle the question of distances between the lexical links and the impact of intervening sentences on the reader's perception of topic focus of the text. The number of intervening sentences between lexical links may also be used as a

measurement of information focus of the text. So further research may pursue this line of inquiry and measure the extent of aboutness of the text.

On the practical side, there are at least three areas for further research. Firstly, further research may seek to design more sophisticated computer programs to recognise more complex repetition patterns than simple lexical repetition. For example, lexicalisation of pronouns might be automated, or at least computer aided. In this study, only pronouns in Sample Text A1 were lexicalised, and this was done manually. As most pronouns with references across the sentence boundary are in the Theme area of the sample texts used in this study, more cohesive links in the Theme area might have been established if all pronouns were lexicalised. Recent research in automatic lexicalisation of pronouns shows that it is of course not a simple task (McCarthy 1994, Kehler 1997, McEney et. al. 1997, Mitkov 1998). Intuitively, a pronoun often stands either for a nominal group immediately preceding the pronoun, or for a nominal group in the Theme of the immediately preceding sentence. Further research might set up hypotheses on the basis of the intuition and test the hypotheses by obtaining quantitative evidence, for the purpose of designing computer programs for automatic lexicalisation of pronouns.

Secondly, further research may be carried out to design computer programs for automatic Theme delimitation. One possible solution may be exploiting the parsing facility to automatically tag the part of speech of each word, and then selecting the stretch of words between a sentence end marker, such as a full-

stop, and the first verb in the main clause. It may also be possible to program the computer to identify lexical verbs or modal primary auxiliaries and to discount any word occurring immediately after relative pronouns such as 'that', 'who' or 'which'. Even if no computer program is able to produce accurate Theme delimitation in the near future, it should at least be possible for the computer to accomplish part of the job and leave the human researcher to edit the computer output. This will save a lot of the researcher's time and produce more consistent output, so as to allow the study of the Theme-Rheme system to expand to a larger corpus.

Thirdly, further research may be carried out to design new computer programs or improve existing ones for automatic summarisation. As shown in Chapter 4 of this thesis, lexical links in the Theme area of the text might be doubly powerful in creating coherence in the text. Therefore, it is possible to credit more weight to lexical links in the Theme area of the text, counting every single T-T link as if it equals two links. There may be two ways to improve Hoey's system of abridgement. The first way is to lower Hoey's (1991a) threshold of three links for the sentences to be bonded down to two links, on condition that one of the links is a T-T link. As shown in Chapter 4, sentence pairs and groups collected in such a way may still read coherently in spite of a lower threshold of links. This may allow more sentences to be included in the abridgement. The second way is to program the computer to count T-T links as double links, but maintain the threshold of three links for the sentences to be bonded. Therefore, a sentence pair with more T-T links would have priority to

be selected in the abridgement over a sentence pair with fewer T-T links. This may produce more reliable abridgements by solving a common problem with current abridgement programs, namely that longer sentences in the source text are more likely to be selected than shorter sentences. This is because longer sentences have a higher likelihood to contain more lexical items and thus more repetition of the lexical items. Sentences are made longer often by an increase of words in the Rheme rather than by an increase of words in the Theme. Therefore, doubling the value of T-T links could reduce the likelihood of longer sentences being selected while increasing the likelihood of the selection of topically focused sentences.

7.8 Final remarks

In an exploration of the Theme-Rheme system and patterns of lexis in text, the present study has succeeded in discovering an interrelationship between the two aspects of language, thus shedding some light on the understanding of text organisation, as well as clause or sentence organisation. Returning to Mathesius's metaphor, cited at the beginning of this thesis, the present study is only an attempt to join the attack on the fortress of language from the interface of lexis, grammar and text. It is believed that with the impact of what Leech described as 'corpus revolution' (1997: 22) and the development of more sophisticated analytical tools to handle more evidence of language data, the attack on the fortress of language will be more effective.

Bibliography

Aarts, Jan and Willem Meijs (1990). (eds.). *Theory and practice in corpus linguistics.* Amsterdam: Rodopi.

Aijmer, Karin and Bengt Altenberg (1991). (eds.) *English corpus linguistics.* London: Longman.

Andor, József (1989). 'Strategies, tactics and realistic methods of text analysis', in Wolfgang Heydrich, Fritz Neubauer, János S. Petöfi and Emel Sözer (eds.). pp. 28-36.

Bacon, Francis (1597-1625). 'Of studies'. Quote in *The Columbia dictionary of quotations* (1993). Columbia University Press. .

Baker, Mona, Gill Francis and Elena Tognini-Bonelli. (1993). (eds.). *Text and technology: in honour of John Sinclair.* Philadelphia/Amsterdam: John Benjamins Publishing Company.

Bargiela-Chiappini, F. and S. Harris (1997). (eds.). *The language of business: an international perspective.* Edinburgh: Edinburgh University Press .

Bazell, Charles E., John C. Catford, M. A. K. Halliday and Robert H. Robins (1966). (eds.). *In memory of J. R. Firth.* London: Longman.

Bell, Alan (1991). *The language of news media.* Oxford: Blackwell Publishers. (reprinted 1993).

Benbrahim, Mohamed (1996). *Automatic text summarisation through lexical cohesion analysis.* PhD thesis. Artificial Intelligence Group, Department of Mathematical and Computing Sciences. Surrey: University of Surrey.

Benbrahim, Mohamed and Khurshid Ahmad (1994). 'Computer-aided lexical cohesion analysis and text abridgement'. *Knowledge processing,* 18. Surrey: University of Surrey.

Beneš, E. (1968). 'On two aspects of functional sentence perspective'. *TLP* 3. pp. 267-274.

Benson, James D., Michael J. Cummings and William S. Greaves (1988). (eds.). *Linguistics in a systemic perspective.* Amsterdam: John Benjamins Publishing Company.

Berber Sardinha, A. P. (1993). 'Lexis in annual reports: paragraph linkage and cohesion distance'. (Working Paper 5). *DIRECT Papers.* CEPRIL, PUC-SP, Sao Paulo, Brazil/AELSU, English Department, University of Liverpool.

Berber Sardinha, A. P. (1995). 'Annual business reports sections: key words'. (Working paper 25). *DIRECT papers.* CEPRIL, PUC-SP Brazil/AELSU, English Department, University of Liverpool.

Berber Sardinha, A. P. (1997). *Automatic identification of segments in written texts.* PhD thesis. Liverpool: University of Liverpool.

Berry, Margaret (1975). *An introduction to systemic linguistics. Vol. 1. Structures and systems.* London: Batsford Ltd.

Berry, Margaret (1977). *An introduction to systemic linguistics. Vol. 2. Levels and links.* London: Batsford Ltd.

Berry, Margaret (1982). 'Review of Halliday (1978)'. *Nottingham linguistic circular 11.* pp. 64 -94.

Berry, Margaret (1989). 'Thematic options and success in writing', in Christopher S. Butler, R. A. Cardwell and Joanna M. Channell (eds.). pp. 62-80 .

Berry, Margaret (1990). *Thematic analysis and stylistic preferences.* Paper read to the 17th International Systemic Congress, University of Stirling, 3rd-7th July 1990.

Berry, Margaret (1995). 'Towards a study of the relevance of thematic options to success in writing', in Mohsen Ghadessy (ed.). pp. 1-19.

Berry, Margaret (1996). 'What is Theme - a(nother). personal view', in Margaret Berry, Christopher Butler, Robin Fawcett and Guowen Huang (eds.). pp. 1-64.

Berry, Margaret, Christopher Butler, Robin Fawcett and Guowen Huang (1996). (eds.). *Meaning and form: systemic functional interpretations.* (Vol. III of Meaning and choice in language: Studies for Michael Halliday). Norwood, New Jersey: Ablex.

Berry, R. (1986). *How to write a research paper.* Oxford: Pergamon Press

Biber, Douglas (1993). 'Using register-diversified corpora for general language studies', in *Computational Linguistics.* Vol. 19, No. 2: 219-241.

Biber, Douglas (1995). *Dimensions of register variation - a cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad and Randi Reppen (1998). *Corpus linguistics, investigating languaeg structure and use.* Cambridge: Cambridge University Press.

Bloor, Meriel and Thomas Bloor (1992). 'Given and new information in the thematic organisation of text: an application to the teaching of academic writing'. *Occasional papers in systemic linguistics.* 6. pp. 33-43.

Bolc, Leonard (1980). (ed.). *Natural language question answering systems.* München Wien: Carl Hanser Verlag.

Bolivar, Adriana (1986). *Interaction through written text: a discourse analysis of newspaper editorials.* PhD thesis. Birmingham: University of Birmingham.

Booth, B. (1987). 'Text input and pre-processing: dealing with the orthographic form of texts', in Roger Garside, Geoffrey N. Leech and Geoffrey Sampson (eds.). pp. 97-109.

Brend, Ruth M. (1972). (ed.). *Kenneth L. Pike - selected writings.* The Hague: Mouton.

Brown, Gillian and George Yule (1983). *Discourse analysis.* Cambridge: Cambridge University Press.

Burnard, Lou (1992). 'Tools and techniques for computer-assisted text processing', in Christopher S. Butler (ed.).

Butler, Christopher S. (1985a). *Systemic linguistics: theory and applications.* London: Batsford Academic.

Butler, Christopher S. (1985b). *Computers in linguistics.* Oxford: Blackwell Publishers.

Butler, Christopher S. (1992). (ed.). *Computers and written texts.* Oxford: Blackwell Publishers.

Butler, Christopher, R. A. Cardwell and Joanna M. Channell (1989). (eds.). *Language and literature - theory and practice: a tribute to Walter Grauberg (University of Nottingham monographs in the humanities VI).* Nottingham: University of Nottingham.

Carmen Rosa Caldas-Coulthard and Malcolm Coulthard (1996). (eds.). *Texts and practices.* London: Routledge.

Carrell, Patricia L. and J. Eisterhold (1988). 'Schema theory and ESL reading pedagogy', in Patricia L. Carrell, Joanne Devine and David E. Eskey (eds.). pp. 73-92.

Carrell, Patricia L., Joanne Devine and David E. Eskey (1988). (eds.). *Interactive approaches to second language reading.* Cambridge: Cambridge University Press.

Carter, Ronald A. (1987). *Vocabulary.* London: Allen & Unwin.

Carter, Ronald A. (1995). *Keywords in language and literacy.* London: Routledge.

Carter, Ronald A. and McCarthy, Michael J. (1988). *Vocabulary and language teaching.* London: Longman.

Chafe, Wallace (1991). 'Grammatical subjects in speaking and writing'. *Text* 11/1. pp. 45-72.

Chafe, Wallace (1992). 'The flow of ideas in a sample of written language', in William C. Mann and S. A. Thompson (eds.). pp. 267-294.

Charolles, Michael (1989). 'Coherence as a principle in the regulation of discursive production', in Wolfgang Heydrich, Fritz Neubauer, János S. Petöfi and Emel Sözer (eds.). pp. 3-15.

Chomsky, Noam (1957). *Syntactic Structures.* The Hague: Mouton.

Clear, Jeremy (1987). 'Computing', in John M. Sinclair (ed.). pp. 41-61.

Collier, Alex (1998). *The automatic selection of concordance lines.* PhD thesis. Liverpool: University of Liverpool.

Collins, Heloisa and Mike Scott (1997). 'Lexical landscaping in business meetings'. *DIRECT papers 32.* CEPRIL, PUC-SP Brazil/ Liverpool: AELSU, English Department, University of Liverpool.

Collison, R. L. (1971). *Abstracts and abstracting services.* Oxford: ABC Clio Press.

Conte, Maria-Elisabeth, János S. Petöfi and Emel Sözer (1989). (eds.). *Text and discourse connectedness.* Amsterdam: John Benjamins Publishing Company.

Cook, Guy and Barbara Seidlhofer (1995). (eds.). *Principle and practice in applied linguistics: studies in honour of H.G. Widdowson.* Oxford: Oxford University Press,

Coulthard, Malcolm (1985). *An introduction to discourse analysis.* London: Longman.

Coulthard, Malcolm (1986). (ed.). *Talking about text: studies presented to David Brazil on his retirement, Discourse analysis monographs No 13.* Birmingham: English Language Research, University of Birmingham.

Coulthard, Malcolm (1994). (ed.). *Advances in written text analysis.* London: Routledge.

Coulthard, Malcolm and Martin Montgomery (1981). (eds.). *Studies in discourse analysis.* London: Routledge and Kegan Paul.

Couture, Barbara (1986). (ed.). *Functional approaches to writing: research perspectives.* London: Frances Printer Publishers.

Cummings, Michael J. (1995a). 'A systemic functional approach to the thematic structure of the old English clause', in Ruqaiya Hasan and Peter H. Fries (eds.). pp. 275-316.

Cummings, Michael J. (1995b). 'Structural semantics as the basis for Theme/Rheme'. *LACUS Forum* 21. pp. 443-459.

Cummings, Michael J. (1998). 'Lexical repetition and the Given/New distinction in written English'. *LACUS Forum* 24. pp. 253-266.

Daneš, František (1970). 'One instance of Prague School methodology: functional analysis of utterance and text', in Paul L. Garvin (ed.). pp. 132-146.

Daneš, František (1974a). 'Functional sentence perspective and the organisation of text', in František Daneš (ed.). pp. 106-128.

Daneš, František (1974b). (ed.). *Papers on functional sentence perspective.* Prague: Academia.

Daneš, František (1987). 'Sentence patterns and predicate classes', in Ross Steele and Terry Threadgold (eds.).

Daneš, František (1989). '"Functional sentence perspective" and text connectedness', in Maria-Elisabeth Conte, János S. Petöfi and Emel Sözer (eds.). pp. 23-31.

Davies, Florence (1988). 'Reading between the lines: thematic choice as a device for presenting writer viewpoint in academic discourse'. *ESPecialist* 9/1-2. pp. 173-200.

Davies, Martin and Louise J. Ravelli (1992). (eds.). *Advances in systemic linguistics: recent theory and practice.* London: Pinter Publishers.

de Beaugrande, R. and W. Dressler. (1981). *Introduction to text linguistics.* London: Longman.

Downing, Angela (1990). 'On topical Theme in English'. Paper presented at the Seventh International Systemic Congress. Stirling. 4-7 July 1990.

Downing, Angela (1991). 'An alternative approach to theme: a systemic-functional perspective'. *Word* 42/2. pp. 119-143.

Dubois, Betty Lou (1987). 'A reformulation of thematic progression typology'. *Text* 7/2. pp. 89-116.

Dubois, Betty Lou (1990). 'Thematisation across machine and human translation: English to French'. *IRAL* 28/1. pp. 45- 51.

Dunning, Ted (1993). 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* 19/1. pp. 61-74.

Dyer, Michael G. (1983). *In-depth understanding: a computer model of integrated processing and memory for narrative comprehension.* Cambridge, Mass.: MIT Press.

Eggins, Suzanne (1994). *An introduction to systemic functional linguistics.* London: Pinter Publishers.

Eiler, Mary Ann (1986). 'Thematic distribution as a heuristic for written discourse function', in Barbara Couture (ed.). pp. 49-68.

Enkvist, Nils E. (1973). 'Theme dynamics and style: an experiment'. *Studia Anglica Posnaniensia* 5. pp. 127-135.

Enkvist, Nils E. (1991). 'Discourse strategies and discourse types', in Eija Ventola (ed.). pp. 3-22.

Firbas, Jan (1957). 'Kotázce nezákladových podmetu v soucasné anglictine (On the problem of non-thematic subjects in contemporary English)'. *Casopis pro moderní filologii* 39. 22-42. pp. 165-173.

Firbas, Jan (1964). 'On defining the Theme in functional sentence analysis'. *Travaux Linguistiques de Prague* 1. pp. 267-280.

Firbas, Jan (1975). 'On the thematic and non-thematic section of the sentence', in Hakan Ringbom (ed.). pp. 317-334.

Firbas, Jan (1992a). *Functional sentence perspective in written and spoken communication.* Cambridge: Cambridge University Press.

Firbas, Jan (1992b). 'On some basic problems of functional sentence perspective', in Martin Davies and Louise J. Ravelli (eds.). pp. 167-188.

Firbas, Jan (1995). 'On the thematic and the rhematic layers of a text', in B. Warvik, S-K. Tanskanen and R. Hiltunen (eds.). pp. 59-72.

Flood, James (1984). (ed.). *Understanding reading comprehension: cognition, language and the structure of prose.* Newark, Delaware: International Reading Association.

Fox, Barbara A. (1987). *Discourse structure and anaphora: written and conversational English.* Cambridge: Cambridge University Press.

Fox, Barbara A. (1996). (ed.). *Studies in anaphora.* Amsterdam: John Benjamins Publishing Company.

Francis, Gill (1986). *Anaphoric nouns, Discourse analysis monographs No 11.* Birmingham: English Language Research, University of Birmingham.

Francis, Gill (1990). 'Theme in the daily press'. *Occasional papers in systemic linguistics* 4. pp. 51-87.

Francis, Gill (1994). 'Labelling discourse: an aspect of nominal group lexical cohesion', in Malcolm Coulthard (ed.). pp. 83-101.

Fries, Peter H. (1981). 'On the status of Theme in English: arguments from discourse'. *Forum Linguisticum* 6/1. pp. 1-38. (reprinted 1983 in J. S. Petöfi and Emel Sözer (eds.). pp. 116-52).

Fries, Peter H. (1986). 'Language features, textual coherence and reading'. *Word* 37/1-2. pp. 13-29.

Fries, Peter H. (1992a). 'The structuring of information in written English text'. *Language sciences* 14/4. pp. 461-488.

Fries, Peter H. (1992b). 'Lexico-grammatical patterns and the interpretation of texts'. *Discourse processes* 15. pp. 73-91.

Fries, Peter H. (1994). 'On Theme, Rheme and discourse goals', in Malcolm Coulthard (ed.). pp. 229-249.

Fries, Peter H. (1995a). 'Patterns of information in initial position in English', in Peter H. Fries and Michael Gregory (eds.). pp. 47-65.

Fries, Peter H. (1995b). 'Themes, methods of development and texts', in Ruqaiya Hasan and Peter H. Fries (eds.). pp. 317-359. .

Fries, Peter H. and G. Francis. (1992). 'Exploring Theme: problems for research'. *Occasional papers in systemic linguistics* 6. pp. 45-59.

Fries, Peter H. and Michael Gregory (1995). (eds.). *Discourse in society: systemic functional perspectives (Meaning and choice in language: Studies for Michael Halliday).* Norwood, New Jersey: Ablex Publishing Corporation.

Garside, Roger, Geoffrey N. Leech and Geoffrey Sampson (1987). (eds.). *The computational analysis of English: a corpus-based approach.* London: Longman.

Garvin, Paul L. (1970). (ed.). *Method and theory in linguistics.* Mouton: the Hague.

Gibson, Tim R. (1989). 'Review of the linguistics and psychology literature'. *BT automatic text summarisation (Project 322). Deliverable 12.*

Gibson, Tim R. (1993). 'Towards a discourse theory of abstracts and abstracting'. *Monographs in systemic linguistics No 5.* Nottingham: University of Nottingham.

Giora, Rachel (1983). 'Segmentation and segment cohesion: on the thematic organisation of the text'. *Text* 3/2. pp. 155-181.

Gordon H. Tucker (1996). 'Cultural classification and system networks: a systemic functional approach to lexical semantics', in Margaret Berry, Christopher Butler, Robin Fawcett and Guowen Huang (eds.). pp. 533-566.

Gosden, Hugh (1992a). 'Research writing and the NNSs.: from the editors'. *Journal of second language writing* 1/2. pp. 132-39. .

Gosden, Hugh (1992b). 'Discourse functions of subject in scientific research articles'. (draft version).

Gosden, Hugh (1993). 'Discourse functions of Subject in scientific articles'. *Applied linguistics* 14. pp. 56-75.

Gosden, Hugh (1994). *A genre-based investigation of Theme: product and process in scientific research articles written by NNS novice researchers.* PhD thesis. Liverpool: University of Liverpool.

Graetz, Naomi (1985). 'Teaching EFL students to extract structural information from abstracts' in J. M. Ulijn and A. K. Pugh (eds.).

Greenbaum, Sidney and Randolph Quirk (1970). *Elicitation experiments in English: linguistic studies in use and attitude.* London: Longman.

Grishman, Ralph (1986). *Computational linguistics: an introduction.* Cambridge: Cambridge University Press.

Grishman, Ralph and Richard Kittredge (1986). (eds.). *Analyzing language in restricted domains -sublanguage description and processing.* Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers.

Halliday, M. A. K. (1967a). 'Notes on transitivity and Theme in English. Part 1'. *Journal of linguistics* 3/1. pp. 37-198.

Halliday, M. A. K. (1967b). 'Notes on transitivity and Theme in English: Part 2'. *Journal of linguistics* 3/2. pp. 199-244.

Halliday, M. A. K. (1967c). 'Notes on transitivity and Theme in English. Part 3'. *Journal of linguistics* 4/1. pp. 179-216.

Halliday, M. A. K. (1970). 'Language structure and language function', in John Lyons (ed.). pp. 140-164.

Halliday, M. A. K. (1977). 'Ideas about language: aims and perspectives in linguistics'. *Series occasional paper No 1 by Applied Linguistics Association of Australia.*

Halliday, M. A. K. (1985a). *An introduction to functional grammar.* London: Edward Arnold.

Halliday, M. A. K. (1985b). *Spoken and written language.* Geelong, Vic.: Deakin University Press (Reprinted 1989. Oxford: Oxford University Press).

Halliday, M. A. K. (1988). 'On the ineffability of grammatical categories', in James D. Benson, Michael J. Cummings and William S. Greaves (eds.). pp. 27-51.

Halliday, M. A. K. (1992). 'Some lexicogrammatical features of the Zero Population Growth text', in William C. Mann and S. A. Thompson (eds.). pp. 327-358.

Halliday, M. A. K. (1993). 'Quantitative studies and probabilities in grammar', in Michael P. Hoey (ed.). pp. 1-25.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.

Halliday, M. A. K. and Ruqaiya Hasan (1976). *Cohesion in English.* London: Longman.

Halliday, M. A. K. and Ruqaiya Hasan (1985). *Language, context and text: aspects of language in a social-semiotic perspective.* Geelong, Vic.: Deakin University Press. (Reprinted 1989. Oxford: Oxford University Press).

Harris, M. D. (1985). *Introduction to natural language processing.* Reston, Virginia: Reston Publishing Co. Inc.

Hartnett, Carolyn G. (1986). 'Static and dynamic cohesion: signals of thinking in writing', in Barbara Couture (ed.). pp. 142-153.

Hasan, Ruqaiya (1984). 'Coherence and cohesive harmony', in J. Flood (ed.). pp. 181-220.

Hasan, Ruqaiya and Peter H. Fries (1995). (eds.). *On Subject and Theme, a discourse functional perspective.* Amsterdam: John Benjamins Publishing Company .

Hatch, E. and Farhady, H. (1982). *Research design and statistics for applied linguistics.* New York: Newbury House Publishers.

Hausenblas, K. (1964). 'On characterisation and classification of discourses'. *TLP* 1. PP. 67-84.

Heger, Klaus (1976). *Monem, wort, satz, und text [Word, moneme, sentence, and text].* Tübingen: Niemeyer .

Heydrich, Wolfgang, Fritz Neubauer, János S. Petöfi and Emel Sözer (1989). (eds.). *Connexity and coherence: analysis of text and discourse.* Berlin: Walter de Gruyter .

Hoey, Michael P. (1979). 'Signalling in discourse'. *Discourse analysis monographs No 6.* Birmingham: English Language Research, University of Birmingham.

Hoey, Michael P. (1983). *On the surface of discourse.* London: George Allen and Unwin.

Hoey, Michael P. (1985). 'The paragraph boundary as a marker of relations between the parts of a discourse'. *M.A.L.S. Journal 10, Special issue in honour of E. O. Winter.* Birmingham: University of Birmingham. pp. 96-107.

Hoey, Michael P. (1986). 'The discourse colony: a preliminary study of a neglected discourse type', in Malcolm Coulthard (ed.). pp. 1-26.

Hoey, Michael P. (1988). 'Writing to meet the reader's needs: text patterning and reading strategies'. *Trondheim papers in applied linguistics* 4. Trondheim: University of Trondheim. pp. 51-73.

Hoey, Michael P. (1991a). *Patterns of lexis in text.* Oxford: Oxford University Press.

Hoey, Michael P. (1991b). 'Another perspective on coherence and cohesive harmony', in Eija Ventola (ed.). pp. 385-414.

Hoey, Michael P. (1991c). 'The matrix organisation of narrative and non-narrative text'. *Proceedings of the fifth symposium on the description and/or comparison of English and Greek.* Thessaloniki: Aristotle University . pp. 215-254.

Hoey, Michael P. (1993). (ed.). *Data, description, discourse: papers on the English language in honour of John M. Sinclair on his sixtieth birthday.* London: HarperCollins Publishers.

Hoey, Michael P. (1994). 'Patterns of lexis in narrative: a preliminary study', in Sanna-Kaisa Tanskanen and Brita Warvik (eds.). pp. 1-39.

Hoey, Michael P. (1995). 'Cohesive words: a paper of consequence'. Paper delivered at the 'Words' Symposium, Lund. 1995.

Hoey, Michael P. (1996a). 'A clause-relational analysis of selected dictionary entries: contrast and compatibility in the definitions of "man" and "woman"', in Carmen Rosa Caldas-Coulthard and Malcolm Coulthard (eds.). pp. 150-165.

Hoey, Michael P. (1996b). *The inseparability of word, grammar and text.* Inaugural lecture. Liverpool: University of Liverpool.

Hoey, Michael P. (1997). 'The discourse's disappearing (and reappearing). subject: an exploration of the extent of intertextual inference in the production of texts', in Karl Simms (ed.). pp. 245-264.

Hoey, Michael P. and Eugene O. Winter. (1986). 'Clause relations and the writer's communicative task', in Barbara Couture (ed.). pp. 120-141.

Hu, Zhuanglin and Fang Yan (1997). (eds.). *Advances in functional linguistics in China.* Beijing: Tsinghua University Press.

Huang, Guowen (1996). 'Experiential enhanced Theme in English', in Margaret Berry, Christopher Butler, Robin Fawcett and Guowen Huang (eds.). pp. 65-112.

Huang, Guowen (1997). 'A thematic analysis of the existential process in English', in Hu Zhuanglin and Fang Yan (eds.). pp. 106-119.

Huddleston, Rodney D. (1988). 'Constituency, multi-functionality and grammaticalization in Halliday's functional grammar'. [review article]. *Journal of linguistics* 24. pp. 137-174.

Huddleston, Rodney D. (1991). 'Further remarks on Halliday's functional grammar: a reply to Matthiessen and Martin'. *Occasional papers in systemic linguistics* 5. pp. 75-129.

Hudson, Richard A. (1984). *Word grammar.* Oxford: Blackwell Publishers.

Hudson, Richard A. (1994). 'About 37% of word-tokens are nouns'. *Language.* 70. pp. 331-339.

Isenberg, H. (1970). 'Der Begriff "Text" in der Sprachtheorie'. *ASG-Bericht* Nr. 8. Berlin. (Multiplicated.).

Johansson, Stig (1982). 'Computer corpora in English language research'. *ICAME News.*

Johns, Tim (1991). 'Should you be persuaded - two samples of data-driven learning materials', in Johns, Tim and P. King (eds.). pp. 1-16.

Johns, Tim and P. King (1991). (eds.). 'Classroom concordancing 4'. *English language research journal.* Birmingham: University of Birmingham.

Johnson, Roy (1996). *Writing Essays.* Manchester: Clifton Press (Quotes from its electronic version: *HelpDisk! 2.2, Essay-writing program*).

Jordan, Michael P. (1984). 'Complex lexical cohesion in the English clause and sentence', in Alan Manning, Pierre Martin and Kim McCalla (eds.). pp. 224-234.

Jordan, Michael P. (1992). 'An integrated three-pronged analysis of a fund-raising letter', in William C. Mann and S. A. Thompson (eds.). pp. 171-226.

Karlgren, H (1975). 'Text connexitivity and word frequency distribution', in H. Ringbom (ed.). pp. 335-348.

Katz, J. (1980). 'Chomsky on meaning'. *Language* 56. pp. 1-42.

Kehler, A. (1997). 'Current theories for centering for pronoun interpretation: a critical evaluation'. *Computational Linguistics.* 23 (3). pp. 467-475.

Kiefer, Ferenc (1969). (ed.). *Studies in syntax and semantics.* Dordrecht.

Koch, W. (1971). *Taxologie des Englischen.* Munich: Fink.

Kohonen, Viljo and Nils E. Enkvist (1978). (eds.). *Text linguistics, cognitive learning, and language teaching.* (Publications de l'Association Finlandaise de Linguisitque Appliquée, No. 22). Åbo: Åbo Akademi.

Kurzon, D. (1988). 'The Theme in text cohesion', in Yishai Tobin (ed.). pp. 155-162.

La Berge, D. and J. Samuels (1977). (eds.). *Basic processes in reading comprehension.* Hillsdale, N. J.: Erlbaum.

Langacker, Ronald W. (1996). 'Conceptual grouping and pronominal anaphora', in Barbara Fox (ed.). pp. 333-378.

Lau, Hieng Hiong (1992). *Nominalised packaging in scientific journal discourse.* PhD thesis. Birmingham: University of Birmingham.

Lautamatti, Liisa (1978). 'Observations on the development of the topic in simplified discourse', in Viljo Kohonen and Nils E. Enkvist (eds.).

Leech, Geoffrey N. (1997). 'Teaching and language corpora: a convergence', in Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.). pp. 1-23.

Leech, Geoffrey N. and Candlin, C. (1986). (eds.). *Computers in English language teaching and research.* London: Longman.

Lehnert, Wendy G. (1980). 'Question answering in natural language processing', in Leonard Bolc (ed.).

Lehnert, Wendy G. and Martin H. Ringle (1982). (eds.). *Strategies for natural language processing.* Hillsdale, N. J.: Erlbaum.

Lowe, I. (1988). 'Sentence initial elements in English and their discourse function'. *Occasional papers in systemic linguistics* 2. pp. 5-33.

Lyons, John (1970). (ed.). *New horizons in linguistics.* Harmondsworth: Penguin.

Maizell, Robert E., Smith, J. F. and Singer, T. E. R. (1971). *Abstracting scientific and technical literature.* London: John Wiley and Sons.

Mann, William C. and S. A. Thompson (1985). 'Assertions from discourse structure'. *Proceedings of the eleventh annual meeting of the Berkeley Linguistics Society* 11. pp. 245-258.

Mann, William C. and S. A. Thompson (1986). 'Relational propositions in discourse'. *Discourse processes* 9. pp. 57-90.

Mann, William C. and S. A. Thompson (1992). (eds.). *Discourse description: diverse linguistic analyses of a fund raising text.* Amsterdam: John Benjamins Publishing Company.

Manning, Alan, Pierre Martin and Kim McCalla (1984). (eds.). *The tenth LACUS forum (1983)., Univerité Laval.* Columbia, S.C.: Hornbeam Press.

Marcu, D. (1997). 'From discourse structures to text summaries'. Toronto: University of Toronto. (manuscript).

Martin, Jim R. (1992). 'Theme, method of development and existentiality: the price of a reply'. *Occasional papers in systemic linguistics.* 6. pp. 147-183.

Mathesius, Vilem (1939). 'Otak zvaném aktauálním členění věty' [The so-called information-bearing structure of the sentence]. *Slovo a Slovesnost.* 5. pp. 171-174 .

Matthiessen, Christian (1992). 'Interpreting the textual metafunction', in Martin Davies and Louise J. Ravelli (eds.).

Mauranen, Anna (1996). 'Discourse competence - evidence from Thematic development in narrative and non-narrative texts'. in Eija Ventola and Anna Mauranen (eds.). pp. 195-230.

McCarthy, Michael J. (1991). *Discourse analysis for language teachers.* Cambridge: Cambridge University Press.

McCarthy, Michael J. (1994). 'It, this and that' in Coulthard (ed.). pp.266-275.

McCarthy, Michael J. and Ronald A. Carter (1994). *Language as discourse: perspective for language teaching.* London: Longman.

McEney, A., I. Tanaka and S. P. Botley (1997). 'Corpus annotation and reference resolution'. *Proceedings of the ACL Workshop on operational factors in practical, robust anaphora resolution for unrestricted text.* pp.67-74.

McGregor, W. (1990). 'On the notion of Theme in semiotic grammar'. Paper presented at the fourth international conference of functional grammar, Copenhagen, 1990.

McKeown, G. et al. (1983). 'The effects of long-term vocabulary instruction on reading comprehension: a replication'. *Journal of reading behaviour* 15/1. pp. 3-18 .

Minugh, David (1997). 'All the language that's fit to print: using British and American newspaper CD-ROMs as corpora', in Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.). pp. 67 82.

Mitkov, Ruslan (1998). 'Robust pronoun resolution with limited knowledge' *Proceedings of the 18th International Conference on Computational Linguistics (COLING '98)/ACL '98 Conference.* Montreal, Canada.

Mohsen Ghadessy (1995). (ed.). *Thematic development in English texts.* London: Pinter.

Morris, Jane and Graeme Hirst (1991). 'Lexical cohesion computed by thesaural relations as an indicator of the structure of text'. *Computational linguistics.* 17/1. pp. 21-48.

Nation, P. and J. Coady (1988). 'Vocabulary and reading', in Ronald A. Carter and Michael McCarthy (eds.).

Ndahiro, Alfred (1998). *Theme, Rheme, Given and New in Written Discourse: Evidence from Annual Business Reports.* PhD Thesis. Liverpool: University of Liverpool.

Nwogu, K. and Thomas Bloor (1991). 'Thematic progression in professional and popular medical texts', in Eija Ventola (ed.). pp. 369-384.

Parsons, Gerald (1995). *Measuring coherence in English texts: the relationship between cohesion and coherence.* PhD thesis. Nottingham: University of Nottingham.

Parsons, Gerald (1996). 'The development of the concept of cohesive harmony', in Margaret Berry, Christopher Butler, Robin Fawcett and Guowen Huang (eds.). pp. 585-599.

Peng, Wangheng (1994). *Information distribution as reflected by patterns of lexis in text and their implications for teaching reading.* MA dissertation. Liverpool: University of Liverpool.

Peng, Wangheng (1996-1997). 'Evaluation reports' (12 unpublished reports)., for European Foundation and University of Liverpool joint project in automatic text summary. Liverpool: University of Liverpool.

Peng, Wangheng (1997a). 'Patterns of lexis and information distribution', in Hu Zhuanglin and Fang Yan (eds.). pp. 119-135.

Peng, Wangheng (1997b). 'Key words in Theme and Rheme', in *Language and discourse* 5. pp. 43-69.

Peng, Wangheng (1998a). "Keywords in Theme and what the text is about", paper delivered at the 25th International Systemic Functional Congress, Cardiff: Cardiff University. 13th-17th July 1998.

Peng, Wangheng (1998b). "Keywords in Theme and their implications for TEFL reading", in *Teaching and language corpora 98*, Oxford: Keble College, Oxford University. pp. 201-204.

Perren, George E. and John L. M. Trim (1971). (eds.). *Applications of linguistics*. Cambridge: Cambridge University Press.

Petöfi, J. S. and Emel Sözer (1983). (eds.). *Micro and macro connexity of texts.* Hamburg Helmut Baske.

Phillips, M. K. (1983). *Lexical macrostructure in science text.* PhD thesis. Birmingham: University of Birmingham.

Phillips, M. K. (1988). 'Text, terms and meanings: some principles of analysis', in James D. Benson, Michael J. Cummings and William S. Greaves (eds.). pp. 99-118.

Phillips, M. K. (1989). 'The lexical structure of text'. *Discourse analysis monographs No 12.* Birmingham: English Language Research, University of Birmingham.

Pike, Kenneth L and Evelyn G. Pike (1977). *Grammatical analysis.* Arlington. TX: Summer Institute of Linguistics

Pike, Kenneth L. (1967). *Language in relation to a unified theory of the structure of human behaviour.* The Hague: Mouton.

Pike, Kenneth L. (1972). 'Grammar as wave', in Ruth M. Brend (ed.). pp. 231-241.

Prince, Ellen F. (1992). 'The ZPG letter: subjects, definiteness, and information-status', in William C. Mann and S. A. Thompson (eds.). pp. 295-325.

Procter, Paul et. al. (1978). *Longman dictionary of contemporary English.* Oxford: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey N. Leech and Jan Svartvik (1985). *A comprehensive grammar of the English language.* London: Longman.

Ravelli, Louise J. (1991). *Language from a dynamic perspective: models in general and grammar in particular.* Ph.D. thesis. Birmingham: University of Birmingham.

Ravelli, Louise J. (1995). 'A dynamic perspective: implications for metafunctional interaction and an understanding of Theme', in Ruqaiya Hasan and Peter H. Fries (eds.). pp. 187-234.

Reinhart, Tanya (1982). *Pragmatics and linguistics: an analysis of sentence topics.* Bloomington: Indiana University Linguistics Club.

Renouf, Antoinette (1993). 'What the linguist has to say to the information scientist', in *Journal of document and text management.* Vol 1, No.2 1993. pp. 173-190.

Rie, E. D. and Yeh, J. W. (1983). 'The effect of language on abstracting skills in learning disabled children'. *Journal of clinical child psychology* 12/1. pp. 40-45.

Riesbeck, Christopher, K. (1982). 'Realistic language comprehension', in Wendy G. Lehnert and Martin H. Ringle (eds.).

Rinehart, S. D., S. A. Stahl and L. G. Erickson (1986). 'Some effects of summarization training on reading and studying'. *Reading research quarterly* 21/4. pp. 422-438.

Ringbom, Hakan (1975). (ed.). *Style and text: studies presented to Nils Erik Enkvist.* Stockholm: Sprkforlaget Scriptor.

Rumelhart, David E. (1977). 'Understanding and summarizating brief stories', in D. La Berge and J. Samuels (eds.).

Sager, Naomi (1981). *Natural language information processing: a computer grammar of English and its applications.* Reading Mass.: Addison-Wesley.

Sanna Kaisa Tanskanen and Brita Warvik (1994). (eds.). *Topics and comments: papers from the discourse project.* Turku: Univeristy of Turku.

Schank, Roger C. and Christopher K. Riesbeck (1981). *Inside computer understanding: five programs plus miniatures.* Hillsdale, N. J.: Erlbaum.

Scott, Mike (1996). *WordSmith Tools.* Oxford: Oxford University Press.

Scott, Mike (1997a). 'PC analysis of key words - and key-key words'. *System* 25/2: pp. 233-245.

Scott, Mike (1997b). 'The right word in the right place: keyword associates in two languages'. (draft version).

Scott, Nelia (forthcoming). 'Research normalisation in literary translation: a computer investigation', to appear in *Translation and literature.* Edinburgh: Edinburgh University Press.

Sgall, Petr (1969). 'L'ordre des mots et la sémantique', in Ferenc Kiefer (ed.). pp. 231-240.

Sherrard, C. (1986). 'Summary writing: a topographical study'. *Written communication.* 3/3. pp. 324-343.

Shimazumi, M. and A. P. Berber Sardinha (1996). 'Approaching the assessment of performance unit (APU). archive of schoolchildren's writing from the point of view of corpus linguistics'. Paper presented at the TALC96 Conference, Lancaster University, UK. 11 August 1996.

Simms, Karl (1997). (ed.). *Language and the subject* (Critical studies, Vol. 9). Amsterdam: Rodopi.

Sinclair, John M. (1966). 'Beginning the study of lexis', in Charles E. Bazell, John C. Catford, M. A. K. Halliday and Robert H. Robins (eds.). pp. 410-430.

Sinclair, John M. (1986). 'Basic computer processing of long texts', in Geoffrey N. Leech and C. Candlin (eds.). pp. 184-203.

Sinclair, John M (1987). (ed.). *Looking up: an account of the COBUILD project in lexical computing.* London: HarperCollins Publishers.

Sinclair, John M. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, John M. (1992). 'Trust the text', in Martin Davies & Louise J. Ravelli. (eds.). pp. 5-19.

Sinclair, John M. and Malcolm Coulthard (1975). *Towards an analysis of discourse: the English used by teachers and pupils.* Oxford: Oxford University Press.

Sinclair, John M., Michael P. Hoey and G. Fox (1993). (eds.). *Techniques of description: spoken and written discourse.* London: Routledge.

Skalička, V. (1960). Syntax of the enunciation. *SaS* 21. pp. 241-149.

Sparck Jones, K. (1993). 'What might be in a summary?' *Information retrieval 93: Von der modellierung zur anwendung.* Universitätsverlag Konstanz. pp. 9-26.

Speight, F. Y. (1967). (ed.). *Guide to sourceindexing and abstracting of the engineering literature.* New York: Engineers Joint Council.

Stainton, Caroline (1993). *Metadiscourse and the analytical text: a genre-based approach to children's written discourse.* PhD Thesis. Manchester: University of Manchester.

Steele, Ross and Terry Threadgold (1987). (eds.). *Language topics, essays in honour of Michael Halliday Vol. 1.* Amsterdam: John Benjamins Publishing Company.

Stubbs, Michael (1983). *Discourse analysis: the sociolinguistic analysis of natural language.* Oxford: Basil Blackwell.

Stubbs, Michael (1986). 'Lexical density: a technique and some findings', in Malcolm Coulthard (ed.). pp. 27-42.

Stubbs, Michael (1993). 'British traditions in text analysis - from Firth to Sinclair', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.). pp.1-33.

Stubbs, Michael (1995). 'Corpus evidence for norms of lexical collocation', in Guy Cook and Barbara Seidlhofer (eds.). pp. 245-56.

Stubbs, Michael (1996). *Text and corpus analysis: computer-assisted studies of language and culture.* Oxford: Blackwell Publishers.

Svoboda, A. (1981). *Diatheme.* Brno: Masaryk University.

Svoboda, A. (1983). 'Thematic elements'. *Brno studies in English* 15. pp. 49-85.

Tait, J. I. (1982). 'Automatic summarising of English texts'. *Technical report No 47.* Cambridge: University of Cambridge Computer Laboratory.

Taylor, Charles V. (1983). 'Structure and theme in printed school text'. *Text* 3/2. pp. 197-228.

Thompson, Geoff (1996). *Introducing functional grammar.* London: Arnold.

Thompson, Susan. E. (1994a). 'Frameworks and contexts: a genre-based approach to analysing lecture introductions'. *English for specific purposes* 13/2. pp. 171-186.

Thompson, Susan. E. (1994b). 'Aspects of cohesion in monologue'. *Applied linguistics* 15/1. pp. 58-75.

Thorndyke, P. W. (1977). 'Cognitive structures in comprehension and memory of narrative discourse'. *Cognitive Psychology* 9. pp. 77-110.

Tobin, Yishai (1988). (ed.). *The Prague school and its legacy in linguistics, literature, semiotics, folklore and the arts.* Amsterdam: John Benjamins Publishing Company.

Trost, Pavel (1962). 'Subject and predicate'. *Slavica Pragensia* 4. pp. 267-270.

Ulijn, J. M. and A. K. Pugh (1985). (eds.). *Reading for professional purposes.* Leuven, Belgium: ACCO.

Ure, Jean (1971). 'Lexical density and register differentiation', in George E. Perren and John L. M. Trim (eds.). pp. 443-452.

van Dijk, Teun A. (1980). *Macrostructures: an interdisciplinary study of global structures in discourse, interaction and cognition.* Hillsdale, NJ: Lawrence Erlbaum.

van Dijk, Teun A. (1983). 'Discourse analysis: its development and application to the structure of news'. *Journal of communication.* 33. pp20-43.

van Dijk, Teun A. (1988). *News as discourse.* Hillsdale, NJ: Lawrence Erlbaum.

van Dijk, Teun A. and Kintsch, W. (1983). *Strategies of discourse comprehensoin.* New York: Academic Press.

Vande Kopple, W. J. (1991). 'Themes, thematic progressions and some implications for understanding discourse'. *Written communication* 8. pp. 311-347.

Vande Kopple, W. J. (1994). 'Some characteristics and functions of grammatical subjects in scientific discourse'. *Written communication* 11/4. pp. 534-564.

Ventola, Eija (1991). (ed.). *Functional and systemic linguistics.* New York: Mouton de Gruyter.

Ventola, Eija and Anna Mauranen (1996). (eds.). *Academic writing: intercultural and textual issues.* Amsterdam: John Benjamins Publishing Company.

Warvik, B., S-K. Tanskanen and R. Hiltunen (1995). (eds.). *Organisation in discourse, proceedings from the Turku Conference 1995.* Anglicana Turkuensia.

Weil, B. H. (1970). 'Standards for writing abstracts'. *Journal of the American Society for Information Science* 21/5. pp. 351-357.

Weil, B. H., Zarember, I. and Owen, H. (1963). 'technical abstracting fundamentals. Part II. Writing principles and practices'. *Journal of chemical documentation* 3. pp. 125-132.

Weil, H. (1844). *De l'ordre des mots dans les langues anciennes comparées aux langues modernes.* Paris: Joubert.

Weil, H. (1887). *The order of words in the ancient languages compared with that of the modern languages.* Boston: Ginn.

West, P. (1960). *Minimum adequate vocabulary.* London: Longman.

Wichmann, Anne, Steven Fligelstone, Tony McEnery and Gerry Knowles (1997). (eds.). *Teaching and language corpora.* London: Longman.

Wikberg, Kay (1990). 'Topic, Theme, and hierarchical structure in procedural discourse', in Jan Aarts and Willem Meijs (eds.). pp. 231-254.

Wilkins, A. (1972). *Linguistics and language teaching.* London: Edward Arnold.

Winter, Eugene O. (1977). 'A clause-relational approach to English texts: a study of some predictive lexical items in written discourse'. *Instructional science* 6/1. pp. 1-91.

Winter, Eugene O. (1978). 'A look at the role of certain words in information structure', in Jones, K. P. and V. Horsnell (eds.). *Informatics* 3/1. pp. 85-97. London: ASLIB.

Winter, Eugene O. (1979). 'Replacement as a fundamental function of the sentence in context'. *Forum linguisticum* 4/2. pp. 95-133.

Winter, Eugene O. (1982). *Toward a contextual grammar of English: the clause and its place in the definition of sentence.* London: George Allen & Unwin.

Winter, Eugene O. (1994). 'Clause relations as information structure: two basic text structures in English', in Malcolm Coulthard (ed.). pp. 46-68.

Yang, Huizhong (1986). 'A new technique for identifying scientific/technical terms and describing science texts'. *Literary and linguistic computing* 1/2. pp. 93-103.

Zora, S. and Johns-Lewis, C. (1989). 'Lexical density in interview and conversation'. *York papers in linguistics* 14. pp. 89-100.

Appendix 1. Sample Text A

Notes:
1. This text is entitled 'The Invisible Influence of Planet X', taken from The Independent on Sunday, 18th February, 1990.
2. Sentences are numbered. Where a sentence contains more than one clause the clauses are numbered as {Xa}, {Xb}, {Xc}, etc., 'X' standing for the sentence number.
3. Themes are italicised; proforms and ellipses are lexicalised.

(1) *The textbooks* say there are nine planets in our solar system.

(2) *The most distant [planet]* is Pluto, discovered on 18 February 1930 by astronomers at the Lowell Observatory, Arizona.

(3) *But some astronomers* have continued to suspect there may be a tenth planet lurking even further away which has somehow escaped detection.

(4) *One [astronomer], Robert Harrington of the US Naval Observatory in Washington*, has begun a new search for 'Planet X'.

(5) ~~*He*~~ *[Harrington]* is using similar techniques to those [techniques] used by Clyde Tombaugh 60 years ago to discover Pluto.

(6) *The young astronomer [Tombaugh]* had detected a sure sign of an object orbiting the Sun by comparing two photographs showing {6a} *that a speck of light* had shifted position against the stars.

(7) *For a number of years before Mr Tombaugh's discovery, the existence of a ninth planet* was suspected, {7a} *because something large* was affecting the orbital path of Uranus around the Sun.

(8) *The locating of Pluto* was thought to explain the Uranus effect.

(9) *But Dr Harrington and other sceptics* say {9a} *Pluto* is too small to explain the orbits of the planets in the outer regions of the solar system, such as Uranus and Neptune.

(10) *Most astronomers nowadays* work on such exotic problems as the origin of our universe or the properties of black holes, {10a} *but Dr Harrington* is cast in a traditional mould.

(11) *As director of the astrometry section at the Naval Observatory,* ~~*he*~~ *[Harrington]* prefers the classical work of finding the positions of the stars and planets with the greatest accuracy.

(12) *Dr Harrington* is continuing a tradition of planet-searching which began thousands of years ago, {12a} *when ancient astronomers* discovered the five planets visible to the naked eye; {12b} *Jupiter, for instance*, is the brilliant

object high up in the southern sky, {12c} *and Venus* can be easily seen in the east at sunrise.

(13) *In 1781, the British astronomer, Sir William Herschel* found a new planet, Uranus, setting off feverish planet-hunting.

(14) *This [planet-hunting]* led to the discovery of asteroids, or minor planets in great profusion between Mars and Jupiter.

(15) *Meanwhile, mathematicians* had become involved in the great planet hunt.

(16) ~~*They*~~ *[mathematicians]* found {16a} ~~*they*~~ *[mathematicians]* could not match the orbital path of Uranus around the Sun to that [path] predicted from the laws of gravity; there must be another planet, or planets, pulling ~~it~~ [Uranus] off course.

(17) *Neptune, located in 1846*, appeared to offer an incomplete solution.

(18) *This century, Percival Lowell, an American astronomer, who achieved notoriety for suggesting Mars had canals and life*, urged a search for a further planet using wide-angle cameras.

(19) *For 20 years, astronomers at Lowell Observatory* searched the skies.

(20) *Mr Tombaugh finally* spotted ~~his~~ [Tombaugh's] moving speck of light after a year at the job.

(21) *However, Pluto* seemed surprisingly small and faint, {21a} *and astronomers almost immediately* suspected {21b} ~~*it*~~ *[Pluto]* was not massive enough to pull Uranus off-course.

(22) *Mr Tombaugh himself* had doubts.

(23) *Still a professional astronomer at 83,* ~~*he*~~ *[Tombaugh]* said last week: {23a} '~~*It*~~ *[Pluto]* was much fainter than we [Tombaugh] expected, {23b} *and* ~~*I*~~ *[Tombaugh]* carried on searching, just in case there was another one [planet], for 14 years, until May 1943.'

(24) *More recent evidence* confirms {24a} *that Pluto* is much too small to influence Uranus and Neptune's orbits.

(25) *In 1978, James Christy at the US Naval Observatory accidentally* found a moon, subsequently called Charon, orbiting Pluto.

(26) ~~*Its*~~ *[Charon's] motion* showed {26a} *that the mass of Pluto* is a thousand times too small to influence the giant planets.

(27) ~~*This*~~*[Pluto being small]* has become the strongest evidence for the mysterious tenth planet.

(28) *Dr David Dewhurst, of the Cambridge Institute of Astronomy* sees the current search as more promising.

(29) 'There are another 20 years of data for a start, {29a} *and that [data]* helps.

(30) *But more significant perhaps* are the great advances in computing.

(31) *The computer models used by the Jet Propulsion Laboratory in Pasadena* can, for instance, handle much more complex calculations.'

(32) *At JPL, where the tracks of space probes through our solar system are computed with phenomenal accuracy, theorists* find {32a} *that recent*

observations of Uranus and Neptune do not fit computer predictions using a nine-planet model.

(33) *The laboratory's observers* found {33a} *that Uranus* is drifting out of its predicted orbit by 1,000 miles a year.

(34) '*One possible explanation* is an unseen planet,' {34a} *Dr Harrington* says.

(35) *Nevertheless, Mr Tombaugh* remains sceptical.

(36) '~~*I*~~ *[Tombaugh]* did my searching very thoroughly and very slowly.

(37) *If* ~~*it*~~ *[planet X]'s there,* ~~*it*~~ *[planet X]* should have shown on my plates.

(38) *However,* ~~*I*~~ *[Tombaugh]* only covered two-thirds of the sky {38a} *and the weakest part of my search* was in the south.

(39) ~~*I*~~ *[Tombaugh]* think {39a} *the case for Planet X* is marginal; {39b} *maybe* ~~*it*~~ *[Planet X]*'s there, {39c} *and maybe* ~~*it*~~ *[Planet X]* isn't [there].

(40) *Let's* see.'

(41) *Dr Harrington* has now begun work on two fronts, running new computer calculations in Washington and making fresh observations in New Zealand.

(42) '~~*My*~~ *[Harrington's] computer strategy* is to make model solar systems that include the nine known planets plus a guess at Planet X.

(43) ~~*I*~~ *[Harrington]* then run lots of these 10-planet simulations to give the smallest possible deviation of Uranus and Neptune from their observed positions.

(44) *Each time* ~~*we*~~ *[Harrington] do this* ~~*we*~~ *[Harrington]* predict a position for Planet X in 1990.

(45) *What* ~~*we*~~ *[Harrington]* are finding is {45a} *that the permitted positions for Planet X* cluster in a small region of the sky.'

(46) *The inclusion of the irregularities in Neptune's orbit* is new,{46a} *and that [inclusion]* could be {46b} *why computer models* are showing a narrower search area.

(47) *Neptune's true position* is accurately known, following the Voyager 2 encounter in August 1989.

(48) *Dr Harrington* says {48a} *the most remarkable feature predicted for Planet X* is {48b} *that* ~~*its*~~ *[Planet X's] orbit* is tilted 300 degrees away from the ecliptic, the main plane of the solar system, {48c} *where all previous searches* have concentrated.

(49) ~~*His*~~ *[Harrington's] models also* predict a greater distance from the Sun, about 10 billion miles, or between two and three times as distant as Pluto.

(50) *In April the new sweep* starts in earnest at the Black Birch Observatory in New Zealand.

(51) *A modest 8in telescope, similar to that [telescope] used by Mr Tombaugh,* will examine the northern part of the constellation Centaurs.

(52) *Pairs of photographs of the same region of sky taken on successive nights* will be sent to Washington.

(53) *Using a blink comparator, a device that compares two photographs, Dr Harrington* hopes to locate any faint object that has moved during the interval between the two pictures.

(54) *A serious problem* is {54a} *that the search area* falls close to the Milky Way, {54b} *and every plate* will include millions of faint stars in our galaxy.

(55) *The planet, if it exists*, must be picked out from this crowded background.

(56) *Dr Harrington* says {56a} *astronomers still* do not understand the outer regions of our solar system.

(57) ~~*He*~~ *[Harrington]* hopes {57a} *Planet X* will help explain the mysterious 'wobble' of Uranus and Neptune.

(58) '~~*I*~~ *[Harrington]* think {58a} ~~*we*~~ *[Harrington]* have a 50-50 chance of showing {58b} *that the anomalies* are due to another planet orbiting 10 billion miles from the Sun.'

Appendix 2. Lexical items in Sample Text A

(frequency >= 2)

Lexical items	*All*	*Th*	*Rh*
Accuracy/ately	3	1	2
Area	2	1	1
Astrometry/nomer /s	14	13	1
Began/un	3	0	3
Billion	2	0	2
Black	2	0	2
Calculations	2	0	2
Case	2	1	1
Charon/'s	2	1	1
Compares/ing /ator	3	2	1
Computed ing er	7	4	3
Continued ing	2	0	2
Course	2	0	2
Detected ion	2	0	2
Discover ed y	5	1	4
Distance t	3	1	2
Evidence	2	1	1
Exists/ence	2	2	0
Explain anation	4	1	3
Faint/er	4	0	4
Find ing found	7	1	6
Great/er est	5	0	5
Harrington	20	20	0
Help/s	2	0	2
Hope s	2	0	2
Hunt/ing	3	1	2
Included ion	4	2	2
Influence	2	0	2
Jupiter	2	1	1
Known	2	0	2
Laboratory	2	2	0
Light	2	1	1
Locate/ing	3	2	1
Lowell	3	2	1
Make/ing	2	0	2
Mars	2	1	1
Mass/ive	2	1	1
Mathematicians	3	3	0
Miles	3	0	3

Lexical items	*All*	*Th*	*Rh*
Model/s	5	3	2
Moved/ing	2	0	2
Mysterious	2	0	2
Naval	3	3	0
Neptune/'s	8	4	4
New	7	1	6
Nine/th	4	1	3
Object	3	0	3
Observatory	6	4	2
Observed/ers /ations	4	2	2
One	3	2	1
Orbit/s/ing/al	10	2	8
Path	3	0	3
Photographs	3	2	1
Planet/s	36	12	24
Plate/s	2	1	1
Pluto	11	7	4
Position/s	6	2	4
Possible	2	1	1
Predict/ed/ions	6	1	5
Problem/s	2	1	1
Pull/ing	2	0	2
Recent	2	2	0
Region/s	4	1	3
Run/ning	2	0	2
Say/s/aid	6	0	6
Sceptics/al	2	1	1
Search/es/ed/ing	11	3	8
See/s/n	3	0	3
Showed/n/ing	5	0	5
Similar	2	1	1
Sky/ies	5	1	4
Small/er	6	0	6
Solar	6	1	5
South/ern	2	0	2
Speck	2	1	1
Stars	3	0	3
Starts	2	0	2
Sun	5	0	5
Suspect/ed	3	0	3

Lexical items	*All*	*Th*	*Rh*
System/s	6	1	5
Tenth/X	9	4	5
Think/thought	3	0	3
Thousands	2	0	2
Time/s	4	1	3
Tombaugh/'s	14	11	3
Tradition/al	2	0	2
Two	6	1	5
Uranus	12	2	10
Us	2	2	0
Used/ing	7	3	4
Washington	3	1	2
Work	3	0	3
Year/s	8	2	6
Zealand	2	0	2
Total	***423***	***154***	***269***

Appendix 3. Lexical items in Theme and Rheme of Sample Text A

Notes: 1. All figures in the table are sentence numbers;
2. △ = occurring twice in the same Theme or Rheme area of the sentence;
3. ( ) = co-occurring in the Theme and Rheme of the same sentence.

Lexical items	*Theme*	*Rheme*
Accuracy/ately	32	11, 47
Area	54	46
Astrometry nomer s	3, 4, 6, 10, 11, 12, 13, 18, 19, 21, 23, 28, 56	2
Began un		4, 12, 41
Billion		49, 58
Black		10, 50
Calculations		31, 41
Case	39	23
Charon/'s	26	25
Compares ing ator	53△	6
Computed ing er	31, 32, 42, 46	30, 32, 41
Continued ing		3, 12
Course		16, 21
Detected ion		3, 6
Discover ed y	7	2, 5, 12, 14
Distance t	2	49△
Evidence	24,	27
Exists ence	7, 55	
Explain anation	34	8, 9, 57
Faint/er		21, 23, 53, 54
Find/ing found	45	11, 13, 16, 25, 32, 33
Great/er/est		11, 14, 15, 30, 49
Harrington	4, 5, 9, 10, 11, 12, 34, 41, 42, 43, 44△, 45, 48, 49, 53, 56, 57, 58△,	
Help/s		29, 57
Hope/s		53, 57
Hunt/ing	14	13, 15
Included/ion	46△	42, 54
Influence		24, 26
Jupiter	12	14

Lexical items	*Theme*	*Rheme*
Known		42, 47
Laboratory	31, 33	
Light	6	20
Locate/ing	8, 17	53
Lowell	18, 19	2
Make/ing		41, 42
Mars	18	14
Mass/ive	26	21
Mathematicians	15, 16△	
Miles		33, 49, 58
Model/s	31, 46, 49	32, 42
Moved/ing		20, 53
Mysterious		27, 57
Naval	4, 11, 25	
Neptune/'s	17, 32, 46, 47	9, 24, 43, 57
New	50	4, 13, 41△, 46, 50
Nine/th	7	1, 32, 42
Object		6, 12, 53
Observatory	4, 11, 19, 25,	2, 50
Observed ers ations	32, 33	41, 43
One	4, 34	23
Orbit/s/ing al	46, 48	6, 7, 9, 16, 24, 25, 33, 58
Path		7, 16△
Photographs	52, 53	6
Planet/s	2, 7, 14, 37△, 39△, 45, 48△, 55, 57	1, 3, 4, 9, 11, 12△, 13△, (14), 15, 16△, 18, 23, 26, 27, 32, 34, 42△, 43, 44, 58
Plate s	54	37
Pluto	8, 9, 21△, 23, 24, 26	2, 5, 25, 49
Position s	45, 47	6, 11, 43, 44
Possible	34	43
Predict/ed ions	48	16, 32, 33, 44, 49
Problem/s	54	10
Pull ing		16, 21
Recent	24, 32	
Region s	52	9, 45, 56
Run ning		41, 43
Say s aid		1, 9, 23, 34, 48, 56
Sceptics al	9	35
Search es ed ing	38, 48, 54	4, 12, 18, 19, 23, 28, 36, 46
See s n		12, 28, 40
Showed/n ing		6, 26, 37, 46, 58
Similar	51	5
Sky ies	52	12, 19, 38, 45
Small er		9, 21, 24, 26, 43, 45
Solar	32	1, 9, 42, 48, 56
South ern		12, 38
Speck	6	20
Stars		6, 11, 54
Starts		29, 50
Sun		6, 7, 16, 49, 58
Suspect/ed		3, 7, 21
System/s	32	1, 9, 42, 48, 56
Tenth/X	39, 45, 48, 57	3, 4, 27, 42, 44
Think/thought		8, 39, 58
Thousands		12, 26
Time/s	44	26, 49

Lexical items	*Theme*	*Rheme*
Tombaugh/'s	6, 7, 20, 22, (23△), 35, 36, 38, 39, 51	5, (20), 23
Tradition/al		10, 12
Two	53	6, 38, 41, 49, (53)
Uranus	32, 33	7, 8, 9, 13, 16△, 21, 24, 43, 57
Us	4, 25	
Used/ing	31, 51, 53,	5△, 18, 32
Washington	4	41, 52
Work		10, 11, 41
Year/s	7, 19	5, 12, 20, 23, 29, 33
Zealand		41, 50

Appendix 4. Different types of links in the sentences of Sample Text A

Notes: (1) Except those in the square brackets [ ], all figures in the table are sentence numbers;
(2) The figures in the square brackets [ ] following a sentence number indicate the frequency of the particular type of link in that sentence.

S	A (T-T)	B (R-T)	C (R-R)	D (T-R)	E (A+C)	F (B+D)	G (A+D)	H (B+C)	I (A+B)	J (C+D)	Others
1		2, 7, 14, 37, 39, 45, 55, 57	3, 4, 9[×4], 11, 12, 13, 15, 16, 18, 23[×2] 26, 27, 34[×2], 42[×3], 43, 44, 48, 56[×3], 58					32[2+1], 48[1+3]			
2	37, 39, 45, 48, 55, 57	6, 8, 10, 19[×2], 21[×2], 24, 28, 56	50	15, 16, 27, 32, 34, 42, 43, 44, 58	14[×2]	3[×2], 9[×2], 13[×2], 18[×2], 26[×2], 4[2+1], 11[2+1], 23[2+1]		25[×2]	7[×2]	49[×2]	12[b+c+d]
3	10, 28, 56	39[×2], 45[×2], 48[×2], 57[×2], 14, 37, 55	27[×2], 42[2], 44[×2], 4, 9, 15, 16, 26, 32, 34, 43, 58		4[×2], 6 [×2],11[×2], 13[×2], 21[×2], 23[×2], 12[1+2]			7[×2]	19[×2]		18[a+b+c]
4	10[×2], 19[×2], 56[×2], 5, 6, 18, 21, 25[×3], 49, 53	39[×2], 7, 14, 37, 38, 54, 55	13[×2], 18[×2], 27[×2], 46[×2], 15, 16, 19, 26, 32, 36	52	9[×2], 28[×2], 43[×2], 58[×2], 11[4+1], 12[2+2], 13[1+2], 34[2+1], 42[1+2], 44[1+2]	50[×2]			45[1+2], 48[1+3], 57[1+2]		23[a+2c+d], 41[a+2c+d]
5	10, 11, 34, 41, 42, 43, 44, 45, 48, 56, 57, 58	6, 7[×3], 8, 19, 21, 22, 24, 26, 31, 35, 36, 38, 39, 51[×3]	14, 18, 25, 29, 32, 33		49[×2], 12[1+2]			20[×2], 23[1+2]	9[×2], 53[×2]		
6	10, 13, 18, 19, 21, 22, 28, 35, 36, 39, 51, 56	45, 47, 48, 52	16[×2], 49[×2], 9, 24, 25, 26, 33, 37, 41, 43, 44, 54, 58[×3]		38[×2], 12[×2], 11[1+2], 7[1+2]		23[×2], 20[1+2]	46[×2], 53[3+1]			

S	*A (T-T)*	*B (R-T)*	*C (R-R)*	*D (T-R)*	*E (−A+C)*	*F (−B+D)*	*G (A+D)*	*H (B+C)*	*I (A+B)*	*J (C+D)*	*Others*
7	39[×2], 55[×2], 19, 22, 35, 36, 37, 38, 45, 51	46	21[×2], 24[×2], 8, 25, 49	11, 12[×3], 15, 18, 23[×3], 26, 27, 29, 34, 42, 44	57[×2]	32[×2]	14[×2], 20[×2]		48[×2]	13[×2], 43[×2], 9[2+1], 16[4+1], 58[2+1]	33[b+c+d]
8	17, 23, 26	32, 33, 34	57[×2], 13, 16, 39, 43, 58	25, 49, 53	21[×2], 24[×2], 9[1+2]						
9	10, 41, 48, 53	46[×2], 14, 17, 33, 37, 39, 45, 47, 52, 55	13[×2], 15, 16[×3], 18, 27, 33, 56[×3]	35	11[×2], 12[×2], 44[×2], 56[×2], 21[1+2], 23[1+2], 24[1+4], 26[1+2], 42[1+3], 43[1+4], 45[1+2], 58[1+2]		49[×2]	32[4+1], 48[2+3]		25[×2]	34[a+b+2c], 57[a+b+3c]
10	56[×2], 13, 18, 19, 21, 23, 28, 34, 42, 43, 44, 45, 48, 49, 53, 57, 58	54	50		41[×2], 11[2+1], 12[2+1]						
11	19[×2], 56[×2], 21, 28, 53	37, 39, 55	15[×2], 16[×2], 26, 27, 30, 33, 54	50	18[×2], 23[×2], 34[×2], 41[×2], 42[×2], 49[×2], 58[×2], 12[2+1], 13[1+2], 25[2+1], 43[1+2], 44[1+2]			14[×2], 47[×2], 32[1+2]	48[×2], 57[×2], 45[1+3]		

S	A (T-T)	B (R-T)	C (R-R)	D (T-R)	E (A+C)	F (−B+D)	G (A+D)	H (B+C)	I (A+B)	J (C+D)	Others
12	56[×2], 21, 49	37, 39, 52, 54, 55	26[×2], 15, 16, 20, 27, 29, 32, 33, 36, 40, 46		13[×2], 34[×2], 41[×2], 42[×2], 43[×2], 44[×2], 53[×2], 58[×2], 18[1+2], 23[1+3], 28[1+2]			38[1+2]	57[×2], 48[1+2]		14[b+c+d], 19[a+b+2c], 45[a+b+c]
13	19, 28, 56	14[×2], 45[×2], 37, 39, 48, 50, 55	15[×2], 43[×2], 16[×3], 24, 25, 26, 27, 34, 41, 42, 44, 46, 58		18[×2], 21[×2], 23[×2]			33[×2], 57[×2], 32[1+2]			
14	37, 39, 45, 48, 55, 57		30, 49	16, 23, 26, 27, 32, 34, 42, 43, 44, 58		18[×2]				15[1+2]	
15		37, 39, 45, 48, 55, 57	18, 23, 26, 27, 30, 32, 34, 42, 43, 44, 49, 58		16[×2]						
16		45[×2], 32, 33, 37, 39, 46, 48[×3], 55	24[×2], 25[×2], 43[×2], 44[×2], 49[×2], 18, 21[×3], 23, 26, 27, 32[×3], 33[×3], 34, 42, 58[×3]					57[×2]			
17	32, 46, 47			24, 43, 53, 57							
18	21, 56	48[×2], 31, 37, 38, 39, 45, 51, 53, 54, 55, 57	32[×2], 26, 27, 34, 36, 42, 43, 44, 46, 58		28[×2], 19[1+2], 23[1+2]						
19	21, 25, 56	48, 52, 54	36, 46, 45	20, 29, 33, 50	28[×2]			38[×2]			23[a+c+d]
20	22, 35, 36, 38, 39, 51		29, 33, 53							23[×2]	

S	*A (T-T)*	*B (R-T)*	*C (R-R)*	*D (T-R)*	*E (A+C)*	*F (−B+D)*	*G (A+D)*	*H (B+C)*	*I (A+B)*	*J (C+D)*	*Others*
21	28, 56	32, 33	43[×2], 45, 53, 54, 57	25, 49	23[2+1], 24[1+2]						26[a+b+c]
22	35, 36, 38, 39, 51			23							
23	24	38[×2], 35, 37, 39[×3], 45, 51, 55, 57	27, 29, 33, 42, 43, 44, 46, 53, 58	25, 49	26[×2], 28[×2], 56[×2]			36[×2], 54[×2], 34[1+2], 48[2+1]			
24		46[×2], 47, 48	57[×2], 43[×3], 45, 58	27, 49	26[1+2]			33[×2]		25[×2]	32[a+2b+c]
25		26[×2], 45, 46, 48	33[×2], 32, 49, 58	50							
26		39, 48, 55, 57	43[×2], 58[×2], 27, 32, 34, 42, 46					37[×2], 44[×2], 45[×2]		49[×2]	
27		39[×2], 45[×2], 48[×2], 37, 55	42[×2], 44[×2], 32, 34, 43, 58					57[2+1]			
28	56	38, 48, 54	36, 40, 46								
29			33, 50, 57								
30		31, 42, 46	32, 41, 49								
31	46[×2], 33, 49, 51, 53			32[×3]			42[×2]			41[×2]	
32		45[×2], 37, 39, 51, 53, 55	44[×2], 34, 58	56[×2],	33[2+2]	48[2+2], 57[1+2]	47[×2]	49[×2]	46[1+2]	41[×2], 43[1+2]	42[b+c+2d]
33		48[×2], 45, 46	49[×2], 58[×2], 44	43[×2], 41, 57							
34	41, 49, 53	37, 39, 55			42[×2], 44[×2], 56[×2], 58[×2]				45[×2]		43[a+c+d], 48[a+b+c], 57[a+b+d]
35	36, 38, 39, 51										
36	39, 51	48, 54	46						38[×2]		

S	*A (T-T)*	*B (R-T)*	*C (R-R)*	*D (T-R)*	*E (A+C)*	*F (−B+D)*	*G (A+D)*	*H (B+C)*	*I (A+B)*	*J (C+D)*	*Others*
37	39, 45, 48, 55, 57	54	46	42, 43, 44						58[×2]	
38	48[×2], 39, 51, 54, 55, 57	52, 53	41, 49	46	45[×2]						
39	45, 48, 51, 57			42[×2], 44[×2], 43						58[×2]	
40											
41	44, 45, 48, 56, 57, 58		52		42[×2], 49[×2], 43[1+2]			46[×2], 50[×2]	53[×2]		
42	53	55	47, 54		43[×2], 58[×2], 44[1+2], 56[1+2]				49[×2], 45[1+2], 46[1+2], 57[1+2]		48[a+2b+2c]
43	46, 49, 53, 56	55	45		58[×2], 44[1+2]				47[×2], 48[×2], 45[1+2]		57[a+b+2c]
44	49, 53, 56	47, 55	49		58[×2]				45[1+3], 48[1+3], 57[1+2]		
45	47, 48[×3], 49, 53, 55, 57[×3]	52[×2]			56[×2]		58[×2]				
46	47, 49	50		57	58[×2]	54[2+1]			48[×2]		
47				57							
48	53, 54, 55, 57[×3]				56[1+3]		49[×2], 58[2+1]				
49	56, 57				58[1+3]				53[×2]		
50											
51	53										
52	53			56							
53	56, 58		54		57[×2]						

S	*A (T-T)*	*B (R-T)*	*C (R-R)*	*D (T-R)*	*E (−A+C)*	*F (=B+D)*	*G (A+D)*	*H (B+C)*	*I (A+B)*	*J (C+D)*	*Others*
54											
55	57			58							
56	57, 58										
57							58[×2]				
58											

Appendix 5. Number of links in the sentences of Sample Text A

Ss	A (T-T)	B (R-T)	C (R-R)	D (T-R)	E	F	G	H	I	J	O	Total
1		8	28					7				43
2	6	10	1	9	2	19		2	2	2	3	56
3	3	11	15		15			2	2		3	51
4	15	8	14	1	29	2			10		8	87
5	12	18	6		5			5	4			50
6	12	4	17		10		5	6				54
7	12	1	7	15	2	2	4		2	15	3	63
8	3	3	7	3	7							23
9	4	11	12	1	37		2	10		2	9	88
10	18	1	1		8							28
11	7	3	9	1	29			7	8			64
12	4	5	12		26			3	5		10	65
13	3	9	17		6			7				42
14	6		2	10		2				3		23
15		6	12		2							20
16		11	28					2				41
17	3			4								7
18	2	12	11		8							33
19	3	3	3	4	2			2			3	20
20	6		3							2		11
21	2	2	6	2	6						3	21
22	5			1								6
23	1	11	9	2	6			10				39
24		4	7	2	3			2		2	4	24
25		5	5	1								11
26		4	9					6		2		21
27		8	8					3				19
28	1	3	3									7
29			3									3
30		3	3									6
31	6			3			2			2		13
32		7	4	2	4	7	2	2	3	5	4	40
33		4	5	4								13
34	3	3			8				2		9	25
35	4											4
36	2	2	1						2			7
37	5	1	1	3						2		12
38	7	2	2	1	2							14
39	4			5						2		11
40												0
41	6		1		7			4	2			20
42	1	1	2		10				11		5	30
43	4	1	1		5				7		4	22

Ss	*A (T-T)*	*B (R-T)*	*C (R-R)*	*D (T-R)*	*E*	*F*	*G*	*H*	*I*	*J*	*O*	*Total*
44	3	2	1		2				11			*19*
45	10	2			2		2					*16*
46	2	1		1	2	3			2			*11*
47				1								*1*
48	6				4		5					*15*
49	2				4				2			*8*
50												*0*
51	1											*1*
52	1			1								*2*
53	2		1		2							*5*
54												*0*
55	1			1								*2*
56	2											*2*
57							2					*2*
58												*0*
Total	*200*	*190*	*277*	*78*	*255*	*35*	*24*	*80*	*75*	*39*	*68*	*1,321*

Appendix 6. Questionnaire A

QUESTIONNAIRE

AELSU, Liverpool University, August, 1996 (A)

People's intuition about coherence may be varied. For example, the following utterances

A: Here's the phone.
B: I'm in the bath.

may be felt by some people to be more coherent than

A: Here's the phone.
B: I like strong tea.

Now please decide which of the following pairs of sentences is: a) very coherent; b) slightly coherent; c) slightly incoherent; or d) very incoherent, by ticking (✓) the appropriate box ☐.

For each pair, please tick ONE box only.

Please try to give reasons for your choice on the line following the sign →.

1 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) Most astronomers nowadays work on such exotic problems as the origin of our universe or the properties of black holes, but Dr Harrington is cast in a traditional mould.

b) Dr Harrington says astronomers still do not understand the outer regions of our solar system.

PTO→

2 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) But after a year in which the house-buying public has failed to conform to any statistical expectations, the feel-good factor among analysts is also pretty low.
b) Yolande Barnes estimates that 3.6 million people have put off buying over the past six years.

3 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) From the cradle to the grave, your family background was always with you, whether as a help or as a hindrance.
b) Although there is no hard and fast rule, your family background is important mainly insofar as it either inspires or dampens class ambition.

4 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) Pressure on John Major to come off the fence and declare whether he intends to enter a single European currency mounted yesterday when an all-party committee of MPs said there were serious drawbacks in his "wait and see" policy.
b) Mr Redwood will clearly be a key figure in the campaign over the single currency.

5 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) If Grade 3 departments are unlikely to lose all their research funding, they could well lose a lot of it.
b) As research funding means they can afford more staff, better laboratories and libraries, that is hardly surprising.

PTO↠

6 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) Power stations wanting to burn the carcasses would have to seek changes in regulations covering the way they operate.
b) Viability tests are underway at power stations in Didcot, Oxfordshire, and Ralcliffe on Soar, Nottinghamshire.

7 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) The warning call came from a public telephone outside a hotel adjoining the park.
b) Atlanta officials came under pressure to explain why the warning call made to an operator at 1.07am failed to reach the park before the device exploded at 1.25am.

8 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) The body clock is housed in a section of the brain called the hypothalamus, which is close to the optic nerves that transmit light from the eyes to the brain.
b) Doctors are learning more and more about the body clock.

Thank you for your time and attention

Appendix 7. Questionnaire B

QUESTIONNAIRE

AELSU, Liverpool University, August, 1996 (B)

People's intuition about coherence may be varied. For example, the following utterances

A: Here's the phone.
B: I'm in the bath.

may be felt by some people to be more coherent than

A: Here's the phone.
B: I like strong tea.

Now please decide which of the following pairs of sentences is: a) very coherent; b) slightly coherent; c) slightly incoherent; or d) very incoherent, by ticking (✓) the appropriate box ☐.

For each pair, please tick ONE box only.

Please try to give reasons for your choice on the line following the sign → .

1 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) Dr Harrington and other sceptics say Pluto is too small to explain the orbits of the planets in the outer regions of the solar system, such as Uranus and Neptune.

(b) In 1781, the British astronomer, Sir William Herschel found a new planet, Uranus, setting off feverish planet-hunting.

PTO↦

2 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

a) From the cradle to the grave, your family background was always with you, whether as a help or as a hindrance.
b) Although there is no hard and fast rule, your family background is important mainly insofar as it either inspires or dampens class ambition.

3 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) Merger rebel seeks place on Halifax board: Andrew Bibby looks at moves afoot to prevent a marriage with the Leeds society.
(b) The election would take place at the society's annual meeting in Halifax on 22 May.

4 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) After months of bitter arguments, Zoe, 26, and John, 27, decided they needed a break from each other.
(b) Yet according to Relate spokeswoman Zelda West-Meads, Zoe and John's experience is unusual.

5 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) He clearly recognised the symbiotic relationship between house prices and the fell-good factor.
(b) The arrival of the skips coincided with house price rises of 5-10 per cent.

PTO→

6 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) The warning call came from a public telephone outside a hotel adjoining the park.
(b) Atlanta officials came under pressure to explain why the warning call made to an operator at 1.07am failed to reach the park before the device exploded at 1.25am.

7 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) They are usually carried out on women who are given fertility drugs or test-tube treatments that produce multiple pregnancies.
(b) She denied that this meant there was abortion on demand of multiple pregnancies.

8 ☐ very coherent; ☐ slightly coherent; ☐ slightly incoherent; ☐ incoherent
→

(a) If Grade 3 departments are unlikely to lose all their research funding, they could well lose a lot of it.
(b) As research funding means they can afford more staff, better laboratories and libraries, that is hardly surprising.

Thank you for your time and attention

Appendix 8. Analysis of Answers to the Questionnaires

1) Results of the original answers

a) Questionnaire A:

Sentence pairs	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Very coherent.	4		2		1	2	3	
Slightly coherent	1	1	2	2	2	2	1	1
Slightly incoherent			1	2				1
Incoherent		4		1	2	1	1	3

b) Questionnaire B:

Sentence pairs	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Very coherent.	2		1	3		3	4	2
Slightly coherent			4	2	2	2	1	
Slightly incoherent	3	2			3			2
Incoherent		3						

2) Summary by Questionnaires

	Questionnaire A		*Questionnaire B*	
	Type A	*Type D*	*Type A*	*Type D*
Very coherent.	10	2	8	7
Slightly coherent	6	6	4	7
Slightly incoherent	3	3	2	8
Incoherent	1	9	1	3

3) Summary by Link Types

	Type A	Type D
Very coherent.	18	9
Slightly coherent	10	13
Slightly incoherent	5	11
Incoherent	2	12
Total	35	45

Appendix 9. Word lists of Group B Sample Texts

(Frequency >=2)

Sample Text B1: 950319

Words	*Th*	*Rh*	*Overall*
Intelligence	**7**	**9**	**16**
Apes	**5**	**5**	**10**
Gorillas	**4**	**3**	**7**
Ape	**4**	**2**	**6**
Human	**4**	**2**	**6**
Orangutans	**4**	**2**	**6**
Social	**3**	**6**	**9**
Chimps	**3**	**4**	**7**
Ancestors	**3**	**2**	**5**
Primates	**3**	**2**	**5**
Orangutan	**3**	**1**	**4**
Groups	**2**	**4**	**6**
Species	**2**	**4**	**6**
Humans	**2**	**3**	**5**
Live	**2**	**3**	**5**
Primate	**2**	**3**	**5**
Great	**2**	**2**	**4**
Living	**2**	**2**	**4**
Monkeys	**2**	**2**	**4**
Sheep	**2**	**2**	**4**
Borneo	**2**	**1**	**3**
Brain	**2**	**1**	**3**
Says	1	5	6
Time	1	5	6
Behaviour	1	4	5
Animals	1	3	4
See	1	3	4
Evolution	1	2	3
Food	1	2	3
Forest	1	2	3
Imitation	1	2	3
Mental	1	2	3
Particular	1	2	3
Things	1	2	3
Animal	1	1	2
Became	1	1	2
Behaved	1	1	2
Closest	1	1	2
Concept	1	1	2
Evidently	1	1	2

Words	*Th*	*Rh*	*Overall*
Get	1	1	2
Gorilla	1	1	2
Larger	1	1	2
Man	1	1	2
Many	1	1	2
Mirror	1	1	2
Past	1	1	2
Perspective	1	1	2
Pygmy	1	1	2
Question	1	1	2
Rudiments	1	1	2
Shepherds	1	1	2
Supinah	1	1	2
Tools	1	1	2
Trees	1	1	2
Years	1	1	2
Byrne	***9***	***0***	***9***
Department	***3***	***0***	***3***
University	***3***	***0***	***3***
Adult	***2***	***0***	***2***
Birds	***2***	***0***	***2***
Chimpanzees	***2***	***0***	***2***
Easily	***2***	***0***	***2***
Interaction	***2***	***0***	***2***
Neocortex	***2***	***0***	***2***
Psychology	***2***	***0***	***2***
Researcher	***2***	***0***	***2***
Russon	***2***	***0***	***2***
Scientists	***2***	***0***	***2***
Thinking	***2***	***0***	***2***
Traditionally	***2***	***0***	***2***
Complex	*0*	*6*	*6*
Understanding	*0*	*5*	*5*
Believes	*0*	*4*	*4*
Requires	*0*	*4*	*4*
Way	*0*	*4*	*4*
Evolved	*0*	*3*	*3*
Range	*0*	*3*	*3*
World	*0*	*3*	*3*
Bad	*0*	*2*	*2*
Believe	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Clues	*0*	*2*	*2*
Copy	*0*	*2*	*2*
Deception	*0*	*2*	*2*
Different	*0*	*2*	*2*
Difficult	*0*	*2*	*2*
Earlier	*0*	*2*	*2*
Evolve	*0*	*2*	*2*
Fire	*0*	*2*	*2*
Fossils	*0*	*2*	*2*
Go	*0*	*2*	*2*
Good	*0*	*2*	*2*
Highly	*0*	*2*	*2*
Intelligent	*0*	*2*	*2*
Interact	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Key	*0*	*2*	*2*
Limited	*0*	*2*	*2*
Order	*0*	*2*	*2*
Processing	*0*	*2*	*2*
Put	*0*	*2*	*2*
React	*0*	*2*	*2*
Sequence	*0*	*2*	*2*
Speculation	*0*	*2*	*2*
Stakes	*0*	*2*	*2*
Start	*0*	*2*	*2*
Take	*0*	*2*	*2*
Tree	*0*	*2*	*2*
Uses	*0*	*2*	*2*
Work	*0*	*2*	*2*

Sample Text B2: 950514

Words	*Th*	*Rh*	*Overall*
Molecules	**8**	**6**	**14**
Davies	**8**	**4**	**12**
Oxford	**7**	**5**	**12**
Company	**5**	**3**	**8**
Enzymes	**4**	**3**	**7**
Drug	**4**	**1**	**5**
Nature	**4**	**1**	**5**
Chiral	**3**	**5**	**8**
Davies's	**3**	**2**	**5**
Handedness	**3**	**2**	**5**
University	**3**	**2**	**5**
Asymmetry	**3**	**1**	**4**
Year	**3**	**1**	**4**
Left	**2**	**11**	**13**
Handed	**2**	**8**	**10**
Right	**2**	**7**	**9**
Chemical	**2**	**2**	**4**
New	**2**	**2**	**4**
Time	**2**	**2**	**4**
Carbon	**2**	**1**	**3**
Earlier	**2**	**1**	**3**
Research	1	5	6
Made	1	4	5
Form	1	3	4
Hand	1	3	4
Says	1	3	4
Chirality	1	2	3
Drugs	1	2	3
Mirror	1	2	3
Pheromone	1	2	3
Synthetic	1	2	3
Used	1	2	3
Way	1	2	3
Atoms	1	1	2
Auxilliaries	1	1	2
Body	1	1	2
Business	1	1	2
Chemistry	1	1	2
Commercial	1	1	2
Companies	1	1	2
Compounds	1	1	2
Difference	1	1	2
Effective	1	1	2

Words	*Th*	*Rh*	*Overall*
Example	1	1	2
Four	1	1	2
Group	1	1	2
Groups	1	1	2
Hands	1	1	2
Image	1	1	2
Major	1	1	2
Money	1	1	2
Pounds	1	1	2
Problem	1	1	2
Researchers	1	1	2
Sell	1	1	2
Team	1	1	2
Vegetables	1	1	2
Work	1	1	2
Palm	***4***	***0***	***4***
Many	***3***	***0***	***3***
Molecule	***3***	***0***	***3***
Reactions	***3***	***0***	***3***
Atom	***2***	***0***	***2***
Bp	***2***	***0***	***2***
Construct	***2***	***0***	***2***
Identifying	***2***	***0***	***2***
It's	***2***	***0***	***2***
Laboratory	***2***	***0***	***2***
Natural	***2***	***0***	***2***
Forms	*0*	*4*	*4*
Same	*0*	*4*	*4*
Called	*0*	*3*	*3*
Make	*0*	*3*	*3*
Molecular	*0*	*3*	*3*
Scaffolding	*0*	*3*	*3*
Use	*0*	*3*	*3*
Using	*0*	*3*	*3*
Ways	*0*	*3*	*3*
Amino	*0*	*2*	*2*
Amount	*0*	*2*	*2*
Backers	*0*	*2*	*2*
Building	*0*	*2*	*2*
Direction	*0*	*2*	*2*
Found	*0*	*2*	*2*
Fund	*0*	*2*	*2*
Gets	*0*	*2*	*2*
Gives	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Help	*0*	*2*	*2*
Human	*0*	*2*	*2*
Important	*0*	*2*	*2*
Keep	*0*	*2*	*2*
Known	*0*	*2*	*2*
Light	*0*	*2*	*2*
Mixture	*0*	*2*	*2*
Mould	*0*	*2*	*2*
Produces	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Proposals	*0*	*2*	*2*
Put	*0*	*2*	*2*
Selling	*0*	*2*	*2*
Size	*0*	*2*	*2*
Tiny	*0*	*2*	*2*
Took	*0*	*2*	*2*
Twist	*0*	*2*	*2*
Woodworms	*0*	*2*	*2*
Wrong	*0*	*2*	*2*

Sample Text B3: 950709

Words	*Th*	*Rh*	*Overall*
Species	**9**	**3**	**12**
America	**6**	**4**	**10**
Animals	**5**	**9**	**14**
Ioths	**5**	**4**	**9**
Climate	**5**	**2**	**7**
Native	**5**	**1**	**6**
South	**4**	**9**	**13**
Pleistocene	**4**	**1**	**5**
Human	**3**	**1**	**4**
Large	**3**	**1**	**4**
Southern	**3**	**1**	**4**
Years	**2**	**6**	**8**
Mammals	**2**	**3**	**5**
Extinctions	**2**	**2**	**4**
True	**2**	**2**	**4**
Evidence	**2**	**1**	**3**
Great	**2**	**1**	**3**
Many	**2**	**1**	**3**
Pliocene	**2**	**1**	**3**
Time	**2**	**1**	**3**
World	**2**	**1**	**3**
North	1	6	7
Mass	1	5	6
Big	1	3	4
Creatures	1	3	4
Extinct	1	3	4
Modern	1	3	4
Past	1	3	4
Ancestors	1	2	3
Changed	1	2	3
Continent	1	2	3
Elephants	1	2	3
Equator	1	2	3
Ground	1	2	3
Hoofed	1	2	3
New	1	2	3
Americas	1	1	2
Australia	1	1	2
Bears	1	1	2
Beings	1	1	2
Changes	1	1	2
Continents	1	1	2
Epoch	1	1	2
Extinction	1	1	2
Four	1	1	2
Islands	1	1	2
Larger	1	1	2
Marsupials	1	1	2

Words	*Th*	*Rh*	*Overall*
Mere	1	1	2
Met	1	1	2
Place	1	1	2
Present	1	1	2
Sloth	1	1	2
Today	1	1	2
Unique	1	1	2
Weird	1	1	2
Worldwide	1	1	2
Xenarthrans	1	1	2
Northern	***4***	***0***	***4***
Late	***3***	***0***	***3***
Martin	***3***	***0***	***3***
Northerners	***3***	***0***	***3***
Southerners	***3***	***0***	***3***
Biologists	***2***	***0***	***2***
Cause	***2***	***0***	***2***
Destruction	***2***	***0***	***2***
Destructivenes+	***2***	***0***	***2***
Events	***2***	***0***	***2***
Humans	***2***	***0***	***2***
Numbers	***2***	***0***	***2***
Right	***2***	***0***	***2***
University	***2***	***0***	***2***
Vrba	***2***	***0***	***2***
American	*0*	*7*	*7*
Million	*0*	*5*	*5*
Show	*0*	*4*	*4*
Disappeared	*0*	*3*	*3*
Including	*0*	*3*	*3*
Armadilloes	*0*	*2*	*2*
Bear	*0*	*2*	*2*
Coast	*0*	*2*	*2*
Coming	*0*	*2*	*2*
Died	*0*	*2*	*2*
Drier	*0*	*2*	*2*
Elephant	*0*	*2*	*2*
Felt	*0*	*2*	*2*
Formidable	*0*	*2*	*2*
Glyptodonts	*0*	*2*	*2*
Group	*0*	*2*	*2*
Horses	*0*	*2*	*2*
Host	*0*	*2*	*2*
Included	*0*	*2*	*2*
Land	*0*	*2*	*2*
Legs	*0*	*2*	*2*
Light	*0*	*2*	*2*
Myth	*0*	*2*	*2*
Panama	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Previously	*0*	*2*	*2*
Reasons	*0*	*2*	*2*
Resulted	*0*	*2*	*2*
Sea	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Stay	*0*	*2*	*2*
Survive	*0*	*2*	*2*
Toed	*0*	*2*	*2*
Versatile	*0*	*2*	*2*

Sample Text B4: 950820

Words	*Th*	*Rh*	*Overall*
Plants	**5**	**4**	**9**
Rocks	**5**	**1**	**6**
Evolution	**4**	**3**	**7**
Living	**4**	**1**	**5**
Lonsdale	**4**	**1**	**5**
House	**3**	**3**	**6**
Kew's	**3**	**1**	**4**
Million	**2**	**5**	**7**
Years	**2**	**5**	**7**
Period	**2**	**4**	**6**
Kew	**2**	**3**	**5**
Landscape	**2**	**3**	**5**
Time	**2**	**3**	**5**
Exhibition	**2**	**2**	**4**
Blackbird	**2**	**1**	**3**
Long	**2**	**1**	**3**
Visitors	1	4	5
Evolved	1	3	4
Extinct	1	3	4
Mosses	1	3	4
Sound	1	3	4
Air	1	2	3
Cooksonia	1	2	3
Life	1	2	3
Primeval	1	2	3
Age	1	1	2
Carboniferous	1	1	2
Corner	1	1	2
Idea	1	1	2
Land	1	1	2
Liverworts	1	1	2
Metres	1	1	2
Midst	1	1	2
Model	1	1	2
Organisms	1	1	2
Path	1	1	2
Planned	1	1	2
Scientific	1	1	2
Specimens	1	1	2
Swamp	1	1	2

Words	*Th*	*Rh*	*Overall*
Tree	1	1	2
Wonder	1	1	2
Cyanobacteria	***3***	***0***	***3***
Ferns	***3***	***0***	***3***
Glasshouse	***2***	***0***	***2***
Iron	***2***	***0***	***2***
Seed	***2***	***0***	***2***
Team	***2***	***0***	***2***
Ancient	*0*	*4*	*4*
Giant	*0*	*4*	*4*
Look	*0*	*4*	*4*
Dominated	*0*	*3*	*3*
Mud	*0*	*3*	*3*
Stems	*0*	*3*	*3*
Atmosphere	*0*	*2*	*2*
Best	*0*	*2*	*2*
Bubbling	*0*	*2*	*2*
Come	*0*	*2*	*2*
Distant	*0*	*2*	*2*
Grow	*0*	*2*	*2*
Immersion	*0*	*2*	*2*
Lycopsids	*0*	*2*	*2*
Made	*0*	*2*	*2*
Make	*0*	*2*	*2*
Many	*0*	*2*	*2*
Models	*0*	*2*	*2*
Oxygen	*0*	*2*	*2*
Plant	*0*	*2*	*2*
Pool	*0*	*2*	*2*
Pools	*0*	*2*	*2*
Primitive	*0*	*2*	*2*
Red	*0*	*2*	*2*
Says	*0*	*2*	*2*
Scene	*0*	*2*	*2*
Spores	*0*	*2*	*2*
Surface	*0*	*2*	*2*
Take	*0*	*2*	*2*

Sample Text B5: 950903

Words	*Th*	*Rh*	*Overall*
Public	**13**	**4**	**17**
Risk	**12**	**25**	**37**
Nuclear	**8**	**14**	**22**
People	**7**	**3**	**10**
Risks	**5**	**18**	**23**
Chemicals	**5**	**4**	**9**

Words	*Th*	*Rh*	*Overall*
Hse	**5**	**2**	**7**
Plant	**5**	**1**	**6**
Life	**4**	**3**	**7**
Smoking	**4**	**3**	**7**
Professor	**4**	**1**	**5**
Average	**3**	**3**	**6**

Words	*Th*	*Rh*	*Overall*
Many	**3**	**3**	**6**
Members	**3**	**2**	**5**
Person	**3**	**1**	**4**
Time	**3**	**1**	**4**
Death	**2**	**10**	**12**
Cancer	**2**	**7**	**9**
Industry	**2**	**7**	**9**
Year	**2**	**7**	**9**
Accidents	**2**	**5**	**7**
Power	**2**	**5**	**7**
Same	**2**	**3**	**5**
Put	**2**	**2**	**4**
Years	**2**	**2**	**4**
Big	**2**	**1**	**3**
Day	**2**	**1**	**3**
Deaths	**2**	**1**	**3**
Major	**2**	**1**	**3**
Paling	**2**	**1**	**3**
Public's	**2**	**1**	**3**
Today	**2**	**1**	**3**
Says	1	7	8
Accident	1	5	6
Result	1	5	6
Known	1	3	4
Large	1	3	4
Media	1	3	4
Radiation	1	3	4
Biggest	1	2	3
Billion	1	2	3
Cause	1	2	3
Found	1	2	3
Maximum	1	2	3
Relatively	1	2	3
Scale	1	2	3
Society	1	2	3
Water	1	2	3
Ability	1	1	2
Bed	1	1	2
Botulism	1	1	2
Buying	1	1	2
Cigarette	1	1	2
Common	1	1	2
Dangerous	1	1	2
Dying	1	1	2
Environmental	1	1	2
Fear	1	1	2
Fears	1	1	2
Harbison	1	1	2
Herbal	1	1	2
Increasingly	1	1	2
Industries	1	1	2
Involving	1	1	2
Know	1	1	2
Mass	1	1	2
Meat	1	1	2
Occur	1	1	2
Perception	1	1	2
Perspective	1	1	2
Remedies	1	1	2
Safety	1	1	2
Station	1	1	2
Reality	***3***	***0***	***3***
Americans	***2***	***0***	***2***

Words	*Th*	*Rh*	*Overall*
Analytical	***2***	***0***	***2***
Chief	***2***	***0***	***2***
End	***2***	***0***	***2***
Experts	***2***	***0***	***2***
Oxford	***2***	***0***	***2***
Peto	***2***	***0***	***2***
Smokers	***2***	***0***	***2***
Staying	***2***	***0***	***2***
Survey	***2***	***0***	***2***
Worker	***2***	***0***	***2***
Young	***2***	***0***	***2***
Million	*0*	*5*	*5*
Affect	*0*	*4*	*4*
Health	*0*	*4*	*4*
Parts	*0*	*4*	*4*
Argue	*0*	*3*	*3*
Believe	*0*	*3*	*3*
Causes	*0*	*3*	*3*
Come	*0*	*3*	*3*
Difficult	*0*	*3*	*3*
Estimate	*0*	*3*	*3*
Long	*0*	*3*	*3*
Low	*0*	*3*	*3*
Plants	*0*	*3*	*3*
Reduce	*0*	*3*	*3*
Times	*0*	*3*	*3*
Avoidable	*0*	*2*	*2*
Comes	*0*	*2*	*2*
Compared	*0*	*2*	*2*
Construction	*0*	*2*	*2*
Damage	*0*	*2*	*2*
Decades	*0*	*2*	*2*
Die	*0*	*2*	*2*
Dies	*0*	*2*	*2*
Different	*0*	*2*	*2*
Doses	*0*	*2*	*2*
Environment	*0*	*2*	*2*
Exceed	*0*	*2*	*2*
Fairly	*0*	*2*	*2*
Full	*0*	*2*	*2*
High	*0*	*2*	*2*
Increase	*0*	*2*	*2*
Killed	*0*	*2*	*2*
Levels	*0*	*2*	*2*
Limit	*0*	*2*	*2*
Man	*0*	*2*	*2*
Ones	*0*	*2*	*2*
Overestimate	*0*	*2*	*2*
Possible	*0*	*2*	*2*
Practicable	*0*	*2*	*2*
Reasonably	*0*	*2*	*2*
Seconds	*0*	*2*	*2*
Seem	*0*	*2*	*2*
Show	*0*	*2*	*2*
Smaller	*0*	*2*	*2*
Term	*0*	*2*	*2*
Thousand	*0*	*2*	*2*
Thousands	*0*	*2*	*2*
Tobacco	*0*	*2*	*2*
Tolerable	*0*	*2*	*2*
Weight	*0*	*2*	*2*
Workers	*0*	*2*	*2*

Sample Text B6: 951210

Words	*Th*	*Rh*	*Overall*
Bse	**9**	**7**	**16**
Government	**5**	**2**	**7**
Scientists	**5**	**1**	**6**
Problem	**4**	**3**	**7**
Cjd	**4**	**2**	**6**
Scrapie	**4**	**1**	**5**
Agent	**3**	**3**	**6**
Experiment	**3**	**2**	**5**
Scientific	**2**	**7**	**9**
Beef	**2**	**3**	**5**
Farmers	**2**	**3**	**5**
Positive	**2**	**3**	**5**
British	**2**	**2**	**4**
John	**2**	**2**	**4**
Known	**2**	**2**	**4**
Risk	**2**	**2**	**4**
Species	**2**	**2**	**4**
Aids	**2**	**1**	**3**
Animal	**2**	**1**	**3**
Best	**2**	**1**	**3**
Different	**2**	**1**	**3**
Infect	**2**	**1**	**3**
Option	**2**	**1**	**3**
Original	**2**	**1**	**3**
Professor	**2**	**1**	**3**
Working	**2**	**1**	**3**
Disease	1	4	5
Country	1	3	4
Cows	1	3	4
Evidence	1	3	4
Humans	1	3	4
Resources	1	3	4
Body	1	2	3
Bruce s	1	2	3
Cattle	1	2	3
Diseases	1	2	3
Eating	1	2	3
Protein	1	2	3
Same	1	2	3
Sheep	1	2	3
Spread	1	2	3
Wasn't	1	2	3
Years	1	2	3
Animals	1	1	2
Bruce	1	1	2
Cause	1	1	2
Causing	1	1	2
Defend	1	1	2
Effective	1	1	2
Encephalopathy	1	1	2
Full	1	1	2
Gummer	1	1	2
Happy	1	1	2
Health	1	1	2
Helped	1	1	2
International	1	1	2
Issue	1	1	2

Words	*Th*	*Rh*	*Overall*
Larger	1	1	2
Pattison	1	1	2
People	1	1	2
Prions	1	1	2
Proteins	1	1	2
Result	1	1	2
Scientist	1	1	2
Spongiform	1	1	2
Statement	1	1	2
Suggesting	1	1	2
Unit	1	1	2
Minister	***3***	***0***	***3***
Decade	***2***	***0***	***2***
Many	***2***	***0***	***2***
Mice	***2***	***0***	***2***
Results	***2***	***0***	***2***
Set	***2***	***0***	***2***
Sets	***2***	***0***	***2***
Team	***2***	***0***	***2***
There's	***2***	***0***	***2***
Work	***2***	***0***	***2***
Brain	*0*	*4*	*4*
Said	*0*	*4*	*4*
Affected	*0*	*3*	*3*
Passed	*0*	*3*	*3*
Transmitted	*0*	*3*	*3*
Wrong	*0*	*3*	*3*
Called	*0*	*2*	*2*
Caused	*0*	*2*	*2*
Cord	*0*	*2*	*2*
Enough	*0*	*2*	*2*
Epidemic	*0*	*2*	*2*
Equivalent	*0*	*2*	*2*
Food	*0*	*2*	*2*
Forced	*0*	*2*	*2*
Human	*0*	*2*	*2*
Kill	*0*	*2*	*2*
Knew	*0*	*2*	*2*
Know	*0*	*2*	*2*
Led	*0*	*2*	*2*
Mean	*0*	*2*	*2*
Possible	*0*	*2*	*2*
Proof	*0*	*2*	*2*
Public	*0*	*2*	*2*
Reputation	*0*	*2*	*2*
Research	*0*	*2*	*2*
Science	*0*	*2*	*2*
Spinal	*0*	*2*	*2*
Take	*0*	*2*	*2*
Think	*0*	*2*	*2*
Unknown	*0*	*2*	*2*

Sample Text B7: 960407

Words	*Th*	*Rh*	*Overall*
Glucose	**8**	**5**	**13**
Sugars	**7**	**11**	**18**
Sugoid	**6**	**1**	**7**
Virus	**5**	**2**	**7**
Cells	**4**	**2**	**6**
Sugoids	**3**	**6**	**9**
Blood	**3**	**2**	**5**
Molecules	**2**	**2**	**4**
Sugar	**2**	**2**	**4**
Cancer	**2**	**1**	**3**
Coating	**2**	**1**	**3**
Levels	**2**	**1**	**3**
Professor	**2**	**1**	**3**
Proteins	**2**	**1**	**3**
Viruses	**2**	**1**	**3**
Look	1	4	5
Toxic	1	4	5
Make	1	3	4
Work	1	3	4
Enzyme	1	2	3
Example	1	2	3
Fold	1	2	3
Hepatitis	1	2	3
Human	1	2	3
Mannose	1	2	3
Many	1	2	3
Plants	1	2	3
Powerful	1	2	3
Attached	1	1	2
B	1	1	2
Chemicals	1	1	2
Compounds	1	1	2
Disease	1	1	2
Diversity	1	1	2
Dwek	1	1	2
Found	1	1	2
Head	1	1	2
Hiv	1	1	2
Institute	1	1	2
Known	1	1	2
Longer	1	1	2
Made	1	1	2
Mannoid	1	1	2
Mimics	1	1	2
Oxford's	1	1	2
Reach	1	1	2
Seems	1	1	2
Sperm	1	1	2
Sticks	1	1	2
Sugared	1	1	2
System	1	1	2
Weight	1	1	2
World's	1	1	2

Words	*Th*	*Rh*	*Overall*
Coat	***3***	***0***	***3***
George	***3***	***0***	***3***
Hbv	***3***	***0***	***3***
Add	***2***	***0***	***2***
Analogue	***2***	***0***	***2***
Complex	***2***	***0***	***2***
Disrupt	***2***	***0***	***2***
Hiv's	***2***	***0***	***2***
It's	***2***	***0***	***2***
Laboratory	***2***	***0***	***2***
Liver	***2***	***0***	***2***
Louise	***2***	***0***	***2***
Mistaken	***2***	***0***	***2***
Molecular	***2***	***0***	***2***
Nbdnj	***2***	***0***	***2***
Reproducing	***2***	***0***	***2***
Says	*0*	*7*	*7*
Active	*0*	*3*	*3*
Good	*0*	*3*	*3*
Making	*0*	*3*	*3*
Put	*0*	*3*	*3*
Activity	*0*	*2*	*2*
Anti	*0*	*2*	*2*
Biologically	*0*	*2*	*2*
Body	*0*	*2*	*2*
Chemical	*0*	*2*	*2*
Detection	*0*	*2*	*2*
Diabetes	*0*	*2*	*2*
Difference	*0*	*2*	*2*
Glycogen	*0*	*2*	*2*
Glycosidase	*0*	*2*	*2*
Immune	*0*	*2*	*2*
Job	*0*	*2*	*2*
Living	*0*	*2*	*2*
Long	*0*	*2*	*2*
Metabolic	*0*	*2*	*2*
Metabolism	*0*	*2*	*2*
Need	*0*	*2*	*2*
Organism	*0*	*2*	*2*
Promising	*0*	*2*	*2*
Result	*0*	*2*	*2*
Same	*0*	*2*	*2*
Shape	*0*	*2*	*2*
Surface	*0*	*2*	*2*
Target	*0*	*2*	*2*
Term	*0*	*2*	*2*
Therapeutic	*0*	*2*	*2*
Ways	*0*	*2*	*2*

Sample Text B8: 960519

Words	*Th*	*Rh*	*Overall*
Cerro	**7**	**3**	**10**
Paranal	**5**	**2**	**7**

Words	*Th*	*Rh*	*Overall*
Telescope	**4**	**4**	**8**
Observatory	**4**	**2**	**6**

Words	*Th*	*Rh*	*Overall*
Mirror	**3**	**2**	**5**
Site	**3**	**1**	**4**
Light	**2**	**7**	**9**
Space	**2**	**4**	**6**
High	**2**	**3**	**5**
Vlt	**2**	**3**	**5**
Chilean	**2**	**2**	**4**
Astronomers	**2**	**1**	**3**
Difference	**2**	**1**	**3**
Instrument	**2**	**1**	**3**
Observation	**2**	**1**	**3**
Surface	**2**	**1**	**3**
Time	1	3	4
Building	1	2	3
Installed	1	2	3
Little	1	2	3
Mirrors	1	2	3
Use	1	2	3
Area	1	1	2
Astronomy	1	1	2
Atacama	1	1	2
Atmosphere	1	1	2
Babylon	1	1	2
Bank	1	1	2
Big	1	1	2
Clouds	1	1	2
Collecting	1	1	2
Cost	1	1	2
De	1	1	2
Desert	1	1	2
Dust	1	1	2
European	1	1	2
Fraction	1	1	2
Hubble	1	1	2
Hubble s	1	1	2
Infra	1	1	2
Installation	1	1	2
Jodrell	1	1	2
Life	1	1	2
Main	1	1	2
Occasional	1	1	2
Ocean	1	1	2

Words	*Th*	*Rh*	*Overall*
Power	1	1	2
Red	1	1	2
Similar	1	1	2
Size	1	1	2
Southern	1	1	2
Chile	***4***	***0***	***4***
Isaac	***2***	***0***	***2***
Natural	***2***	***0***	***2***
Paris	***2***	***0***	***2***
Precision	***2***	***0***	***2***
Wolfgang	***2***	***0***	***2***
Says	*0*	*6*	*6*
See	*0*	*6*	*6*
Made	*0*	*3*	*3*
Scientists	*0*	*3*	*3*
Tell	*0*	*3*	*3*
Tolerance	*0*	*3*	*3*
Trained	*0*	*3*	*3*
Achieved	*0*	*2*	*2*
Antofagasta	*0*	*2*	*2*
Carried	*0*	*2*	*2*
Come	*0*	*2*	*2*
Controlled	*0*	*2*	*2*
Covered	*0*	*2*	*2*
Cut	*0*	*2*	*2*
Demanding	*0*	*2*	*2*
Different	*0*	*2*	*2*
Earth	*0*	*2*	*2*
Eso	*0*	*2*	*2*
Giant	*0*	*2*	*2*
Give	*0*	*2*	*2*
Planet	*0*	*2*	*2*
Planned	*0*	*2*	*2*
Powerful	*0*	*2*	*2*
Sitting	*0*	*2*	*2*
Stars	*0*	*2*	*2*
System	*0*	*2*	*2*
Take	*0*	*2*	*2*
Totally	*0*	*2*	*2*
Vlt s	*0*	*2*	*2*
Year	*0*	*2*	*2*
Years	*0*	*2*	*2*

Sample Text B9: 960526

Words	*Th*	*Rh*	*Overall*
Shopping	**10**	**8**	**18**
Digital	**5**	**1**	**6**
Store	**4**	**6**	**10**
Technology	**4**	**3**	**7**
Supermarket	**3**	**5**	**8**
List	**3**	**3**	**6**
Customers	**3**	**2**	**5**
Home	**3**	**1**	**4**
Retailers	**3**	**1**	**4**
Stores	**3**	**1**	**4**
Systems	**3**	**1**	**4**
Electronic	**2**	**3**	**5**
New	**2**	**3**	**5**
Customer	**2**	**1**	**3**
Items	**2**	**1**	**3**

Words	*Th*	*Rh*	*Overall*
Research	**2**	**1**	**3**
Virtual	**2**	**1**	**3**
Loyalty	1	4	5
Time	1	4	5
Pc	1	3	4
Screen	1	3	4
Check	1	2	3
Market	1	2	3
People	1	2	3
Service	1	2	3
Shop	1	2	3
Trip	1	2	3
Access	1	1	2
Allows	1	1	2
Appetite	1	1	2

Words	*Th*	*Rh*	*Overall*
Centre	1	1	2
Checklist	1	1	2
Consumer	1	1	2
Consumers	1	1	2
Delivery	1	1	2
Form	1	1	2
Gain	1	1	2
Interactive	1	1	2
Month	1	1	2
Point	1	1	2
Sainsbury's	1	1	2
Trust	1	1	2
Year	1	1	2
Icl	***3***	***0***	***3***
Retail	***3***	***0***	***3***
Alun	***2***	***0***	***2***
British	***2***	***0***	***2***
Goods	***2***	***0***	***2***
Kiosk	***2***	***0***	***2***
Manager	***2***	***0***	***2***
Media	***2***	***0***	***2***
Power	***2***	***0***	***2***
Roberts	***2***	***0***	***2***
Wedding	***2***	***0***	***2***
Card	*0*	*10*	*10*
Believes	*0*	*5*	*5*
Internet	*0*	*4*	*4*
Smart	*0*	*4*	*4*
Stored	*0*	*4*	*4*
Go	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Help	*0*	*3*	*3*
Income	*0*	*3*	*3*
Information	*0*	*3*	*3*
Take	*0*	*3*	*3*
Added	*0*	*2*	*2*
Based	*0*	*2*	*2*
Buy	*0*	*2*	*2*
Buying	*0*	*2*	*2*
Choices	*0*	*2*	*2*
Computer	*0*	*2*	*2*
Convenient	*0*	*2*	*2*
Course	*0*	*2*	*2*
Experience	*0*	*2*	*2*
Found	*0*	*2*	*2*
Get	*0*	*2*	*2*
Getting	*0*	*2*	*2*
Going	*0*	*2*	*2*
Good	*0*	*2*	*2*
Habits	*0*	*2*	*2*
Having	*0*	*2*	*2*
Located	*0*	*2*	*2*
Need	*0*	*2*	*2*
Order	*0*	*2*	*2*
Ready	*0*	*2*	*2*
Relationship	*0*	*2*	*2*
Shops	*0*	*2*	*2*
Sounds	*0*	*2*	*2*
Used	*0*	*2*	*2*
Using	*0*	*2*	*2*
Wants	*0*	*2*	*2*

Sample Text B10: 960623

Words	*Th*	*Rh*	*Overall*
Nature	**15**	**4**	**19**
People	**8**	**3**	**11**
Many	**6**	**4**	**10**
Reading	**4**	**1**	**5**
Countryside	**3**	**5**	**8**
Man	**3**	**2**	**5**
Urban	**3**	**2**	**5**
Buildings	**3**	**1**	**4**
Humans	**3**	**1**	**4**
Example	**2**	**6**	**8**
Made	**2**	**4**	**6**
Air	**2**	**2**	**4**
Largely	**2**	**2**	**4**
See	**2**	**2**	**4**
Sensory	**2**	**2**	**4**
Century	**2**	**1**	**3**
God	**2**	**1**	**3**
Landscape	**2**	**1**	**3**
Open	**2**	**1**	**3**
Say	**2**	**1**	**3**
Natural	1	5	6
Found	1	4	5
Landscapes	1	4	5
Water	1	4	5
World	1	4	5
Called	1	3	4
Experiences	1	3	4

Words	*Th*	*Rh*	*Overall*
Groups	1	3	4
Human	1	3	4
Need	1	3	4
Plants	1	3	4
Developed	1	2	3
Force	1	2	3
Health	1	2	3
Lives	1	2	3
Trees	1	2	3
Artificial	1	1	2
Enjoyed	1	1	2
Environment	1	1	2
Environmental	1	1	2
Fascination	1	1	2
Fresh	1	1	2
Holiday	1	1	2
Kaplans	1	1	2
Light	1	1	2
Mystery	1	1	2
Nearly	1	1	2
New	1	1	2
Park	1	1	2
Patients	1	1	2
Perhaps	1	1	2
Place	1	1	2
Psychologist	1	1	2
Religion	1	1	2
Religious	1	1	2

Words	*Th*	*Rh*	*Overall*
Remarkable	1	1	2
Seen	1	1	2
Soft	1	1	2
Third	1	1	2
Town	1	1	2
Vital	1	1	2
Year	1	1	2
Study	***9***	***0***	***9***
University	***6***	***0***	***6***
Wilderness	***3***	***0***	***3***
American	***2***	***0***	***2***
Annual	***2***	***0***	***2***
Asked	***2***	***0***	***2***
Blues	***2***	***0***	***2***
Childhood	***2***	***0***	***2***
Commission	***2***	***0***	***2***
Cultures	***2***	***0***	***2***
Describe	***2***	***0***	***2***
Hospital	***2***	***0***	***2***
Modern	***2***	***0***	***2***
Movement	***2***	***0***	***2***
Recent	***2***	***0***	***2***
Researchers	***2***	***0***	***2***
Residents	***2***	***0***	***2***
Roots	***2***	***0***	***2***
Science	***2***	***0***	***2***
Society	***2***	***0***	***2***
Sunsets	***2***	***0***	***2***
Surveys	***2***	***0***	***2***
Therapy	***2***	***0***	***2***
William	***2***	***0***	***2***
Words	***2***	***0***	***2***
Showed	*0*	*4*	*4*
Alive	*0*	*3*	*3*
Cities	*0*	*3*	*3*
Country	*0*	*3*	*3*
Known	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Offers	*0*	*3*	*3*
Sense	*0*	*3*	*3*
Aggressive	*0*	*2*	*2*
Attention	*0*	*2*	*2*
Better	*0*	*2*	*2*
City	*0*	*2*	*2*
Colour	*0*	*2*	*2*
Desire	*0*	*2*	*2*
Doors	*0*	*2*	*2*
Experience	*0*	*2*	*2*
Experienced	*0*	*2*	*2*
Followed	*0*	*2*	*2*
Forest	*0*	*2*	*2*
Freedom	*0*	*2*	*2*
Happier	*0*	*2*	*2*
Linked	*0*	*2*	*2*
Living	*0*	*2*	*2*
Mental	*0*	*2*	*2*
Moments	*0*	*2*	*2*
Moods	*0*	*2*	*2*
Move	*0*	*2*	*2*
Mystical	*0*	*2*	*2*
Parks	*0*	*2*	*2*
Potent	*0*	*2*	*2*
Provide	*0*	*2*	*2*
Psychologists	*0*	*2*	*2*
Put	*0*	*2*	*2*
Represent	*0*	*2*	*2*
Retain	*0*	*2*	*2*
Shown	*0*	*2*	*2*
System	*0*	*2*	*2*
Things	*0*	*2*	*2*
Think	*0*	*2*	*2*
Time	*0*	*2*	*2*
Use	*0*	*2*	*2*
Visit	*0*	*2*	*2*
Walk	*0*	*2*	*2*

Sample Text B11: 960630

Words	*Th*	*Rh*	*Overall*
Apes	**13**	**10**	**23**
Language	**12**	**10**	**22**
Professor	**12**	**1**	**13**
Chimps	**7**	**1**	**8**
Sign	**6**	**6**	**12**
Savage	**6**	**1**	**7**
Wild	**5**	**2**	**7**
Rumbaugh	**5**	**1**	**6**
Use	**4**	**10**	**14**
Chimpanzees	**4**	**5**	**9**
Food	**3**	**9**	**12**
Chimp	**3**	**5**	**8**
Research	**3**	**4**	**7**
Signing	**3**	**2**	**5**
Booee	**3**	**1**	**4**
Katharine	**3**	**1**	**4**
Year	**3**	**1**	**4**
Words	**2**	**5**	**7**

Words	*Th*	*Rh*	*Overall*
Children	**2**	**4**	**6**
People	**2**	**3**	**5**
Animals	**2**	**2**	**4**
Great	**2**	**2**	**4**
Kanzi	**2**	**2**	**4**
Centre	**2**	**1**	**3**
Chimpanzee	**2**	**1**	**3**
Different	**2**	**1**	**3**
Mike	**2**	**1**	**3**
Primate	**2**	**1**	**3**
University	**2**	**1**	**3**
Usually	**2**	**1**	**3**
Human	1	7	8
Drink	1	3	4
Used	1	3	4
American	1	2	3
Asked	1	2	3
Done	1	2	3
Name	1	2	3

Words	*Th*	*Rh*	*Overall*
Want	1	2	3
Way	1	2	3
Years	1	2	3
Young	1	2	3
Allowed	1	1	2
Amy	1	1	2
Animal	1	1	2
Atlanta	1	1	2
Better	1	1	2
Bonobo	1	1	2
Companions	1	1	2
Cry	1	1	2
French	1	1	2
Giving	1	1	2
Kanzi's	1	1	2
Left	1	1	2
Means	1	1	2
New	1	1	2
Old	1	1	2
Oldest	1	1	2
Signs	1	1	2
State	1	1	2
Uses	1	1	2
Fouts	***4***	***0***	***4***
Panbanisha	***3***	***0***	***3***
Penny	***3***	***0***	***3***
Rivas	***3***	***0***	***3***
Washoe	***3***	***0***	***3***
Esteban	***2***	***0***	***2***
Michael	***2***	***0***	***2***
Ones	***2***	***0***	***2***
Seen	***2***	***0***	***2***
Sherman	***2***	***0***	***2***
Word	***2***	***0***	***2***
Said	*0*	*7*	*7*
States	*0*	*5*	*5*
Ask	*0*	*4*	*4*
English	*0*	*4*	*4*
Mind	*0*	*4*	*4*
Portions	*0*	*4*	*4*
Speak	*0*	*4*	*4*
Understand	*0*	*4*	*4*
Ape	*0*	*3*	*3*
Behaviour	*0*	*3*	*3*
Care	*0*	*3*	*3*
Communicate	*0*	*3*	*3*
Communicating	*0*	*3*	*3*
Give	*0*	*3*	*3*
Hand	*0*	*3*	*3*
Looked	*0*	*3*	*3*
Mental	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Question	*0*	*3*	*3*
Rudimentary	*0*	*3*	*3*
World	*0*	*3*	*3*
Called	*0*	*2*	*2*
Child	*0*	*2*	*2*
Continue	*0*	*2*	*2*
Cup	*0*	*2*	*2*
Deaf	*0*	*2*	*2*
Demonstrator	*0*	*2*	*2*
Describe	*0*	*2*	*2*
Eating	*0*	*2*	*2*
Egotistical	*0*	*2*	*2*
Emotions	*0*	*2*	*2*
Enough	*0*	*2*	*2*
Found	*0*	*2*	*2*
Gestures	*0*	*2*	*2*
Get	*0*	*2*	*2*
Going	*0*	*2*	*2*
Gorillas	*0*	*2*	*2*
Hepatitis	*0*	*2*	*2*
Institute	*0*	*2*	*2*
Internal	*0*	*2*	*2*
Lemsip	*0*	*2*	*2*
Life	*0*	*2*	*2*
Lying	*0*	*2*	*2*
Make	*0*	*2*	*2*
Money	*0*	*2*	*2*
National	*0*	*2*	*2*
Outlook	*0*	*2*	*2*
Played	*0*	*2*	*2*
Pygmy	*0*	*2*	*2*
Raise	*0*	*2*	*2*
Same	*0*	*2*	*2*
Saw	*0*	*2*	*2*
Say	*0*	*2*	*2*
Science	*0*	*2*	*2*
Seems	*0*	*2*	*2*
Share	*0*	*2*	*2*
Shown	*0*	*2*	*2*
Signed	*0*	*2*	*2*
Sweet	*0*	*2*	*2*
Symbol	*0*	*2*	*2*
Symbols	*0*	*2*	*2*
Taught	*0*	*2*	*2*
Thought	*0*	*2*	*2*
Trained	*0*	*2*	*2*
True	*0*	*2*	*2*
Using	*0*	*2*	*2*
Wanted	*0*	*2*	*2*
Ways	*0*	*2*	*2*

Appendix 10. Word lists of Group C Sample Texts

(Frequency >=2)

Sample Text C1: 960707

Words	*Th*	*Rh*	*Overall*
Earth	19	11	30
Mantle	13	22	35
Core	13	17	30
Seismic	7	6	13
Waves	6	4	10
Centre	6	3	9
Earth's	6	2	8
Structure	5	4	9
Crust	5	1	6
Magnetic	4	6	10
Change	4	4	8
Liquid	4	2	6
Phase	4	2	6
Surface	3	7	10
Iron	3	5	8
Layer	3	4	7
Boundary	3	3	6
Diamonds	3	3	6
Earthquakes	3	3	6
Ocean	3	3	6
Pressures	3	3	6
Depth	3	2	5
Geologists	3	2	5
Heat	3	2	5
Material	3	2	5
Interior	3	1	4
Minerals	3	1	4
Perovskite	3	1	4
Scientists	3	1	4
Temperature	3	1	4
Pressure	2	12	14
Field	2	7	9
Deep	2	6	8
Years	2	6	8
Diamond	2	5	7
Hot	2	5	7
Cent	2	3	5
Crystal	2	3	5
Miles	2	3	5

Words	*Th*	*Rh*	*Overall*
Temperatures	2	3	5
Base	2	2	4
Entire	2	2	4
Evidence	2	2	4
Great	2	2	4
Journey	2	2	4
Mineral	2	2	4
Rocks	2	2	4
Come	2	1	3
Drill	2	1	3
Earthquake	2	1	3
Gun	2	1	3
Laboratory	2	1	3
Shock	2	1	3
Space	2	1	3
Sun	2	1	3
Time	2	1	3
Tomography	2	1	3
Rock	1	8	9
Solid	1	6	7
Know	1	4	5
Little	1	4	5
Lower	1	4	5
Many	1	4	5
Same	1	4	5
Sample	1	4	5
Cold	1	3	4
Hotter	1	3	4
Made	1	3	4
Reveal	1	3	4
Slowly	1	3	4
Thick	1	3	4
Bulk	1	2	3
Cooler	1	2	3
Form	1	2	3
Molten	1	2	3
Planet's	1	2	3
Probably	1	2	3
Record	1	2	3
Spinel	1	2	3
Vast	1	2	3

Words	Th	Rh	Overall
Work	1	2	3
Anvil	1	1	2
Areas	1	1	2
Bolivia	1	1	2
California	1	1	2
Canada	1	1	2
Chemical	1	1	2
Churning	1	1	2
Colder	1	1	2
Comes	1	1	2
Comparatively	1	1	2
Conditions	1	1	2
D	1	1	2
Data	1	1	2
Decay	1	1	2
Earlier	1	1	2
Explanation	1	1	2
Floor	1	1	2
Four	1	1	2
Geology	1	1	2
Ground	1	1	2
Human	1	1	2
Knowledge	1	1	2
Long	1	1	2
Packed	1	1	2
Particles	1	1	2
Perhaps	1	1	2
Plates	1	1	2
Presses	1	1	2
Published	1	1	2
Rate	1	1	2
Released	1	1	2
Scum	1	1	2
Simulate	1	1	2
South	1	1	2
Stations	1	1	2
Thin	1	1	2
Undergo	1	1	2
Wave	1	1	2
Window	1	1	2
World	1	1	2
Professor	***3***	***0***	***3***
Simulations	***3***	***0***	***3***
Today	***3***	***0***	***3***
Volcanoes	***3***	***0***	***3***
Average	***2***	***0***	***2***
Body	***2***	***0***	***2***
Controversies	***2***	***0***	***2***
Conveniently	***2***	***0***	***2***
Descending	***2***	***0***	***2***
Feet	***2***	***0***	***2***
Gained	***2***	***0***	***2***
Masters	***2***	***0***	***2***
People	***2***	***0***	***2***
Recent	***2***	***0***	***2***
Ruby	***2***	***0***	***2***
Slab	***2***	***0***	***2***
Spite	***2***	***0***	***2***
Technique	***2***	***0***	***2***
Circulation	*0*	*6*	*6*
Seems	*0*	*6*	*6*
Hard	*0*	*5*	*5*
Times	*0*	*5*	*5*

Words	Th	Rh	Overall
Way	*0*	*5*	*5*
Different	*0*	*4*	*4*
Place	*0*	*4*	*4*
Points	*0*	*4*	*4*
Travel	*0*	*4*	*4*
Called	*0*	*3*	*3*
Continents	*0*	*3*	*3*
Freezing	*0*	*3*	*3*
High	*0*	*3*	*3*
Known	*0*	*3*	*3*
Materials	*0*	*3*	*3*
Melting	*0*	*3*	*3*
Millions	*0*	*3*	*3*
Moon	*0*	*3*	*3*
Planet	*0*	*3*	*3*
Point	*0*	*3*	*3*
Possible	*0*	*3*	*3*
Process	*0*	*3*	*3*
Represents	*0*	*3*	*3*
Simple	*0*	*3*	*3*
Takes	*0*	*3*	*3*
Typically	*0*	*3*	*3*
Upper	*0*	*3*	*3*
Uses	*0*	*3*	*3*
Acts	*0*	*2*	*2*
Africa	*0*	*2*	*2*
Analyse	*0*	*2*	*2*
Appear	*0*	*2*	*2*
Atmospheric	*0*	*2*	*2*
Becomes	*0*	*2*	*2*
Beginning	*0*	*2*	*2*
Big	*0*	*2*	*2*
Break	*0*	*2*	*2*
Can't	*0*	*2*	*2*
Clear	*0*	*2*	*2*
Compared	*0*	*2*	*2*
Concentric	*0*	*2*	*2*
Crack	*0*	*2*	*2*
Created	*0*	*2*	*2*
Cross	*0*	*2*	*2*
Denser	*0*	*2*	*2*
Dynamic	*0*	*2*	*2*
Dynamo	*0*	*2*	*2*
Expected	*0*	*2*	*2*
Face	*0*	*2*	*2*
Fiction	*0*	*2*	*2*
Flattened	*0*	*2*	*2*
Formed	*0*	*2*	*2*
Giant	*0*	*2*	*2*
Green	*0*	*2*	*2*
Happens	*0*	*2*	*2*
Layers	*0*	*2*	*2*
Left	*0*	*2*	*2*
Look	*0*	*2*	*2*
Make	*0*	*2*	*2*
Marks	*0*	*2*	*2*
Microscopic	*0*	*2*	*2*
Millionth	*0*	*2*	*2*
Motion	*0*	*2*	*2*
New	*0*	*2*	*2*
Olivine	*0*	*2*	*2*
Onion	*0*	*2*	*2*
Perfectly	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Processes	*0*	*2*	*2*
Provides	*0*	*2*	*2*
Question	*0*	*2*	*2*
Quickly	*0*	*2*	*2*
Reach	*0*	*2*	*2*
Reversal	*0*	*2*	*2*
Rocky	*0*	*2*	*2*
See	*0*	*2*	*2*
Series	*0*	*2*	*2*
Similar	*0*	*2*	*2*
Size	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Speeds	*0*	*2*	*2*
Suggest	*0*	*2*	*2*
Technology	*0*	*2*	*2*
Tell	*0*	*2*	*2*
Thought	*0*	*2*	*2*
Tons	*0*	*2*	*2*
Using	*0*	*2*	*2*
West	*0*	*2*	*2*

Sample Text C2: 960714

Words	*Th*	*Rh*	*Overall*
Pleasure	**8**	**8**	**16**
Research	**6**	**1**	**6**
Warburton	**6**	**1**	**7**
Professor	**4**	**1**	**5**
Stress	**3**	**3**	**6**
Chocolate	**3**	**1**	**4**
University	**3**	**1**	**4**
Guilt	**2**	**3**	**5**
People	**2**	**2**	**4**
Arise	**2**	**1**	**3**
Blood	**2**	**1**	**3**
Depressed	**2**	**1**	**3**
Says	1	3	4
Better	1	2	3
Many	1	2	3
Pathway	1	2	3
Study	1	2	3
System	1	2	3
Alcohol	1	1	2
Body's	1	1	2
Caffeine	1	1	2
Coffee	1	1	2
Combination	1	1	2
Conducted	1	1	2
Experiencing	1	1	2
Explains	1	1	2
Happiness	1	1	2
Indulgence	1	1	2
Infections	1	1	2
Journal	1	1	2
Key	1	1	2
Levels	1	1	2
Measured	1	1	2
Pleasures	1	1	2
Respiratory	1	1	2
Survey	1	1	2
Systems	1	1	2
Taste	1	1	2

Words	*Th*	*Rh*	*Overall*
Techniques	1	1	2
Triggers	1	1	2
Arise's	***3***	***0***	***3***
Control	***3***	***0***	***3***
Group	***3***	***0***	***3***
Carruthers	***2***	***0***	***2***
Forbidden	***2***	***0***	***2***
Laughter	***2***	***0***	***2***
Nicotine	***2***	***0***	***2***
Patients	***2***	***0***	***2***
Philosophers	***2***	***0***	***2***
Pleasurable	***2***	***0***	***2***
Reading	***2***	***0***	***2***
Brain	*0*	*4*	*3*
Good	*0*	*3*	*3*
Measurably	*0*	*3*	*3*
Nervous	*0*	*3*	*3*
Produces	*0*	*3*	*3*
Absolute	*0*	*2*	*2*
Areas	*0*	*2*	*2*
Central	*0*	*2*	*2*
Feel	*0*	*2*	*2*
Foods	*0*	*2*	*2*
Healthy	*0*	*2*	*2*
Humans	*0*	*2*	*2*
Improvement	*0*	*2*	*2*
Long	*0*	*2*	*2*
Lower	*0*	*2*	*2*
Mind	*0*	*2*	*2*
Natural	*0*	*2*	*2*
Performance	*0*	*2*	*2*
Problems	*0*	*2*	*2*
Proven	*0*	*2*	*2*
Rises	*0*	*2*	*2*
Things	*0*	*2*	*2*
Uk	*0*	*2*	*2*
Work	*0*	*2*	*2*

Sample Text C3: 960721

Words	*Th*	*Rh*	*Overall*
Creatures	**11**	**13**	**24**
Software	**4**	**1**	**5**
Computer	**3**	**5**	**8**

Words	*Th*	*Rh*	*Overall*
Cliff	**3**	**2**	**5**
Tierra	**3**	**1**	**4**
Net	**2**	**4**	**6**

Words	*Th*	*Rh*	*Overall*
Evolution	**2**	**3**	**5**
Currently	**2**	**1**	**3**
Genes	**2**	**1**	**3**
Memory	**2**	**1**	**3**
Species	**2**	**1**	**3**
Switched	**2**	**1**	**2**
Tierrans	**2**	**1**	**3**
Internet	1	3	4
Time	1	3	4
Breeding	1	2	3
Don't	1	2	3
Eat	1	2	3
Entertainment	1	2	3
Four	1	2	3
Host	1	2	3
Hours	1	2	3
Long	1	2	3
Neurons	1	2	3
Simpson	1	2	3
Use	1	2	3
Versions	1	2	3
Worked	1	2	3
Advanced	1	1	2
Available	1	1	2
Behaviour	1	1	2
Body	1	1	2
Brains	1	1	2
Creating	1	1	2
Designed	1	1	2
Done	1	1	2
Dream	1	1	2
Enough	1	1	2
Example	1	1	2
Home	1	1	2
Hormones	1	1	2
Idea	1	1	2
Lights	1	1	2
Monster	1	1	2
Program	1	1	2
Result	1	1	2
Technology	1	1	2
Toby	1	1	2
World	1	1	2
Millennium	***7***	***0***	***7***
Ray	***4***	***0***	***4***
Various	***3***	***0***	***3***
Biologist	***2***	***0***	***2***
Creature	***2***	***0***	***2***
Habitats	***2***	***0***	***2***
It's	***2***	***0***	***2***
Little	***2***	***0***	***2***
Pack	***2***	***0***	***2***
Pet	***2***	***0***	***2***
University	***2***	***0***	***2***
Life	*0*	*9*	*9*
Says	*0*	*6*	*6*
Complex	*0*	*4*	*4*
Learn	*0*	*4*	*4*
Neural	*0*	*4*	*4*
Artificial	*0*	*3*	*3*
Brain	*0*	*3*	*3*
Eggs	*0*	*3*	*3*
Offspring	*0*	*3*	*3*
Parents	*0*	*3*	*3*
Sex	*0*	*3*	*3*
Simple	*0*	*3*	*3*
Sophisticated	*0*	*3*	*3*
Better	*0*	*2*	*2*
Biologists	*0*	*2*	*2*
Breed	*0*	*2*	*2*
Cause	*0*	*2*	*2*
Creature's	*0*	*2*	*2*
Determined	*0*	*2*	*2*
Different	*0*	*2*	*2*
Evolve	*0*	*2*	*2*
Get	*0*	*2*	*2*
Hard	*0*	*2*	*2*
Hopes	*0*	*2*	*2*
Human	*0*	*2*	*2*
Individuals	*0*	*2*	*2*
Inherited	*0*	*2*	*2*
Know	*0*	*2*	*2*
Let	*0*	*2*	*2*
Million	*0*	*2*	*2*
Model	*0*	*2*	*2*
New	*0*	*2*	*2*
Power	*0*	*2*	*2*
Programme	*0*	*2*	*2*
Programmed	*0*	*2*	*2*
Rudimentary	*0*	*2*	*2*
Screens	*0*	*2*	*2*
Sell	*0*	*2*	*2*
Show	*0*	*2*	*2*
Speak	*0*	*2*	*2*
Start	*0*	*2*	*2*
Traffic	*0*	*2*	*2*
Traits	*0*	*2*	*2*
Types	*0*	*2*	*2*
Unpredictable	*0*	*2*	*2*
Used	*0*	*2*	*2*
Years	*0*	*2*	*2*

Sample Text C4: 960728

Words	*Th*	*Rh*	*Overall*
Large	**13**	**11**	**24**
Blue	**12**	**8**	**20**
Caterpillars	**10**	**5**	**15**
Butterfly	**8**	**7**	**15**
Ants	**7**	**10**	**17**
Thomas	**7**	**2**	**9**
Ant	**5**	**8**	**13**
Colony	**5**	**4**	**9**
Red	**5**	**4**	**9**
Years	**5**	**2**	**7**
Rabbits	**4**	**1**	**5**
Species	**3**	**4**	**7**

Words	*Th*	*Rh*	*Overall*
Caterpillar	**3**	**2**	**5**
Land	**2**	**4**	**6**
Larvae	**2**	**2**	**4**
Nature	**2**	**2**	**4**
Queen	**2**	**2**	**4**
Smell	**2**	**2**	**4**
Thyme	**2**	**2**	**4**
Cent	**2**	**1**	**3**
Climate	**2**	**1**	**3**
Britain	1	7	8
Grubs	1	4	5
Enough	1	3	4
Numbers	1	3	4
Creatures	1	2	3
Established	1	2	3
Host	1	2	3
Aesop	1	1	2
Began	1	1	2
Behaviour	1	1	2
Believed	1	1	2
Bizarre	1	1	2
Cycle	1	1	2
Fertiliser	1	1	2
Final	1	1	2
Food	1	1	2
Grazing	1	1	2
Hot	1	1	2
Life	1	1	2
Longer	1	1	2
Moral	1	1	2
Order	1	1	2
Parasite	1	1	2
Reintroduce	1	1	2
Short	1	1	2
Site	1	1	2
South	1	1	2
Steep	1	1	2
Type	1	1	2
Virus	1	1	2
Grass	***4***	***0***	***4***
Conservationis+	***2***	***0***	***2***
Exacting	***2***	***0***	***2***
Ichneumon	***2***	***0***	***2***
People	***2***	***0***	***2***
Story	***2***	***0***	***2***
Summer	***2***	***0***	***2***
Butterflies	*0*	*5*	*5*
Nest	*0*	*4*	*4*

Words	*Th*	*Rh*	*Overall*
Extinct	*0*	*3*	*3*
Feed	*0*	*3*	*3*
Human	*0*	*3*	*3*
Mimic	*0*	*3*	*3*
Says	*0*	*3*	*3*
Sites	*0*	*3*	*3*
Survive	*0*	*3*	*3*
Using	*0*	*3*	*3*
Abundant	*0*	*2*	*2*
Beings	*0*	*2*	*2*
Bell	*0*	*2*	*2*
Bred	*0*	*2*	*2*
Die	*0*	*2*	*2*
Downfall	*0*	*2*	*2*
Fall	*0*	*2*	*2*
Foraging	*0*	*2*	*2*
Found	*0*	*2*	*2*
Fritillary	*0*	*2*	*2*
Grazed	*0*	*2*	*2*
Ground	*0*	*2*	*2*
Grow	*0*	*2*	*2*
Keep	*0*	*2*	*2*
Kill	*0*	*2*	*2*
Killed	*0*	*2*	*2*
Make	*0*	*2*	*2*
Meadow	*0*	*2*	*2*
Months	*0*	*2*	*2*
New	*0*	*2*	*2*
Perfect	*0*	*2*	*2*
Plant	*0*	*2*	*2*
Plants	*0*	*2*	*2*
Population	*0*	*2*	*2*
Proved	*0*	*2*	*2*
Re	*0*	*2*	*2*
Return	*0*	*2*	*2*
Small	*0*	*2*	*2*
Suitable	*0*	*2*	*2*
Survival	*0*	*2*	*2*
Swedish	*0*	*2*	*2*
Thin	*0*	*2*	*2*
Time	*0*	*2*	*2*
Together	*0*	*2*	*2*
Warm	*0*	*2*	*2*

Sample Text C5: 960804

Words	*Th*	*Rh*	*Overall*
Lovelock	**9**	**1**	**10**
Gaia	**6**	**5**	**11**
Granite	**6**	**3**	**9**
Water	**5**	**6**	**11**
Earth	**5**	**5**	**10**
Organisms	**4**	**8**	**12**
Biosphere	**4**	**4**	**8**
System	**3**	**5**	**8**
Oxygen	**3**	**3**	**6**

Words	*Th*	*Rh*	*Overall*
Ants	**3**	**2**	**5**
Global	**3**	**1**	**4**
Order	**3**	**1**	**4**
Life	**2**	**7**	**9**
Atmosphere	**2**	**6**	**8**
Gaian	**2**	**4**	**6**
New	**2**	**4**	**6**
Earth's	**2**	**3**	**5**
Model	**2**	**3**	**5**

Words	*Th*	*Rh*	*Overall*
Self	**2**	**3**	**5**
University	**2**	**3**	**5**
Atmospheric	**2**	**2**	**4**
World	**2**	**2**	**4**
Basalt	**2**	**1**	**3**
Bogs	**2**	**1**	**3**
Carnivore	**2**	**1**	**3**
Charlson	**2**	**1**	**3**
Clouds	**2**	**1**	**3**
Figures	**2**	**1**	**3**
Hypothesis	**2**	**1**	**3**
Models	**2**	**1**	**3**
Solar	**2**	**1**	**3**
Species	**2**	**1**	**3**
Superorganism	**2**	**1**	**3**
Says	1	12	13
Carbon	1	4	5
Climate	1	4	5
Formed	1	4	5
Long	1	4	5
Make	1	4	5
Regulating	1	4	5
Science	1	4	5
Things	1	4	5
Daisies	1	3	4
Greenhouse	1	3	4
Regulation	1	3	4
Vital	1	3	4
Adds	1	2	3
Argues	1	2	3
Behaviour	1	2	3
Coccolithospor+	1	2	3
Daisyworld	1	2	3
Deep	1	2	3
Droplets	1	2	3
Emerges	1	2	3
Environment	1	2	3
Feedbacks	1	2	3
Holland	1	2	3
Hotter	1	2	3
Kinds	1	2	3
Klinger	1	2	3
Living	1	2	3
Look	1	2	3
Marine	1	2	3
Ocean	1	2	3
Oceans	1	2	3
Petford	1	2	3
Phosphate	1	2	3
Wrong	1	2	3
Years	1	2	3
Areas	1	1	2
Average	1	1	2
Believes	1	1	2
Better	1	1	2
Concentration	1	1	2
Concept	1	1	2
Crust	1	1	2
Daisy	1	1	2
Doing	1	1	2
Evolved	1	1	2
Ferric	1	1	2
Forest	1	1	2

Words	*Th*	*Rh*	*Overall*
Gases	1	1	2
Geophysiology	1	1	2
Geosphere	1	1	2
Goodwin	1	1	2
Greater	1	1	2
Happy	1	1	2
Hard	1	1	2
Harding	1	1	2
Herbivore	1	1	2
Hydrosphere	1	1	2
Hydroxide	1	1	2
Individual	1	1	2
Many	1	1	2
Organism	1	1	2
Powerful	1	1	2
Put	1	1	2
Rain	1	1	2
Records	1	1	2
Regional	1	1	2
Respectable	1	1	2
Scientist	1	1	2
Sea	1	1	2
State	1	1	2
Study	1	1	2
Tiny	1	1	2
Valdes	1	1	2
Version	1	1	2
View	1	1	2
Warming	1	1	2
Example	***6***	***0***	***6***
Clearly	***2***	***0***	***2***
College	***2***	***0***	***2***
Continents	***2***	***0***	***2***
Jim	***2***	***0***	***2***
Multi	***2***	***0***	***2***
Research	***2***	***0***	***2***
Scientists	***2***	***0***	***2***
Suggestion	***2***	***0***	***2***
Temperature	*0*	*5*	*5*
Means	*0*	*4*	*4*
Need	*0*	*4*	*4*
Scientific	*0*	*4*	*4*
See	*0*	*4*	*4*
Term	*0*	*4*	*4*
Complex	*0*	*3*	*3*
Found	*0*	*3*	*3*
Free	*0*	*3*	*3*
Gas	*0*	*3*	*3*
Heat	*0*	*3*	*3*
High	*0*	*3*	*3*
Interactions	*0*	*3*	*3*
Planet	*0*	*3*	*3*
Reflect	*0*	*3*	*3*
Regulate	*0*	*3*	*3*
Sediments	*0*	*3*	*3*
Stability	*0*	*3*	*3*
Sunshine	*0*	*3*	*3*
Age	*0*	*2*	*2*
Animals	*0*	*2*	*2*
Biotic	*0*	*2*	*2*
Change	*0*	*2*	*2*
Circulation	*0*	*2*	*2*
Cloud	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Comes	*0*	*2*	*2*
Contains	*0*	*2*	*2*
Conventional	*0*	*2*	*2*
Covered	*0*	*2*	*2*
Derived	*0*	*2*	*2*
Distributed	*0*	*2*	*2*
Essential	*0*	*2*	*2*
Feedback	*0*	*2*	*2*
Finding	*0*	*2*	*2*
Forests	*0*	*2*	*2*
Formation	*0*	*2*	*2*
Gcms	*0*	*2*	*2*
Geologist	*0*	*2*	*2*
Get	*0*	*2*	*2*
Going	*0*	*2*	*2*
Ice	*0*	*2*	*2*
Important	*0*	*2*	*2*
Know	*0*	*2*	*2*
Lay	*0*	*2*	*2*
Leading	*0*	*2*	*2*
Limited	*0*	*2*	*2*
Maintain	*0*	*2*	*2*
Past	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Peat	*0*	*2*	*2*
Photosynthesis	*0*	*2*	*2*
Point	*0*	*2*	*2*
Professor	*0*	*2*	*2*
Range	*0*	*2*	*2*
Regions	*0*	*2*	*2*
Regulated	*0*	*2*	*2*
Regulatory	*0*	*2*	*2*
Religious	*0*	*2*	*2*
Result	*0*	*2*	*2*
Right	*0*	*2*	*2*
Rocks	*0*	*2*	*2*
Space	*0*	*2*	*2*
Standard	*0*	*2*	*2*
Strong	*0*	*2*	*2*
Super	*0*	*2*	*2*
Test	*0*	*2*	*2*
Tightly	*0*	*2*	*2*
Together	*0*	*2*	*2*
Way	*0*	*2*	*2*

Sample Text C6: 960811

Words	*Th*	*Rh*	*Overall*
Galileo	**13**	**1**	**14**
Probe	**13**	**1**	**14**
Jupiter	**11**	**10**	**21**
Planet	**6**	**7**	**13**
Time	**6**	**3**	**9**
Pictures	**6**	**1**	**7**
Ganymede	**5**	**1**	**6**
Voyager	**5**	**1**	**6**
Years	**4**	**7**	**11**
Scientists	**4**	**4**	**8**
Atmosphere	**4**	**3**	**7**
Space	**4**	**3**	**7**
Launch	**4**	**1**	**5**
Earth	**3**	**23**	**26**
Io	**3**	**4**	**7**
Million	**3**	**3**	**6**
Left	**3**	**2**	**5**
Months	**3**	**1**	**4**
Made	**2**	**5**	**7**
Surface	**2**	**5**	**7**
Asteroid	**2**	**3**	**5**
Gravitational	**2**	**3**	**5**
Hydrogen	**2**	**3**	**5**
Jovian	**2**	**3**	**5**
Jupiter's	**2**	**3**	**5**
Main	**2**	**3**	**5**
Make	**2**	**3**	**5**
Antenna	**2**	**2**	**4**
Core	**2**	**2**	**4**
Nuclear	**2**	**2**	**4**
Same	**2**	**2**	**4**
Sulphur	**2**	**2**	**4**
Craft	**2**	**1**	**3**

Words	*Th*	*Rh*	*Overall*
Data	**2**	**1**	**3**
Fact	**2**	**1**	**3**
Impacts	**2**	**1**	**3**
Released	**2**	**1**	**3**
Squyres	**2**	**1**	**3**
Steve	**2**	**1**	**3**
Moon	1	8	9
Many	1	7	8
Solar	1	6	7
Clouds	1	5	6
Sun	1	5	6
Times	1	5	6
Gas	1	4	5
Cloud	1	3	4
Field	1	3	4
Massive	1	3	4
Moons	1	3	4
Venus	1	3	4
Atmospheric	1	2	3
Bombardment	1	2	3
Clear	1	2	3
Cold	1	2	3
Coming	1	2	3
Cracks	1	2	3
Deep	1	2	3
Form	1	2	3
Formed	1	2	3
Galileo's	1	2	3
Giant	1	2	3
Heat	1	2	3
Material	1	2	3
Passing	1	2	3
Probably	1	2	3
Rocky	1	2	3
Seems	1	2	3

Words	Th	Rh	Overall
Tape	1	2	3
Ammonia	1	1	2
Big	1	1	2
Called	1	1	2
Comet	1	1	2
Composition	1	1	2
Compounds	1	1	2
Craters	1	1	2
Destroyed	1	1	2
Dinosaur	1	1	2
Direction	1	1	2
Energy	1	1	2
Enough	1	1	2
Exposed	1	1	2
Fault	1	1	2
Future	1	1	2
Gravity	1	1	2
Great	1	1	2
Helium	1	1	2
Hour	1	1	2
Metal	1	1	2
Orange	1	1	2
Pressure	1	1	2
Previously	1	1	2
Probes	1	1	2
Protection	1	1	2
Recorder	1	1	2
Route	1	1	2
Say	1	1	2
Shine	1	1	2
Size	1	1	2
Started	1	1	2
Suggest	1	1	2
Tops	1	1	2
Torn	1	1	2
Typical	1	1	2
Visible	1	1	2
Work	1	1	2
Yellow	1	1	2
Young	1	1	2
Astronomers	***3***	***0***	***3***
Evidence	***3***	***0***	***3***
Mission	***3***	***0***	***3***
Began	***2***	***0***	***2***
Biggest	***2***	***0***	***2***
Challenger	***2***	***0***	***2***
December	***2***	***0***	***2***
Degree	***2***	***0***	***2***
Entered	***2***	***0***	***2***
Eros	***2***	***0***	***2***
History	***2***	***0***	***2***
Interior	***2***	***0***	***2***
Latest	***2***	***0***	***2***
Measurements	***2***	***0***	***2***
Nasa's	***2***	***0***	***2***
Protective	***2***	***0***	***2***
Sensors	***2***	***0***	***2***
Shuttle	***2***	***0***	***2***
Site	***2***	***0***	***2***
Smooth	***2***	***0***	***2***
Temperature	***2***	***0***	***2***
System	*0*	*8*	*8*
Orbit	*0*	*6*	*6*

Words	Th	Rh	Overall
Way	*0*	*6*	*6*
Ice	*0*	*5*	*5*
Forces	*0*	*4*	*4*
Largest	*0*	*4*	*4*
Life	*0*	*4*	*4*
Show	*0*	*4*	*4*
Water	*0*	*4*	*4*
Comets	*0*	*3*	*3*
Earth's	*0*	*3*	*3*
Europa	*0*	*3*	*3*
Known	*0*	*3*	*3*
Layers	*0*	*3*	*3*
Liquid	*0*	*3*	*3*
Network	*0*	*3*	*3*
Oceans	*0*	*3*	*3*
Passed	*0*	*3*	*3*
Past	*0*	*3*	*3*
Planets	*0*	*3*	*3*
Radiation	*0*	*3*	*3*
Revealed	*0*	*3*	*3*
Rock	*0*	*3*	*3*
Sea	*0*	*3*	*3*
Similar	*0*	*3*	*3*
Slingshot	*0*	*3*	*3*
Sufficiently	*0*	*3*	*3*
Tidal	*0*	*3*	*3*
Volcanic	*0*	*3*	*3*
White	*0*	*3*	*3*
Active	*0*	*2*	*2*
Asteroids	*0*	*2*	*2*
Believed	*0*	*2*	*2*
Belts	*0*	*2*	*2*
Black	*0*	*2*	*2*
Caused	*0*	*2*	*2*
Churned	*0*	*2*	*2*
Closer	*0*	*2*	*2*
Come	*0*	*2*	*2*
Comparatively	*0*	*2*	*2*
Cosmic	*0*	*2*	*2*
Covered	*0*	*2*	*2*
Descended	*0*	*2*	*2*
Directly	*0*	*2*	*2*
Discovered	*0*	*2*	*2*
Elliptical	*0*	*2*	*2*
Four	*0*	*2*	*2*
Hazard	*0*	*2*	*2*
Hope	*0*	*2*	*2*
Hot	*0*	*2*	*2*
Including	*0*	*2*	*2*
Know	*0*	*2*	*2*
Long	*0*	*2*	*2*
Look	*0*	*2*	*2*
Magnetic	*0*	*2*	*2*
Making	*0*	*2*	*2*
Marked	*0*	*2*	*2*
Mars	*0*	*2*	*2*
Means	*0*	*2*	*2*
Meant	*0*	*2*	*2*
Melted	*0*	*2*	*2*
Mercury	*0*	*2*	*2*
Millions	*0*	*2*	*2*
Object	*0*	*2*	*2*
Orbits	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Perhaps	*0*	*2*	*2*
Place	*0*	*2*	*2*
Pock	*0*	*2*	*2*
Pose	*0*	*2*	*2*
Possible	*0*	*2*	*2*
Powerful	*0*	*2*	*2*
Precise	*0*	*2*	*2*
Reached	*0*	*2*	*2*
Says	*0*	*2*	*2*
Seem	*0*	*2*	*2*
Smaller	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Speed	*0*	*2*	*2*
Star	*0*	*2*	*2*
Sunlight	*0*	*2*	*2*
Telescope	*0*	*2*	*2*
Thick	*0*	*2*	*2*
Together	*0*	*2*	*2*
Turned	*0*	*2*	*2*
Umbrella	*0*	*2*	*2*
Vapour	*0*	*2*	*2*
Volcanically	*0*	*2*	*2*

Sample Text C7: 960818

Words	*Th*	*Rh*	*Overall*
Science	**10**	**4**	**14**
Century	**6**	**6**	**13**
Dutch	**5**	**5**	**9**
Natural	**3**	**2**	**5**
Genetics	**3**	**1**	**4**
Philosophy	**3**	**1**	**4**
Modern	**2**	**3**	**5**
TRUE	**2**	**2**	**4**
Dulwich	**2**	**1**	**3**
Exhibition	**2**	**1**	**3**
Mendel	**2**	**1**	**3**
Painting	**2**	**1**	**3**
Words	**2**	**1**	**3**
New	1	3	4
Art	1	2	3
Horticulturali+	1	2	3
Latin	1	2	3
Nature	1	2	3
Painted	1	2	3
Understood	1	2	3
Centuries	1	1	2
Established	1	1	2
Fortunes	1	1	2
Giants	1	1	2
Intriguing	1	1	2
Laws	1	1	2
Life	1	1	2
Long	1	1	2
Perception	1	1	2
Saw	1	1	2
Short	1	1	2
Species	1	1	2
Stripes	1	1	2
Technology	1	1	2
Time	1	1	2
Together	1	1	2
Virus	1	1	2

Words	*Th*	*Rh*	*Overall*
Viruses	1	1	2
Well	1	1	2
Work	1	1	2
Tulips	***5***	***0***	***5***
Breeding	***3***	***0***	***3***
Newton	***3***	***0***	***3***
Plant	***3***	***0***	***3***
Age	***2***	***0***	***2***
Breughel	***2***	***0***	***2***
Breughel's	***2***	***0***	***2***
Decades	***2***	***0***	***2***
General	***2***	***0***	***2***
Little	***2***	***0***	***2***
Mendel's	***2***	***0***	***2***
Ray	***2***	***0***	***2***
World	*0*	*3*	*3*
Allowed	*0*	*2*	*2*
Began	*0*	*2*	*2*
Benefit	*0*	*2*	*2*
Bulbs	*0*	*2*	*2*
Clone	*0*	*2*	*2*
Deeply	*0*	*2*	*2*
Genes	*0*	*2*	*2*
God	*0*	*2*	*2*
Hands	*0*	*2*	*2*
Host	*0*	*2*	*2*
Ideas	*0*	*2*	*2*
Insects	*0*	*2*	*2*
Lost	*0*	*2*	*2*
Peas	*0*	*2*	*2*
See	*0*	*2*	*2*
Serious	*0*	*2*	*2*
Ways	*0*	*2*	*2*
Works	*0*	*2*	*2*
Later	*0*	*1*	*2*
Linnaeus	*0*	*1*	*2*
Years	*0*	*1*	*2*

Sample Text C8: 960825

Words	*Th*	*Rh*	*Overall*
Body	**3**	**8**	**11**
X	**3**	**1**	**4**
Images	**2**	**4**	**6**

Words	*Th*	*Rh*	*Overall*
Computer	**2**	**2**	**4**
Patient's	**2**	**2**	**4**
Imaging	**2**	**1**	**3**

Words	*Th*	*Rh*	*Overall*
Rays	**2**	**1**	**3**
Scans	**2**	**1**	**3**
Voyage	**2**	**1**	**3**
Anatomists	1	2	3
Field	1	2	3
Layer	1	2	3
Physicians	1	2	3
Renaissance	1	2	3
Technicians	1	2	3
Artists	1	1	2
Blood	1	1	2
Cadavers	1	1	2
Cruise	1	1	2
Dissection	1	1	2
Fantastic	1	1	2
Human	1	1	2
Magnetic	1	1	2
Magnify	1	1	2
Powerful	1	1	2
Raquel	1	1	2
Real	1	1	2
Science	1	1	2
Skills	1	1	2
Sperm	1	1	2
Text	1	1	2
Time	1	1	2

Words	*Th*	*Rh*	*Overall*
Welch	1	1	2
Workings	1	1	2
Cat	***3***	***0***	***3***
Imagery	***3***	***0***	***3***
Medical	***3***	***0***	***3***
Students	***3***	***0***	***3***
Times	***3***	***0***	***3***
Conventional	***2***	***0***	***2***
New	***2***	***0***	***2***
Scan	***2***	***0***	***2***
Technology	***2***	***0***	***2***
White	***2***	***0***	***2***
Years	***2***	***0***	***2***
Organs	*0*	*3*	*3*
See	*0*	*3*	*3*
Dreams	*0*	*2*	*2*
Free	*0*	*2*	*2*
Gives	*0*	*2*	*2*
Image	*0*	*2*	*2*
Looks	*0*	*2*	*2*
Patient	*0*	*2*	*2*
Photography	*0*	*2*	*2*
Skin	*0*	*2*	*2*
Surgeons	*0*	*2*	*2*
Vesalius	*0*	*2*	*2*
Walk	*0*	*2*	*2*

Sample Text C9: 960901

Words	*Th*	*Rh*	*Overall*
Fault	**6**	**8**	**14**
Hierapolis	**6**	**3**	**9**
Professor	**6**	**2**	**8**
Plane	**5**	**4**	**9**
Hancock	**5**	**1**	**6**
Earthquake	**4**	**11**	**15**
Quake	**4**	**1**	**5**
Earthquakes	**3**	**3**	**6**
Earth's	**3**	**2**	**5**
Town	**2**	**5**	**7**
Quakes	**2**	**2**	**4**
Surface	**2**	**2**	**4**
Water	**2**	**2**	**4**
Artifacts	**2**	**1**	**3**
Buildings	**2**	**1**	**3**
Energy	**2**	**1**	**3**
Scientists	**2**	**1**	**3**
Roman	1	4	5
Century	1	3	4
Ambraseys	1	2	3
Lies	1	2	3
Passing	1	2	3
Released	1	2	3
Scarp	1	2	3
Times	1	2	3
Travertine	1	2	3
Columns	1	1	2
Cracked	1	1	2
Destruction	1	1	2
Displaced	1	1	2
Fallen	1	1	2

Words	*Th*	*Rh*	*Overall*
Intensity	1	1	2
Ix	1	1	2
Made	1	1	2
Planes	1	1	2
Predict	1	1	2
Records	1	1	2
Relics	1	1	2
Runs	1	1	2
Sand	1	1	2
Sarcophagus	1	1	2
Scale	1	1	2
Sites	1	1	2
Take	1	1	2
Technique	1	1	2
Time	1	1	2
Turkish	1	1	2
Channel	***3***	***0***	***3***
Local	***3***	***0***	***3***
Altunel	***2***	***0***	***2***
Damage	***2***	***0***	***2***
Geologists	***2***	***0***	***2***
Jericho	***2***	***0***	***2***
Layer	***2***	***0***	***2***
Pollen	***2***	***0***	***2***
Probably	***2***	***0***	***2***
Years	*0*	*4*	*4*
Ancient	*0*	*3*	*3*
Built	*0*	*3*	*3*
Cause	*0*	*3*	*3*
Caused	*0*	*3*	*3*
Fort	*0*	*3*	*3*
Indicate	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Known	*0*	*3*	*3*
Possible	*0*	*3*	*3*
Area	*0*	*2*	*2*
Clues	*0*	*2*	*2*
Damaged	*0*	*2*	*2*
Date	*0*	*2*	*2*
Directly	*0*	*2*	*2*
Epicentre	*0*	*2*	*2*
Extend	*0*	*2*	*2*
Ground	*0*	*2*	*2*
Happened	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Historical	*0*	*2*	*2*
History	*0*	*2*	*2*
Hit	*0*	*2*	*2*
Jagged	*0*	*2*	*2*
Limestone	*0*	*2*	*2*
Passes	*0*	*2*	*2*
Past	*0*	*2*	*2*
Place	*0*	*2*	*2*
Record	*0*	*2*	*2*
Reveal	*0*	*2*	*2*
Says	*0*	*2*	*2*

Sample Text C10: 960908

Words	*Th*	*Rh*	*Overall*
Hubble	**3**	**4**	**7**
Pictures	**3**	**3**	**6**
Universe	**2**	**5**	**7**
Stars	**2**	**2**	**4**
Earth	**2**	**1**	**3**
Ho	**2**	**1**	**3**
Telescope	**2**	**1**	**3**
Light	1	6	7
Ground	1	5	6
Based	1	3	4
Correct	1	3	4
Years	1	3	4
Constant	1	2	3
Kilometres	1	2	3
Problems	1	2	3
Atmosphere	1	1	2
December	1	1	2
Distances	1	1	2
Enough	1	1	2
Example	1	1	2
Field	1	1	2
Hubble's	1	1	2
Inspiring	1	1	2
Modern	1	1	2
Precision	1	1	2
Pyramids	1	1	2
Technology	1	1	2
Testament	1	1	2
Hst	***9***	***0***	***9***
Astronomers	***2***	***0***	***2***
D	***2***	***0***	***2***
Lights	***2***	***0***	***2***
Scientists	***2***	***0***	***2***

Words	*Th*	*Rh*	*Overall*
Street	***2***	***0***	***2***
V	***2***	***0***	***2***
Work	*0*	*5*	*5*
Space	*0*	*4*	*4*
Billion	*0*	*3*	*3*
Human	*0*	*3*	*3*
Lets	*0*	*3*	*3*
See	*0*	*3*	*3*
Atoms	*0*	*2*	*2*
Better	*0*	*2*	*2*
Electron	*0*	*2*	*2*
Gaze	*0*	*2*	*2*
Genome	*0*	*2*	*2*
Make	*0*	*2*	*2*
Metre	*0*	*2*	*2*
Microscope	*0*	*2*	*2*
Mirror	*0*	*2*	*2*
Observation	*0*	*2*	*2*
Observe	*0*	*2*	*2*
Observed	*0*	*2*	*2*
Possible	*0*	*2*	*2*
Project	*0*	*2*	*2*
Puts	*0*	*2*	*2*
Same	*0*	*2*	*2*
Scanning	*0*	*2*	*2*
Sensitive	*0*	*2*	*2*
Signals	*0*	*2*	*2*
Sky	*0*	*2*	*2*
Take	*0*	*2*	*2*
Value	*0*	*2*	*2*
World	*0*	*2*	*2*

Sample Text C11: 960915

Words	*Th*	*Rh*	*Overall*
Heart	**10**	**19**	**29**
People	**7**	**5**	**12**
Disease	**5**	**11**	**16**
Attack	**5**	**3**	**8**
C	**4**	**10**	**14**
Gupta	**4**	**1**	**5**
Pneumoniae	**3**	**10**	**13**

Words	*Th*	*Rh*	*Overall*
Arteries	**3**	**7**	**10**
Bug	**3**	**6**	**9**
Antibiotics	**3**	**5**	**8**
Factors	**3**	**5**	**8**
Infection	**3**	**5**	**8**
Plaque	**3**	**4**	**7**
High	**3**	**3**	**6**

Words	Th	Rh	Overall
Research	3	2	5
Patients	3	1	4
Blood	2	12	14
System	2	8	10
Immune	2	5	7
Key	2	4	6
Cholesterol	2	3	5
Pressure	2	3	5
Way	2	3	5
Caused	2	2	4
Pylori	2	2	4
Risk	2	2	4
Time	2	2	4
Doctors	2	1	3
Find	2	1	3
H	2	1	3
Population	2	1	3
Says	2	1	3
Theory	2	1	3
Attacks	1	3	4
Cause	1	3	4
Clotting	1	3	4
Course	1	3	4
Get	1	3	4
Atherosclerosi+	1	2	3
Causes	1	2	3
Chlamydia	1	2	3
Day	1	2	3
Gene	1	2	3
Link	1	2	3
Living	1	2	3
Macrophages	1	2	3
Proteins	1	2	3
Scarring	1	2	3
Trigger	1	2	3
Antibodies	1	1	2
Carried	1	1	2
Chronic	1	1	2
Damaged	1	1	2
Died	1	1	2
Easily	1	1	2
Exercise	1	1	2
Fat	1	1	2
Fats	1	1	2
Fatty	1	1	2
Having	1	1	2
Home	1	1	2
New	1	1	2
Production	1	1	2
Red	1	1	2
Revealed	1	1	2
Sludge	1	1	2
Suggests	1	1	2
Surface	1	1	2
Theories	1	1	2
Thought	1	1	2
Wine	1	1	2
Researchers	***5***	***0***	***5***
Britain	***3***	***0***	***3***
Exactly	***3***	***0***	***3***
Go	***3***	***0***	***3***
Scientists	***3***	***0***	***3***
St	***3***	***0***	***3***

Words	Th	Rh	Overall
Bacteria	***2***	***0***	***2***
Clogged	***2***	***0***	***2***
Days	***2***	***0***	***2***
Example	***2***	***0***	***2***
Foods	***2***	***0***	***2***
George's	***2***	***0***	***2***
Hospital	***2***	***0***	***2***
It's	***2***	***0***	***2***
Long	***2***	***0***	***2***
Low	***2***	***0***	***2***
Recently	***2***	***0***	***2***
States	***2***	***0***	***2***
Studies	***2***	***0***	***2***
Studying	***2***	***0***	***2***
Thinking	***2***	***0***	***2***
University	***2***	***0***	***2***
Usually	***2***	***0***	***2***
Year	***2***	***0***	***2***
Artery	*0*	*8*	*8*
Found	*0*	*5*	*5*
Healthy	*0*	*4*	*4*
Prevent	*0*	*4*	*4*
Smoking	*0*	*4*	*4*
Walls	*0*	*4*	*4*
Coronary	*0*	*3*	*3*
Respiratory	*0*	*3*	*3*
Wall	*0*	*3*	*3*
Abnormal	*0*	*2*	*2*
Antibiotic	*0*	*2*	*2*
Bacterium	*0*	*2*	*2*
Breakthrough	*0*	*2*	*2*
Cells	*0*	*2*	*2*
Class	*0*	*2*	*2*
Common	*0*	*2*	*2*
Damage	*0*	*2*	*2*
Diet	*0*	*2*	*2*
Doing	*0*	*2*	*2*
Factor	*0*	*2*	*2*
Finding	*0*	*2*	*2*
Findings	*0*	*2*	*2*
Flow	*0*	*2*	*2*
Happens	*0*	*2*	*2*
Implicated	*0*	*2*	*2*
Infected	*0*	*2*	*2*
Infections	*0*	*2*	*2*
Inflammation	*0*	*2*	*2*
Know	*0*	*2*	*2*
Levels	*0*	*2*	*2*
Organism	*0*	*2*	*2*
Parts	*0*	*2*	*2*
Pathogen	*0*	*2*	*2*
Question	*0*	*2*	*2*
Questions	*0*	*2*	*2*
Raise	*0*	*2*	*2*
Rid	*0*	*2*	*2*
Search	*0*	*2*	*2*
Sections	*0*	*2*	*2*
Signal	*0*	*2*	*2*
Therapy	*0*	*2*	*2*
Tissue	*0*	*2*	*2*
Tubes	*0*	*2*	*2*
World	*0*	*2*	*2*

Sample Text C12: 960922

Words	*Th*	*Rh*	*Overall*
Bracken	**25**	**6**	**31**
Spores	**8**	**5**	**13**
Water	**4**	**5**	**9**
Autumn	**4**	**3**	**7**
Year	**4**	**3**	**7**
Animals	**4**	**1**	**5**
Bac	**3**	**1**	**4**
Numbers	**3**	**1**	**4**
Walking	**3**	**1**	**4**
Areas	**2**	**4**	**6**
Health	**2**	**3**	**5**
Disease	**2**	**2**	**4**
Early	**2**	**2**	**4**
Parts	**2**	**2**	**4**
Plants	**2**	**2**	**4**
Carcinogens	**2**	**1**	**3**
Cattle	**2**	**1**	**3**
Filters	**2**	**1**	**3**
Getting	**2**	**1**	**3**
Iii	**2**	**1**	**3**
It's	**2**	**1**	**3**
Louping	**2**	**1**	**3**
Lyme	**2**	**1**	**3**
Problem	**2**	**1**	**3**
Sheep	**2**	**1**	**3**
Used	**2**	**1**	**3**
Years	**2**	**1**	**3**
Cancers	1	4	5
Land	1	3	4
Many	1	3	4
Summer	1	3	4
Uplands	1	3	4
Britain	1	2	3
Brown	1	2	3
Causing	1	2	3
Cost	1	2	3
Digestive	1	2	3
Directive	1	2	3
Drinking	1	2	3
Professor	1	2	3
Quickly	1	2	3
Risk	1	2	3
Same	1	2	3
Supplies	1	2	3
Tract	1	2	3
Asulox	1	1	2
Attractive	1	1	2
Carcinogenic	1	1	2
Catchments	1	1	2
Common	1	1	2
Crathorne	1	1	2
Done	1	1	2
Eat	1	1	2
Eaten	1	1	2
England	1	1	2
Especially	1	1	2
Extensive	1	1	2
Fern	1	1	2

Words	*Th*	*Rh*	*Overall*
Harbours	1	1	2
Healthy	1	1	2
Herbicide	1	1	2
Hills	1	1	2
Infection	1	1	2
Moors	1	1	2
Ptaquilosides	1	1	2
Says	1	1	2
Sporing	1	1	2
Spread	1	1	2
Supply	1	1	2
Toxic	1	1	2
Vast	1	1	2
Viral	1	1	2
Walkers	1	1	2
Warm	1	1	2
World's	1	1	2
Grants	***4***	***0***	***4***
Large	***4***	***0***	***4***
Plant	***4***	***0***	***4***
Dairy	***3***	***0***	***3***
Farmers	***3***	***0***	***3***
Time	***3***	***0***	***3***
British	***2***	***0***	***2***
Centre	***2***	***0***	***2***
Commission	***2***	***0***	***2***
Countries	***2***	***0***	***2***
Cutting	***2***	***0***	***2***
Dense	***2***	***0***	***2***
Heyworth	***2***	***0***	***2***
Livestock	***2***	***0***	***2***
Isn't	*0*	*4*	*4*
Reservoirs	*0*	*4*	*4*
Spreading	*0*	*4*	*4*
Adds	*0*	*3*	*3*
Available	*0*	*3*	*3*
Don't	*0*	*3*	*3*
Humans	*0*	*3*	*3*
Milk	*0*	*3*	*3*
Arthritis	*0*	*2*	*2*
Bad	*0*	*2*	*2*
Believes	*0*	*2*	*2*
Bite	*0*	*2*	*2*
Caused	*0*	*2*	*2*
Cent	*0*	*2*	*2*
Concerned	*0*	*2*	*2*
Contain	*0*	*2*	*2*
Difficult	*0*	*2*	*2*
Diseases	*0*	*2*	*2*
Green	*0*	*2*	*2*
Hazard	*0*	*2*	*2*
Heavy	*0*	*2*	*2*
Including	*0*	*2*	*2*
Lowland	*0*	*2*	*2*
North	*0*	*2*	*2*
Particularly	*0*	*2*	*2*
Pastures	*0*	*2*	*2*
Persist	*0*	*2*	*2*
Places	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Produces	*0*	*2*	*2*
Range	*0*	*2*	*2*
Rash	*0*	*2*	*2*
Remove	*0*	*2*	*2*
Sprayed	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Tick	*0*	*2*	*2*
Ticks	*0*	*2*	*2*
Wales	*0*	*2*	*2*
Washed	*0*	*2*	*2*
Widely	*0*	*2*	*2*

Sample Text C13: 960929

Words	*Th*	*Rh*	*Overall*
Taylor	**8**	**7**	**15**
Sexual	**6**	**5**	**11**
Years	**5**	**5**	**10**
Evidence	**4**	**5**	**9**
Female	**4**	**5**	**9**
Women	**4**	**4**	**8**
Sex	**3**	**9**	**12**
Age	**3**	**5**	**8**
Ice	**3**	**5**	**8**
Many	**3**	**3**	**6**
Million	**3**	**3**	**6**
Fact	**3**	**2**	**5**
Four	**3**	**2**	**5**
Species	**3**	**1**	**3**
Females	**2**	**4**	**5**
Men	**2**	**4**	**6**
Found	**2**	**3**	**5**
Plants	**2**	**3**	**5**
Today	**2**	**2**	**4**
Ancient	**2**	**1**	**3**
Culture	**2**	**1**	**3**
Human	**2**	**1**	**3**
Past	**2**	**1**	**3**
Penis	**2**	**1**	**3**
Male	1	6	7
Account	1	2	3
Ancestors	1	2	3
Greek	1	2	3
Hand	1	2	3
Hands	1	2	3
Large	1	2	3
Objects	1	2	3
Tail	1	2	3
Tied	1	2	3
Urine	1	2	3
Woman	1	2	3
Art	1	1	2
Attractive	1	1	2
Burial	1	1	2
Carvings	1	1	2
Cave	1	1	2
Comes	1	1	2
Devices	1	1	2
Example	1	1	2
Explicitly	1	1	2
Farming	1	1	2
Figures	1	1	2
Find	1	1	2
Herodotus	1	1	2
Images	1	1	2
Interpretation	1	1	2
Look	1	1	2

Words	*Th*	*Rh*	*Overall*
Major	1	1	2
Necessarily	1	1	2
Neolithic	1	1	2
Period	1	1	2
Plant	1	1	2
Pregnant	1	1	2
Prudishness	1	1	2
Pygmy	1	1	2
Recently	1	1	2
Rock	1	1	2
Sculptor	1	1	2
Selection	1	1	2
Sense	1	1	2
Shaped	1	1	2
Short	1	1	2
Straps	1	1	2
Today's	1	1	2
Together	1	1	2
Troop	1	1	2
Animals	***4***	***0***	***4***
Humans	***4***	***0***	***4***
Figurine	***3***	***0***	***3***
Graves	***3***	***0***	***3***
Hormones	***3***	***0***	***3***
Beaker	***2***	***0***	***2***
Black	***2***	***0***	***2***
Century	***2***	***0***	***2***
Chimpanzees	***2***	***0***	***2***
Depict	***2***	***0***	***2***
Eden	***2***	***0***	***2***
Figurines	***2***	***0***	***2***
Fur	***2***	***0***	***2***
Gender	***2***	***0***	***2***
Oestrogen	***2***	***0***	***2***
Posture	***2***	***0***	***2***
Primates	***2***	***0***	***2***
Scientists	***2***	***0***	***2***
Sculptures	***2***	***0***	***2***
Sea	***2***	***0***	***2***
Time	***2***	***0***	***2***
Venus	***2***	***0***	***2***
Says	*0*	*7*	*7*
Reproduction	*0*	*4*	*4*
Suggest	*0*	*4*	*4*
Argues	*0*	*3*	*3*
Believes	*0*	*3*	*3*
Breasts	*0*	*3*	*3*
Buried	*0*	*3*	*3*
Carved	*0*	*3*	*3*
Induce	*0*	*3*	*3*
Little	*0*	*3*	*3*
Made	*0*	*3*	*3*
Seems	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Actually	*0*	*2*	*2*
Archaeological	*0*	*2*	*2*
Archaeologists	*0*	*2*	*2*
Argue	*0*	*2*	*2*
Began	*0*	*2*	*2*
Best	*0*	*2*	*2*
Chimpanzee	*0*	*2*	*2*
Clearly	*0*	*2*	*2*
Clothes	*0*	*2*	*2*
Complete	*0*	*2*	*2*
Considered	*0*	*2*	*2*
Contain	*0*	*2*	*2*
Contraceptives	*0*	*2*	*2*
Copper	*0*	*2*	*2*
Drug	*0*	*2*	*2*
Facing	*0*	*2*	*2*
Fall	*0*	*2*	*2*
Function	*0*	*2*	*2*
Functional	*0*	*2*	*2*
Hard	*0*	*2*	*2*
Having	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Larger	*0*	*2*	*2*
Longer	*0*	*2*	*2*
Lost	*0*	*2*	*2*
Make	*0*	*2*	*2*
Man	*0*	*2*	*2*
Mixed	*0*	*2*	*2*
Modern	*0*	*2*	*2*
Nature	*0*	*2*	*2*
Obvious	*0*	*2*	*2*
People	*0*	*2*	*2*
Place	*0*	*2*	*2*
Pottery	*0*	*2*	*2*
Reason	*0*	*2*	*2*
Recognised	*0*	*2*	*2*
Ritual	*0*	*2*	*2*
Skeleton	*0*	*2*	*2*
Strong	*0*	*2*	*2*
Tails	*0*	*2*	*2*
Took	*0*	*2*	*2*
Usually	*0*	*2*	*2*
Way	*0*	*2*	*2*

Sample Text C14: 961006

Words	*Th*	*Rh*	*Overall*
Mithen	**16**	**1**	**17**
Mind	**13**	**10**	**23**
Modern	**5**	**5**	**10**
Evolutionary	**5**	**4**	**9**
Human	**5**	**2**	**7**
Hand	**4**	**7**	**11**
Tools	**4**	**7**	**11**
Cognitive	**4**	**4**	**8**
Social	**4**	**4**	**8**
Modules	**4**	**2**	**6**
Axes	**4**	**1**	**5**
Early	**4**	**1**	**5**
Humans	**4**	**1**	**5**
Intelligence	**3**	**5**	**8**
Domain	**3**	**3**	**6**
General	**3**	**2**	**5**
Mithen's	**3**	**2**	**5**
Thought	**3**	**1**	**4**
Years	**2**	**6**	**8**
Making	**2**	**3**	**5**
Stone	**2**	**3**	**5**
Archaeology	**2**	**2**	**4**
Record	**2**	**2**	**4**
Specialised	**2**	**2**	**4**
Archaeologist	**2**	**1**	**3**
Behaviour	**2**	**1**	**3**
Language	**2**	**1**	**3**
Modularity	**2**	**1**	**3**
Psychologists	**2**	**1**	**3**
Reading	**2**	**1**	**3**
Theory	**2**	**1**	**3**
Vision	**2**	**1**	**3**
World	**2**	**1**	**3**
Minds	1	5	6
Cathedral	1	3	4

Words	*Th*	*Rh*	*Overall*
Evolution	1	3	4
Specific	1	3	4
Swiss	1	3	4
Account	1	2	3
Analogy	1	2	3
Archaeological	1	2	3
Army	1	2	3
Complex	1	2	3
Create	1	2	3
Fact	1	2	3
Form	1	2	3
Knife	1	2	3
Module	1	2	3
Natural	1	2	3
Points	1	2	3
Populations	1	2	3
Psychology	1	2	3
Tool	1	2	3
Ancient	1	1	2
Argument	1	1	2
Built	1	1	2
Certain	1	1	2
Culture	1	1	2
Era	1	1	2
Forms	1	1	2
Gatherers	1	1	2
Hunter	1	1	2
Imposed	1	1	2
Learn	1	1	2
Leda	1	1	2
Made	1	1	2
Neanderthals	1	1	2
New	1	1	2
People	1	1	2
Powerful	1	1	2
Prehistory	1	1	2
Similar	1	1	2

Words	*Th*	*Rh*	*Overall*
Sound	1	1	2
Space	1	1	2
Speed	1	1	2
Steven	1	1	2
Sudden	1	1	2
Try	1	1	2
Types	1	1	2
It's	***5***	***0***	***5***
Archaeologists	***2***	***0***	***2***
Barriers	***2***	***0***	***2***
Book	***2***	***0***	***2***
Characteristic	***2***	***0***	***2***
Children	***2***	***0***	***2***
Chimpanzee	***2***	***0***	***2***
Computer	***2***	***0***	***2***
Cosmides	***2***	***0***	***2***
Creativity	***2***	***0***	***2***
Discipline	***2***	***0***	***2***
Distinguishing	***2***	***0***	***2***
End	***2***	***0***	***2***
List	***2***	***0***	***2***
Point	***2***	***0***	***2***
Purpose	***2***	***0***	***2***
Systems	***2***	***0***	***2***
Tooby	***2***	***0***	***2***
Used	***2***	***0***	***2***
Young	***2***	***0***	***2***
Different	*0*	*6*	*6*
Axe	*0*	*5*	*5*
Knowledge	*0*	*5*	*5*
Integrate	*0*	*4*	*4*
Use	*0*	*4*	*4*
Better	*0*	*3*	*3*
Can't	*0*	*3*	*3*
Developed	*0*	*3*	*3*
Evolved	*0*	*3*	*3*
Fluidity	*0*	*3*	*3*
Got	*0*	*3*	*3*
Objects	*0*	*3*	*3*
Past	*0*	*3*	*3*
Process	*0*	*3*	*3*
Says	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Sense	*0*	*3*	*3*
Structures	*0*	*3*	*3*
Symbolic	*0*	*3*	*3*
Together	*0*	*3*	*3*
Understand	*0*	*3*	*3*
Way	*0*	*3*	*3*
Agrees	*0*	*2*	*2*
Appear	*0*	*2*	*2*
Appears	*0*	*2*	*2*
Applied	*0*	*2*	*2*
Argue	*0*	*2*	*2*
Argues	*0*	*2*	*2*
Biologists	*0*	*2*	*2*
Capacity	*0*	*2*	*2*
Conceive	*0*	*2*	*2*
Cosmides's	*0*	*2*	*2*
Course	*0*	*2*	*2*
Develop	*0*	*2*	*2*
Devoted	*0*	*2*	*2*
Domains	*0*	*2*	*2*
Earlier	*0*	*2*	*2*
History	*0*	*2*	*2*
Information	*0*	*2*	*2*
Intuitive	*0*	*2*	*2*
Knew	*0*	*2*	*2*
Large	*0*	*2*	*2*
Limitless	*0*	*2*	*2*
Limits	*0*	*2*	*2*
Make	*0*	*2*	*2*
Makers	*0*	*2*	*2*
Material	*0*	*2*	*2*
Modified	*0*	*2*	*2*
Need	*0*	*2*	*2*
Obsolete	*0*	*2*	*2*
Ones	*0*	*2*	*2*
Order	*0*	*2*	*2*
Produced	*0*	*2*	*2*
Ready	*0*	*2*	*2*
Seem	*0*	*2*	*2*
Suggests	*0*	*2*	*2*
Tasks	*0*	*2*	*2*
Vital	*0*	*2*	*2*

Sample Text C15: 961013

Words	*Th*	*Rh*	*Overall*
Hiv	**17**	**11**	**28**
Protease	**13**	**3**	**16**
Aids	**10**	**11**	**21**
Virus	**9**	**9**	**18**
Inhibitors	**8**	**1**	**9**
Drugs	**7**	**4**	**11**
New	**7**	**2**	**9**
Treatment	**6**	**7**	**13**
Cells	**6**	**5**	**11**
Levels	**5**	**5**	**10**
Cancer	**5**	**4**	**9**
Drug	**5**	**4**	**9**
Therapy	**5**	**2**	**7**
Tissue	**5**	**1**	**6**
Eradication	**4**	**1**	**5**

Words	*Th*	*Rh*	*Overall*
Ho	**4**	**1**	**5**
Years	**4**	**1**	**5**
Viral	**3**	**10**	**13**
Combination	**3**	**3**	**6**
Infection	**3**	**3**	**6**
Doctors	**3**	**2**	**5**
Blood	**3**	**1**	**4**
Cent	**3**	**1**	**4**
Inhibitor	**3**	**1**	**4**
Low	**3**	**1**	**4**
Triple	**3**	**1**	**4**
Anti	**2**	**4**	**6**
Britain	**2**	**4**	**6**
Known	**2**	**4**	**6**
Scientists	**2**	**4**	**6**

Words	*Th*	*Rh*	*Overall*
Load	**2**	**3**	**5**
Patients	**2**	**3**	**5**
Vancouver	**2**	**3**	**5**
Year	**2**	**3**	**5**
British	**2**	**2**	**4**
Conference	**2**	**2**	**4**
Disease	**2**	**2**	**4**
Immune	**2**	**2**	**4**
Months	**2**	**2**	**4**
Story	**2**	**2**	**4**
Zidovudine	**2**	**2**	**4**
Analogues	**2**	**1**	**3**
Cell	**2**	**1**	**3**
Enzymes	**2**	**1**	**3**
Health	**2**	**1**	**3**
High	**2**	**1**	**3**
Hype	**2**	**1**	**3**
Infected	**2**	**1**	**3**
Many	**2**	**1**	**3**
Model	**2**	**1**	**3**
Nucleoside	**2**	**1**	**3**
Results	**2**	**1**	**3**
Spread	**2**	**1**	**3**
System	**2**	**1**	**3**
Treated	**2**	**1**	**3**
Used	1	5	6
Long	1	4	5
Bloodstream	1	3	4
Called	1	3	4
Coverage	1	2	3
Diseases	1	2	3
Early	1	2	3
Journalists	1	2	3
Licensed	1	2	3
Live	1	2	3
Patient's	1	2	3
People	1	2	3
Proteins	1	2	3
Term	1	2	3
Weeks	1	2	3
Advanced	1	1	2
Amount	1	1	2
Antigens	1	1	2
Approach	1	1	2
Associated	1	1	2
Available	1	1	2
Big	1	1	2
Cause	1	1	2
Company	1	1	2
Effects	1	1	2
Fire	1	1	2
Free	1	1	2
Gay	1	1	2
Goal	1	1	2
Good	1	1	2
Heterosexuals	1	1	2
Including	1	1	2
Indinavir	1	1	2
Initial	1	1	2
Investigation	1	1	2
Lived	1	1	2
Longer	1	1	2

Words	*Th*	*Rh*	*Overall*
Maintained	1	1	2
Mathematical	1	1	2
Month	1	1	2
National	1	1	2
Organisations	1	1	2
Precursor	1	1	2
Professor	1	1	2
Retroviral	1	1	2
Science	1	1	2
Scientific	1	1	2
Sex	1	1	2
Sexually	1	1	2
Short	1	1	2
Stages	1	1	2
Suppression	1	1	2
Theoretical	1	1	2
Total	1	1	2
Transmitted	1	1	2
Treatments	1	1	2
Tumour	1	1	2
Update	1	1	2
Word	1	1	2
London	***3***	***0***	***3***
Waxman	***3***	***0***	***3***
Based	***2***	***0***	***2***
Centre	***2***	***0***	***2***
Class	***2***	***0***	***2***
Daily	***2***	***0***	***2***
Ho's	***2***	***0***	***2***
Hospital	***2***	***0***	***2***
Issue	***2***	***0***	***2***
Lymphoid	***2***	***0***	***2***
Medical	***2***	***0***	***2***
Production	***2***	***0***	***2***
Research	***2***	***0***	***2***
Royal	***2***	***0***	***2***
Saquinavir	***2***	***0***	***2***
School	***2***	***0***	***2***
Subsequent	***2***	***0***	***2***
Use	***2***	***0***	***2***
Views	***2***	***0***	***2***
Williams	***2***	***0***	***2***
Works	***2***	***0***	***2***
World	***2***	***0***	***2***
Body	*0*	*4*	*4*
Failed	*0*	*4*	*4*
Possible	*0*	*4*	*4*
Reduced	*0*	*4*	*4*
Replication	*0*	*4*	*4*
Certainly	*0*	*3*	*3*
Cure	*0*	*3*	*3*
Pounds	*0*	*3*	*3*
Reported	*0*	*3*	*3*
Says	*0*	*3*	*3*
Stopped	*0*	*3*	*3*
Theory	*0*	*3*	*3*
Undetectable	*0*	*3*	*3*
Advance	*0*	*2*	*2*
Block	*0*	*2*	*2*
Brief	*0*	*2*	*2*
Cautious	*0*	*2*	*2*
Dramatically	*0*	*2*	*2*
Earlier	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Enzyme	*0*	*2*	*2*
Fell	*0*	*2*	*2*
Healthy	*0*	*2*	*2*
Infectious	*0*	*2*	*2*
Interest	*0*	*2*	*2*
International	*0*	*2*	*2*
Little	*0*	*2*	*2*
Made	*0*	*2*	*2*
Make	*0*	*2*	*2*
News	*0*	*2*	*2*
Particles	*0*	*2*	*2*
Potent	*0*	*2*	*2*
Potential	*0*	*2*	*2*
Probably	*0*	*2*	*2*
Prognosis	*0*	*2*	*2*
Prove	*0*	*2*	*2*
Put	*0*	*2*	*2*
Said	*0*	*2*	*2*
Space	*0*	*2*	*2*
Stage	*0*	*2*	*2*
Study	*0*	*2*	*2*
Suggested	*0*	*2*	*2*
Sustained	*0*	*2*	*2*

Sample Text C16: 961020

Words	*Th*	*Rh*	*Overall*
Cassava	**6**	**14**	**20**
Genes	**6**	**10**	**16**
Crops	**6**	**5**	**11**
Gene	**6**	**5**	**11**
Resistance	**4**	**2**	**6**
Engineering	**3**	**3**	**6**
Genetic	**3**	**3**	**6**
Natural	**3**	**3**	**5**
Pests	**3**	**3**	**6**
Resistant	**3**	**2**	**5**
Scientists	**3**	**2**	**5**
Ishikazi	**3**	**1**	**4**
New	**2**	**7**	**9**
Food	**2**	**5**	**7**
Make	**2**	**5**	**7**
Temperate	**2**	**3**	**5**
Temperatures	**2**	**2**	**4**
Tobacco	**2**	**2**	**4**
Varieties	**2**	**2**	**4**
Dna	**2**	**1**	**3**
Life	**2**	**1**	**3**
Living	**2**	**1**	**3**
Nishizawa	**2**	**1**	**3**
Opportunities	**2**	**1**	**3**
Swiss	**2**	**1**	**3**
Term	**2**	**1**	**3**
Plants	1	7	8
Crop	1	6	7
Plant	1	6	7
Tropical	1	6	7
Fats	1	5	6
Countries	1	4	5
Grown	1	4	5
Made	1	4	5
Used	1	4	5
Add	1	3	4
Freezing	1	3	4
Means	1	3	3
People	1	3	3
Saturated	1	3	4
Well	1	3	4
World's	1	3	4
Added	1	2	3
Diseases	1	2	3
Low	1	2	3
Making	1	2	3
Pesticide	1	2	3
Plastics	1	2	3
Research	1	2	3
Way	1	2	3
Available	1	1	2
Breakthrough	1	1	2
Butter	1	1	2
Called	1	1	2
Colleagues	1	1	2
Coming	1	1	2
Companies	1	1	2
Competition	1	1	2
Engineered	1	1	2
Forms	1	1	2
Give	1	1	2
Ground	1	1	2
Institute	1	1	2
Nidulans	1	1	2
Past	1	1	2
Recent	1	1	2
Roots	1	1	2
Short	1	1	2
Show	1	1	2
Similar	1	1	2
Success	1	1	2
Techniques	1	1	2
Cereals	***3***	***0***	***3***
Taylor	***3***	***0***	***3***
We'd	***3***	***0***	***3***
Ability	***2***	***0***	***2***
Breeders	***2***	***0***	***2***
Experts	***2***	***0***	***2***
Nigel	***2***	***0***	***2***
Problems	***2***	***0***	***2***
Provide	***2***	***0***	***2***
Viruses	***2***	***0***	***2***
Wanted	***2***	***0***	***2***
Possible	*0*	*6*	*6*
Staple	*0*	*4*	*4*
Use	*0*	*4*	*4*
Using	*0*	*4*	*4*
Bacterium	*0*	*3*	*3*
Biodegradable	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Essential	*0*	*3*	*3*
Feed	*0*	*3*	*3*
Get	*0*	*3*	*3*
Gun	*0*	*3*	*3*
Insect	*0*	*3*	*3*
Put	*0*	*3*	*3*
Taken	*0*	*3*	*3*
Tropics	*0*	*3*	*3*
Tumefaciens	*0*	*3*	*3*
World	*0*	*3*	*3*
Africa	*0*	*2*	*2*
American	*0*	*2*	*2*
Basis	*0*	*2*	*2*
Believes	*0*	*2*	*2*
Breeding	*0*	*2*	*2*
California	*0*	*2*	*2*
Come	*0*	*2*	*2*
Content	*0*	*2*	*2*
Cyanide	*0*	*2*	*2*
Developing	*0*	*2*	*2*
Disadvantage	*0*	*2*	*2*
Discovered	*0*	*2*	*2*
Easier	*0*	*2*	*2*
Engineer	*0*	*2*	*2*
Enzyme	*0*	*2*	*2*
Farmers	*0*	*2*	*2*
Good	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Grow	*0*	*2*	*2*
Growers	*0*	*2*	*2*
Growing	*0*	*2*	*2*
Industries	*0*	*2*	*2*
Introduce	*0*	*2*	*2*
Literally	*0*	*2*	*2*
Makes	*0*	*2*	*2*
Market	*0*	*2*	*2*
Material	*0*	*2*	*2*
Nations	*0*	*2*	*2*
Needed	*0*	*2*	*2*
Number	*0*	*2*	*2*
Opens	*0*	*2*	*2*
Population	*0*	*2*	*2*
Processed	*0*	*2*	*2*
Produce	*0*	*2*	*2*
Producing	*0*	*2*	*2*
Profitable	*0*	*2*	*2*
Raw	*0*	*2*	*2*
Renewable	*0*	*2*	*2*
Richer	*0*	*2*	*2*
Root	*0*	*2*	*2*
Says	*0*	*2*	*2*
Starch	*0*	*2*	*2*
Time	*0*	*2*	*2*
Viral	*0*	*2*	*2*
Worked	*0*	*2*	*2*

Sample Text C17: 961027

Words	*Th*	*Rh*	*Overall*
Plankton	**13**	**8**	**21**
Ocean	**8**	**7**	**15**
Water	**7**	**5**	**12**
Carbon	**5**	**5**	**10**
Nutrients	**5**	**3**	**8**
Oceans	**4**	**6**	**10**
Coccolithophor+	**4**	**4**	**8**
Cent	**4**	**3**	**7**
Surface	**4**	**3**	**7**
Bloom	**4**	**2**	**6**
Woods	**4**	**1**	**5**
Sea	**3**	**6**	**9**
Says	**3**	**5**	**8**
Dms	**3**	**3**	**6**
Phytoplankton	**3**	**3**	**6**
Diatoms	**3**	**2**	**5**
Cold	**3**	**1**	**4**
Example	**3**	**1**	**4**
Climate	**2**	**3**	**5**
Nutrient	**2**	**3**	**5**
Dioxide	**2**	**2**	**4**
Global	**2**	**2**	**4**
Oceanography	**2**	**2**	**4**
Warm	**2**	**2**	**4**
Ice	**2**	**1**	**3**
Viruses	**2**	**1**	**3**
Waters	**2**	**1**	**3**
Atmosphere	1	4	5
Change	1	3	4

Words	*Th*	*Rh*	*Overall*
Drawdown	1	3	4
North	1	3	4
Square	1	3	4
Areas	1	2	3
Atlantic	1	2	3
Dense	1	2	3
Growth	1	2	3
Hundreds	1	2	3
Silicate	1	2	3
Tonnes	1	2	3
Way	1	2	3
Biodiversity	1	1	2
Blooms	1	1	2
Consider	1	1	2
Day	1	1	2
Deep	1	1	2
Depleted	1	1	2
Get	1	1	2
Giving	1	1	2
Group	1	1	2
Holligan	1	1	2
Huxleyii	1	1	2
Including	1	1	2
Kilometre	1	1	2
Little	1	1	2
Mann	1	1	2
Osmo	1	1	2
Osmotic	1	1	2
Plants	1	1	2
Production	1	1	2
Research	1	1	2

Words	*Th*	*Rh*	*Overall*
Response	1	1	2
Result	1	1	2
Shifts	1	1	2
Solar	1	1	2
Solutes	1	1	2
Stronger	1	1	2
Succession	1	1	2
Sulphur	1	1	2
Sunlight	1	1	2
Using	1	1	2
Winds	1	1	2
Years	1	1	2
Iron	***5***	***0***	***5***
Large	***3***	***0***	***3***
Altitude	***2***	***0***	***2***
Average	***2***	***0***	***2***
Classical	***2***	***0***	***2***
Emissions	***2***	***0***	***2***
Experiment	***2***	***0***	***2***
John	***2***	***0***	***2***
Life	***2***	***0***	***2***
Low	***2***	***0***	***2***
Metre	***2***	***0***	***2***
Model	***2***	***0***	***2***
Professor	***2***	***0***	***2***
Stop	***2***	***0***	***2***
Theory	***2***	***0***	***2***
University	***2***	***0***	***2***
Watson	***2***	***0***	***2***
Species	*0*	*7*	*7*
Data	*0*	*4*	*4*
Know	*0*	*4*	*4*
Light	*0*	*4*	*4*
Abundance	*0*	*3*	*3*
Bacteria	*0*	*3*	*3*
Based	*0*	*3*	*3*
Ecosystems	*0*	*3*	*3*
Long	*0*	*3*	*3*
Major	*0*	*3*	*3*
Responsible	*0*	*3*	*3*
Significant	*0*	*3*	*3*
Act	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Availability	*0*	*2*	*2*
Billion	*0*	*2*	*2*
Biological	*0*	*2*	*2*
Cloud	*0*	*2*	*2*
Conditions	*0*	*2*	*2*
Contain	*0*	*2*	*2*
Deal	*0*	*2*	*2*
Diatom	*0*	*2*	*2*
Different	*0*	*2*	*2*
Droplets	*0*	*2*	*2*
Encountered	*0*	*2*	*2*
Enough	*0*	*2*	*2*
Factors	*0*	*2*	*2*
Full	*0*	*2*	*2*
Gas	*0*	*2*	*2*
Great	*0*	*2*	*2*
Important	*0*	*2*	*2*
Increasing	*0*	*2*	*2*
Kilometres	*0*	*2*	*2*
Look	*0*	*2*	*2*
Make	*0*	*2*	*2*
Massive	*0*	*2*	*2*
Move	*0*	*2*	*2*
Particular	*0*	*2*	*2*
Phenomena	*0*	*2*	*2*
Play	*0*	*2*	*2*
Pressure	*0*	*2*	*2*
Put	*0*	*2*	*2*
Role	*0*	*2*	*2*
Sink	*0*	*2*	*2*
Storage	*0*	*2*	*2*
Sulphate	*0*	*2*	*2*
Term	*0*	*2*	*2*
Thousands	*0*	*2*	*2*
Tiny	*0*	*2*	*2*
Vast	*0*	*2*	*2*
Virtual	*0*	*2*	*2*
Warmer	*0*	*2*	*2*
Warming	*0*	*2*	*2*
World	*0*	*2*	*2*

Sample Text C18: 961103

Words	*Th*	*Rh*	*Overall*
Eye	**13**	**16**	**29**
Drivers	**5**	**3**	**8**
Tracking	**5**	**3**	**8**
Professor	**4**	**2**	**6**
Land	**3**	**1**	**4**
Information	**2**	**6**	**8**
Movement	**2**	**2**	**4**
Work	**2**	**2**	**4**
It's	**2**	**1**	**3**
Moves	**2**	**1**	**3**
Research	**2**	**1**	**3**
New	1	4	5
Video	1	3	4
Camera	1	2	3
Look	1	2	3

Words	*Th*	*Rh*	*Overall*
Movements	1	2	3
Department	1	1	2
Design	1	1	2
Devices	1	1	2
Driver's	1	1	2
Driving	1	1	2
Exactly	1	1	2
Experienced	1	1	2
Fixations	1	1	2
Found	1	1	2
Head	1	1	2
Line	1	1	2
Looking	1	1	2
Looks	1	1	2
Package	1	1	2
Points	1	1	2

Words	*Th*	*Rh*	*Overall*
Shoppers	1	1	2
Siebert	1	1	2
Transport	1	1	2
Visual	1	1	2
World	1	1	2
Worn	1	1	2
Eyes	***4***	***0***	***4***
Long	***3***	***0***	***3***
University	***3***	***0***	***3***
Gaze	***2***	***0***	***2***
Id	***2***	***0***	***2***
Magasin	***2***	***0***	***2***
Pianists	***2***	***0***	***2***
Recorders	***2***	***0***	***2***
Shopping	***2***	***0***	***2***
Specific	***2***	***0***	***2***
Student	***2***	***0***	***2***
There's	***2***	***0***	***2***
Time	***2***	***0***	***2***
Frame	*0*	*5*	*5*
Road	*0*	*5*	*5*
Says	*0*	*5*	*5*
Providing	*0*	*4*	*4*
See	*0*	*4*	*4*

Words	*Th*	*Rh*	*Overall*
Data	*0*	*3*	*3*
Done	*0*	*3*	*3*
Using	*0*	*3*	*3*
Activity	*0*	*2*	*2*
Analysed	*0*	*2*	*2*
Brain	*0*	*2*	*2*
Chart	*0*	*2*	*2*
Eager	*0*	*2*	*2*
Face	*0*	*2*	*2*
Getting	*0*	*2*	*2*
Hazards	*0*	*2*	*2*
Music	*0*	*2*	*2*
Screen	*0*	*2*	*2*
Search	*0*	*2*	*2*
Shelf	*0*	*2*	*2*
Sight	*0*	*2*	*2*
Stay	*0*	*2*	*2*
Steering	*0*	*2*	*2*
Supermarket	*0*	*2*	*2*
Table	*0*	*2*	*2*
Take	*0*	*2*	*2*
Tennis	*0*	*2*	*2*
Translate	*0*	*2*	*2*
Used	*0*	*2*	*2*
Yielding	*0*	*2*	*2*

Sample Text C19: 961110

Words	*Th*	*Rh*	*Overall*
Mars	**23**	**25**	**48**
Surveyor	**7**	**2**	**9**
Life	**6**	**6**	**12**
Missions	**5**	**1**	**6**
Martian	**4**	**11**	**15**
Earth	**4**	**5**	**9**
Mission	**4**	**3**	**7**
Pathfinder	**4**	**3**	**7**
Lander	**4**	**2**	**6**
Global	**4**	**1**	**5**
Surface	**3**	**22**	**25**
Russian	**3**	**5**	**8**
Craft	**3**	**4**	**7**
Spacecraft	**3**	**2**	**5**
Come	**2**	**6**	**8**
Launch	**2**	**6**	**8**
American	**2**	**4**	**6**
Orbit	**2**	**4**	**6**
Rocks	**2**	**4**	**6**
Time	**2**	**4**	**6**
Water	**2**	**4**	**6**
Years	**2**	**4**	**6**
Rock	**2**	**3**	**5**
Entire	**2**	**2**	**4**
Experiments	**2**	**2**	**4**
Found	**2**	**2**	**4**
Nasa	**2**	**2**	**4**
Orbiter	**2**	**2**	**4**
Samples	**2**	**2**	**4**
Year's	**2**	**1**	**3**
Atmosphere	1	6	7

Words	*Th*	*Rh*	*Overall*
Observer	1	4	5
Evidence	1	3	4
Landing	1	3	4
Solar	1	3	4
Called	1	2	3
Heat	1	2	3
Landers	1	2	3
Neighbour	1	2	3
Planetary	1	2	3
Provide	1	2	3
Shield	1	2	3
Size	1	2	3
System	1	2	3
August	1	1	2
Bacteria	1	1	2
Cable	1	1	2
Camera	1	1	2
Day	1	1	2
Feature	1	1	2
Find	1	1	2
Fossil	1	1	2
Future	1	1	2
Go	1	1	2
July	1	1	2
Launched	1	1	2
Many	1	1	2
Meteorite	1	1	2
Moon	1	1	2
Organisms	1	1	2
Parts	1	1	2
Planets	1	1	2
Possibly	1	1	2
Return	1	1	2

Words	*Th*	*Rh*	*Overall*
Rockets	1	1	2
Same	1	1	2
Seismic	1	1	2
Sent	1	1	2
Series	1	1	2
Seventies	1	1	2
Sorts	1	1	2
Successful	1	1	2
Tether	1	1	2
Viking	1	1	2
Perhaps	***4***	***0***	***4***
Arrives	***3***	***0***	***3***
Robot	***3***	***0***	***3***
Rover	***3***	***0***	***3***
Scientists	***3***	***0***	***3***
September	***3***	***0***	***3***
Sojourner	***3***	***0***	***3***
Ares	***2***	***0***	***2***
Dan	***2***	***0***	***2***
Early	***2***	***0***	***2***
Goldin	***2***	***0***	***2***
Hope	***2***	***0***	***2***
Humans	***2***	***0***	***2***
It's	***2***	***0***	***2***
Meteorites	***2***	***0***	***2***
Parachute	***2***	***0***	***2***
Sensors	***2***	***0***	***2***
Vallis	***2***	***0***	***2***
Year	***2***	***0***	***2***
Weather	*0*	*7*	*7*
Planet	*0*	*6*	*6*
Carry	*0*	*5*	*5*
Well	*0*	*5*	*5*
Cover	*0*	*4*	*4*
Ice	*0*	*4*	*4*
Set	*0*	*4*	*4*
Study	*0*	*4*	*4*
Composition	*0*	*3*	*3*
Data	*0*	*3*	*3*
Marsquakes	*0*	*3*	*3*
Metres	*0*	*3*	*3*
Red	*0*	*3*	*3*
Take	*0*	*3*	*3*
Aerobraking	*0*	*2*	*2*
Area	*0*	*2*	*2*
Bit	*0*	*2*	*2*

Words	*Th*	*Rh*	*Overall*
Build	*0*	*2*	*2*
Cape	*0*	*2*	*2*
Clear	*0*	*2*	*2*
Contact	*0*	*2*	*2*
Contamination	*0*	*2*	*2*
Continues	*0*	*2*	*2*
Descended	*0*	*2*	*2*
Detail	*0*	*2*	*2*
Different	*0*	*2*	*2*
Discovery	*0*	*2*	*2*
Dry	*0*	*2*	*2*
Fall	*0*	*2*	*2*
Fire	*0*	*2*	*2*
Free	*0*	*2*	*2*
Look	*0*	*2*	*2*
Looking	*0*	*2*	*2*
Looks	*0*	*2*	*2*
Made	*0*	*2*	*2*
Make	*0*	*2*	*2*
Manoeuvres	*0*	*2*	*2*
Mapping	*0*	*2*	*2*
Measure	*0*	*2*	*2*
Move	*0*	*2*	*2*
New	*0*	*2*	*2*
November	*0*	*2*	*2*
Open	*0*	*2*	*2*
Operate	*0*	*2*	*2*
Penetrators	*0*	*2*	*2*
Photograph	*0*	*2*	*2*
Photographs	*0*	*2*	*2*
Picture	*0*	*2*	*2*
Planet's	*0*	*2*	*2*
Possible	*0*	*2*	*2*
Present	*0*	*2*	*2*
Radio	*0*	*2*	*2*
Relay	*0*	*2*	*2*
Returning	*0*	*2*	*2*
Risk	*0*	*2*	*2*
Volcanic	*0*	*2*	*2*
Wind	*0*	*2*	*2*
Window	*0*	*2*	*2*

Sample Text C20: 961117

Words	*Th*	*Rh*	*Overall*
It's	**6**	**1**	**7**
Science	**4**	**3**	**7**
People	**4**	**1**	**5**
Public	**4**	**1**	**5**
Know	**2**	**4**	**6**
Theory	**2**	**3**	**5**
Understanding	**2**	**3**	**5**
Dawkins	**2**	**1**	**3**
Need	**2**	**1**	**3**
Nina	**2**	**1**	**2**
Reaction	**2**	**1**	**3**

Words	*Th*	*Rh*	*Overall*
Test	**2**	**1**	**3**
Get	1	6	7
Don't	1	4	5
Technology	1	3	4
Group	1	2	3
Man	1	2	3
Place	1	2	3
Quantum	1	2	3
Scientific	1	2	3
Well	1	2	3
Abstract	1	1	2
B	1	1	2

Words	*Th*	*Rh*	*Overall*
Course	1	1	2
Dangerous	1	1	2
Hell	1	1	2
Law	1	1	2
Makes	1	1	2
Members	1	1	2
Real	1	1	2
Take	1	1	2
Theories	1	1	2
Argues	***2***	***0***	***2***
Breaks	***2***	***0***	***2***
Jack	***2***	***0***	***2***
Lewis	***2***	***0***	***2***
Machine	***2***	***0***	***2***
Olivia	***2***	***0***	***2***
Problem	***2***	***0***	***2***
Robin	***2***	***0***	***2***
Shame	***2***	***0***	***2***
Washing	***2***	***0***	***2***
Wolpert	***2***	***0***	***2***
You're	***2***	***0***	***2***
Things	*0*	*4*	*4*
Think	*0*	*4*	*4*
Thought	*0*	*4*	*3*
Come	*0*	*3*	*3*
Ignorance	*0*	*3*	*3*
Knowledge	*0*	*3*	*3*

Words	*Th*	*Rh*	*Overall*
Seems	*0*	*3*	*3*
Star	*0*	*3*	*3*
Ask	*0*	*2*	*2*
Big	*0*	*2*	*2*
C	*0*	*2*	*2*
Culture	*0*	*2*	*2*
Didn't	*0*	*2*	*2*
Fix	*0*	*2*	*2*
Forms	*0*	*2*	*2*
Fridge	*0*	*2*	*2*
Going	*0*	*2*	*2*
Got	*0*	*2*	*2*
Holes	*0*	*2*	*2*
Ideas	*0*	*2*	*2*
Informed	*0*	*2*	*2*
Lack	*0*	*2*	*2*
Life	*0*	*2*	*2*
Many	*0*	*2*	*2*
System	*0*	*2*	*2*
Takes	*0*	*2*	*2*
Technological	*0*	*2*	*2*
Universe	*0*	*2*	*2*
Use	*0*	*2*	*2*
Way	*0*	*2*	*2*
Works	*0*	*2*	*2*

Appendix 11. Key words of Group B Sample Texts

Sample Text B1: 950319

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Apes	10	0.64	145		146.3	0.000000
2	Intelligence	16	1.03	6,718		127.7	0.000000
3	Orangutans	6	0.38	3		120.8	0.000000
4	Gorillas	7	0.45	85		104.8	0.000000
5	Byrne	9	0.58	602		104.6	0.000000
6	Chimps	7	0.45	101		102.4	0.000000
7	Ape	6	0.38	205		77.7	0.000000
8	Orangutan	4	0.26	4		77.1	0.000000
9	Primates	5	0.32	69		73.6	0.000000
10	Primate	5	0.32	135		67.1	0.000000
11	Ancestors	5	0.32	485		54.4	0.000000
12	Monkeys	4	0.26	269		46.4	0.000000
13	Humans	5	0.32	1,125		46.1	0.000000
14	Species	6	0.38	3,322		44.5	0.000000
15	Borneo	3	0.19	121		37.9	0.000000
16	Complex	6	0.38	5,931		37.7	0.000000
17	Social	9	0.58	31,973	0.03	34.3	0.000000
18	Understanding	5	0.32	4,864		31.6	0.000000
19	Sheep	4	0.26	1,901		30.9	0.000000
20	Imitation	3	0.19	450		30.0	0.000000
21	Behaviour	5	0.32	6,352		28.9	0.000000
22	Requires	4	0.26	2,833		27.7	0.000000
23	Evolved	3	0.19	695		27.5	0.000000
24	Groups	6	0.38	14,225	0.01	27.5	0.000000
25	Evolution	3	0.19	832		26.4	0.000000
26	Human	6	0.38	17,036	0.02	25.4	0.000000
27	Animals	4	0.26	4,251		24.5	0.000001

Sample Text B2: 950514

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Molecules	14	1.01	419		188.4	0.000000
2	Chiral	8	0.58	0		178.3	0.000000
3	Enzymes	7	0.51	195		95.2	0.000000
4	Davies	12	0.87	7,035		90.7	0.000000
5	Oxford	12	0.87	7,954		87.7	0.000000
6	Handed	10	0.72	4,653		80.1	0.000000

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
7	Handedness	5	0.36	120		69.4	0.000000
8	Chirality	3	0.22	0		66.8	0.000000
9	Asymmetry	4	0.29	35		63.3	0.000000
10	Davies's	5	0.36	386		57.9	0.000000
11	Left	13	0.94	47,241	0.05	51.9	0.000000
12	Pheromone	3	0.22	39		45.2	0.000000
13	Palm	4	0.29	858		38.2	0.000000
14	Molecule	3	0.22	170		36.6	0.000000
15	Scaffolding	3	0.22	266		33.9	0.000000
16	Molecular	3	0.22	343		32.4	0.000000
17	Synthetic	3	0.22	425		31.1	0.000000
18	Right	9	0.65	50,914	0.05	28.5	0.000000
19	Chemical	4	0.29	3,326		27.4	0.000000
20	Nature	5	0.36	8,714		27.0	0.000000
21	Research	6	0.43	16,892	0.02	26.9	0.000000
22	Drug	5	0.36	9,590	0.01	26.1	0.000000
23	Reactions	3	0.22	1,057		25.7	0.000000
24	Company	8	0.58	45,677	0.05	25.2	0.000001
25	Forms	4	0.29	4,495		25.1	0.000001

Sample Text B3: 950709

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Sloths	9	0.61	12		170.8	0.000000
2	Animals	14	0.95	4,251		122.3	0.000000
3	Species	12	0.82	3,322		107.0	0.000000
4	Pleistocene	5	0.34	5		96.9	0.000000
5	Extinctions	4	0.27	29		64.3	0.000000
6	America	10	0.68	11,918	0.01	60.3	0.000000
7	Pliocene	3	0.20	2		59.7	0.000000
8	Mammals	5	0.34	308		59.5	0.000000
9	Climate	7	0.48	3,259		55.2	0.000000
10	South	13	0.88	40,982	0.04	53.9	0.000000
11	Native	6	0.41	2,303		49.6	0.000000
12	Hoofed	3	0.20	24		47.6	0.000000
13	Extinct	4	0.27	341		45.0	0.000000
14	Northerners	3	0.20	116		38.5	0.000000
15	Equator	3	0.20	130		37.8	0.000000
16	Mass	6	0.41	6,399		37.5	0.000000
17	Creatures	4	0.27	901		37.3	0.000000
18	Southerners	3	0.20	151		36.9	0.000000
19	Ancestors	3	0.20	485		30.0	0.000000
20	Elephants	3	0.20	490		29.9	0.000000
21	North	7	0.48	26,549	0.03	26.6	0.000000

Sample Text B4: 950820

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Kew's	4	0.33	3		80.7	0.000000
2	Evolution	7	0.58	832		77.0	0.000000
3	Plants	9	0.75	4,350		73.9	0.000000
4	Lonsdale	5	0.42	99		72.7	0.000000
5	Cooksonia	3	0.25	0		67.7	0.000000
6	Kew	5	0.42	219		64.9	0.000000

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
7	Mosses	4	0.33	49		61.9	0.000000
8	Rocks	6	0.50	1,090		60.9	0.000000
9	Cyanobacteria	3	0.25	5		57.1	0.000000
10	Extinct	4	0.33	341		46.6	0.000000
11	Primeval	3	0.25	84		41.6	0.000000
12	Landscape	5	0.42	2,381		41.2	0.000000
13	Evolved	4	0.33	695		41.0	0.000000
14	Ferns	3	0.25	104		40.3	0.000000
15	Blackbird	3	0.25	114		39.8	0.000000
16	Visitors	5	0.42	4,897		34.0	0.000000
17	Period	6	0.50	14,116	0.01	30.6	0.000000
18	Stems	3	0.25	837		27.9	0.000000
19	Ancient	4	0.33	3,894		27.3	0.000000
20	Exhibition	4	0.33	4,319		26.5	0.000000
21	Giant	4	0.33	4,710		25.8	0.000000
22	Mud	3	0.25	1,516		24.4	0.000001

Sample Text B5: 950903

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Risk	37	1.74	12,928	0.01	286.3	0.000000
2	Risks	23	1.08	3,239		219.3	0.000000
3	Nuclear	22	1.04	12,526	0.01	148.8	0.000000
4	Hse	7	0.33	240		86.3	0.000000
5	Chemicals	9	0.42	1,887		78.6	0.000000
6	Cancer	9	0.42	5,088		61.0	0.000000
7	Accidents	7	0.33	1,621		59.8	0.000000
8	Public	17	0.80	53,719	0.06	58.7	0.000000
9	Death	12	0.57	24,176	0.03	51.6	0.000000
10	Smoking	7	0.33	3,035		51.1	0.000000
11	Paling	3	0.14	18		47.0	0.000000
12	Accident	6	0.28	4,802		36.5	0.000000
13	Radiation	4	0.19	1,094		32.8	0.000000
14	Plant	6	0.28	6,909		32.3	0.000000
15	Industry	9	0.42	29,387	0.03	30.5	0.000000
16	Public's	3	0.14	632		26.2	0.000000
17	Affect	4	0.19	2,842		25.3	0.000000
18	Average	6	0.28	13,899	0.01	24.2	0.000001

Sample Text B6: 951210

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Bse	16	1.02	451		213.3	0.000000
2	Cjd	6	0.38	85		88.0	0.000000
3	Scrapie	5	0.32	57		75.4	0.000000
4	Scientific	9	0.58	3,543		72.9	0.000000
5	Beef	5	0.32	1,460		43.4	0.000000
6	Agent	6	0.38	3,811		42.9	0.000000
7	Scientists	6	0.38	4,312		41.4	0.000000
8	Experiment	5	0.32	2,045		40.1	0.000000
9	Bruce's	3	0.19	108		38.5	0.000000
10	Cows	4	0.26	765		38.1	0.000000
11	Infect	3	0.19	124		37.7	0.000000
12	Humans	4	0.26	1,125		35.0	0.000000

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
13	Farmers	5	0.32	3,803		34.0	0.000000
14	Disease	5	0.32	5,287		30.7	0.000000
15	Protein	3	0.19	558		28.7	0.000000
16	Positive	5	0.32	6,898		28.1	0.000000
17	Problem	7	0.45	23,738	0.02	27.2	0.000000
18	Transmitted	3	0.19	755		26.9	0.000000
19	Species	4	0.26	3,322		26.5	0.000000
20	Brain	4	0.26	3,590		25.9	0.000000

Sample Text B7: 960407

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Sugars	18	1.50	117		300.4	0.000000
2	Glucose	13	1.09	106		211.4	0.000000
3	Sugoids	9	0.75	0		203.1	0.000000
4	Sugoid	7	0.58	0		158.0	0.000000
5	Mannose	3	0.25	0		67.7	0.000000
6	Virus	7	0.58	1,930		65.3	0.000000
7	Hbv	3	0.25	3		59.4	0.000000
8	Cells	6	0.50	2,300		52.1	0.000000
9	Toxic	5	0.42	852		51.4	0.000000
10	Molecules	4	0.33	419		45.0	0.000000
11	Coating	3	0.25	162		37.7	0.000000
12	Enzyme	3	0.25	167		37.5	0.000000
13	Proteins	3	0.25	236		35.5	0.000000
14	Hepatitis	3	0.25	275		34.6	0.000000
15	Viruses	3	0.25	377		32.7	0.000000
16	Sugar	4	0.33	3,127		29.0	0.000000
17	Blood	5	0.42	8,303		28.9	0.000000
18	Fold	3	0.25	982		27.0	0.000000
19	Coat	3	0.25	1,160		26.0	0.000000

Sample Text B8: 960519

	Keywords	freq.	°o	ref. freq.	°o	keyness	p-value
1	Cerro	10	0.71	14		190.0	0.000000
2	Paranal	7	0.50	5		139.5	0.000000
3	Telescope	8	0.57	350		101.4	0.000000
4	Vlt	5	0.36	3		100.7	0.000000
5	Observatory	6	0.43	202		79.2	0.000000
6	Light	9	0.64	14,110	0.01	50.3	0.000000
7	Chilean	4	0.29	436		43.5	0.000000
8	Chile	4	0.29	735		39.3	0.000000
9	Astronomers	3	0.21	263		33.9	0.000000
10	Space	6	0.43	10,041	0.01	32.8	0.000000
11	Mirror	5	0.36	5,010		32.3	0.000000
12	Mirrors	3	0.21	652		28.5	0.000000
13	Tolerance	3	0.21	838		27.0	0.000000
14	Observation	3	0.21	1,223		24.7	0.000001
15	Instrument	3	0.21	1,379		24.0	0.000001

Sample Text B9: 960526

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Shopping	18	1.49	3,894		176.6	0.000000
2	Store	10	0.83	3,506		88.5	0.000000
3	Supermarket	8	0.66	2,055		75.7	0.000000
4	Card	10	0.83	7,150		74.3	0.000000
5	Digital	6	0.50	1,612		56.2	0.000000
6	Technology	7	0.58	9,527	0.01	43.1	0.000000
7	Loyalty	5	0.41	1,960		43.1	0.000000
8	Stored	4	0.33	837		39.5	0.000000
9	Electronic	5	0.41	3,364		37.7	0.000000
10	Internet	4	0.33	1,535		34.6	0.000000
11	Retailers	4	0.33	1,675		33.9	0.000000
12	List	6	0.50	11,552	0.01	32.9	0.000000
13	Icl	3	0.25	462		31.4	0.000000
14	Smart	4	0.33	2,692		30.2	0.000000
15	Stores	4	0.33	3,235		28.7	0.000000
16	Customers	5	0.41	8,818		28.2	0.000000
17	Pc	4	0.33	4,085		26.9	0.000000
18	Believes	5	0.41	10,471	0.01	26.6	0.000000
19	Screen	4	0.33	5,654		24.3	0.000001

Sample Text B10: 960623

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
1	Nature	19	1.07	8,714		143.5	0.000000
2	Countryside	8	0.45	2,228		68.3	0.000000
3	Landscapes	5	0.28	489		53.1	0.000000
4	Study	9	0.51	10,229	0.01	51.8	0.000000
5	Sensory	4	0.23	136		50.8	0.000000
6	Example	8	0.45	15,057	0.02	38.2	0.000000
7	Humans	4	0.23	1,125		34.1	0.000000
8	Natural	6	0.34	8,049		32.6	0.000000
9	Urban	5	0.28	3,898		32.5	0.000000
10	Experiences	4	0.23	2,064		29.2	0.000000
11	Wilderness	3	0.17	683		26.8	0.000000

Sample Text B11: 960630

	Keywords	freq.	°o	ref. freq.	°o	keyness	p-value
1	Apes	23	1.27	145		366.0	0.000000
2	Language	22	1.21	10,317	0.01	164.1	0.000000
3	Chimpanzees	9	0.50	94		134.6	0.000000
4	Chimp	8	0.44	58		125.1	0.000000
5	Rumbaugh	6	0.33	4		117.0	0.000000
6	Chimps	8	0.44	101		116.7	0.000000
7	Fouts	4	0.22	0		86.9	0.000000
8	Booee	4	0.22	0		86.9	0.000000
9	Professor	13	0.72	9,241		86.2	0.000000
10	Sign	12	0.66	9,635	0.01	76.7	0.000000
11	Kanzi	4	0.22	6		73.5	0.000000
12	Savage	7	0.39	1,257		65.5	0.000000
13	Panbanisha	3	0.17	0		65.2	0.000000
14	Food	12	0.66	17,453	0.02	62.7	0.000000

	Keywords	freq.	%	ref. freq.	%	keyness	p-value
15	Use	14	0.77	32,129	0.03	60.9	0.000000
16	Washoe	3	0.17	3		56.9	0.000000
17	Rivas	3	0.17	14		49.4	0.000000
18	Portions	4	0.22	199		47.6	0.000000
19	Wild	7	0.39	4,870		46.7	0.000000
20	Katharine	4	0.22	344		43.3	0.000000
21	Chimpanzee	3	0.17	96		38.3	0.000000
22	Primate	3	0.17	135		36.3	0.000000
23	Human	8	0.44	17,036	0.02	35.9	0.000000
24	Signing	5	0.28	2,687		35.9	0.000000
25	Rudimentary	3	0.17	191		34.2	0.000000
26	Ape	3	0.17	205		33.8	0.000000
27	Communicating	3	0.17	285		31.9	0.000000
28	Words	7	0.39	14,541	0.02	31.8	0.000000
29	Research	7	0.39	16,892	0.02	29.8	0.000000
30	Communicate	3	0.17	831		25.5	0.000000

Appendix 12. Key words of Group C Sample Texts

Sample Text C1: 960707

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Mantle	35	1.11	396		479.7	0.000000
2 Core	30	0.95	3,821		268.4	0.000000
3 Earth	30	0.95	6,524		236.6	0.000000
4 Seismic	13	0.41	187		172.1	0.000000
5 Magnetic	10	0.32	549		106.1	0.000000
6 Waves	10	0.32	1,708		83.6	0.000000
7 Earth's	8	0.25	478		83.5	0.000000
8 Perovskite	4	0.13	0		82.5	0.000000
9 Earthquakes	6	0.19	221		68.4	0.000000
10 Surface	10	0.32	4,088		66.3	0.000000
11 Crust	6	0.19	274		65.8	0.000000
12 Geologists	5	0.16	92		63.8	0.000000
13 Pressure	14	0.44	17,368	0.02	62.6	0.000000
14 Diamond	7	0.22	1,087		59.8	0.000000
15 Layer	7	0.22	1,148		59.1	0.000000
16 Diamonds	6	0.19	538		57.8	0.000000
17 Spinel	3	0.1	3		53.6	0.000000
18 Structure	9	0.29	5,542		52.4	0.000000
19 Iron	8	0.25	3,642		51.4	0.000000
20 Liquid	6	0.19	1,146		48.8	0.000000
21 Rock	9	0.29	7,635		46.8	0.000000
22 Boundary	6	0.19	1,357		46.8	0.000000
23 Solid	7	0.22	2,934		46.1	0.000000
24 Ocean	6	0.19	1,505		45.6	0.000000
25 Hotter	4	0.13	155		45.2	0.000000
26 Circulation	6	0.19	1,714		44	0.000000
27 Tomography	3	0.1	22		43.6	0.000000
28 Phase	6	0.19	2,344		40.3	0.000000
29 Field	9	0.29	11,622	0.01	39.5	0.000000
30 Minerals	4	0.13	318		39.5	0.000000
31 Pressures	6	0.19	2,593		39.1	0.000000
32 Deep	8	0.25	9,279		36.8	0.000000
33 Temperatures	5	0.16	1,533		36	0.000000
34 Hot	7	0.22	6,259		35.7	0.000000
35 Simulations	3	0.1	96		35	0.000000

Keywords	freq.	%	ref. freq.	%	keyness	p-value
36 Sample	5	0.16	1,759		34.6	0.000000
37 Molten	3	0.1	103		34.6	0.000000
38 Mineral	4	0.13	624		34.1	0.000000
39 Depth	5	0.16	2,070		33	0.000000
40 Volcanoes	3	0.1	135		33	0.000000
41 Planet's	3	0.1	140		32.8	0.000000
42 Crystal	5	0.16	2,406		31.5	0.000000
43 Rocks	4	0.13	1,090		29.7	0.000000
44 Cooler	3	0.1	246		29.4	0.000000
45 Continents	3	0.1	274		28.8	0.000000
46 Heat	5	0.16	3,963		26.7	0.000000
47 Centre	9	0.29	25,511	0.03	26.3	0.000000
48 Melting	3	0.1	443		25.9	0.000000
49 Drill	3	0.1	447		25.9	0.000000

Sample Text C2: 960714

Keywords	freq.	%	ref. freq.	%	keyness	p-value
1 Pleasure	16	1.32	3,169		159.5	0.000000
2 Warburton	7	0.58	84		108.4	0.000000
3 Arise's	3	0.25	1		63.1	0.000000
4 Measurably	3	0.25	17		50.7	0.000000
5 Stress	6	0.49	3,369		47.3	0.000000
6 Guilt	5	0.41	1,869		43.5	0.000000
7 Pathway	3	0.25	70		42.6	0.000000
8 Research	8	0.66	16,893	0.02	42.3	0.000000
9 Chocolate	4	0.33	1,712		33.7	0.000000
10 Coffee	4	0.33	3,035		29.2	0.000000
11 Brain	4	0.33	3,591		27.8	0.000000
12 Professor	5	0.41	9,243		27.7	0.000000
13 Arise	3	0.25	1,020		26.7	0.000000

Sample Text C3: 960721

Keywords	freq.	%	ref. freq.	%	keyness	p-value
1 Flowers	14	0.77	3,164		124.4	0.000000
2 Native	10	0.55	2,305		88.5	0.000000
3 Plants	11	0.6	4,347		85.6	0.000000
4 Flowering	6	0.33	533		64.4	0.000000
5 Heather	6	0.33	580		63.4	0.000000
6 Species	8	0.44	3,322		61.4	0.000000
7 Habitats	5	0.27	245		59.6	0.000000
8 Esas	3	0.16	4		55.6	0.000000
9 Ploughed	5	0.27	382		55.2	0.000000
10 Rich	9	0.49	8,551		54.4	0.000000
11 Grassland	4	0.22	85		54.2	0.000000
12 Yellow	7	0.38	3,276		52.1	0.000000
13 Farm	7	0.38	4,115		48.9	0.000000
14 Countryside	6	0.33	2,229		47.4	0.000000
15 Hay	5	0.27	897		46.7	0.000000
16 Cowslips	3	0.16	31		44.9	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
17 Meadows	4	0.22	298		44.3	0.000000
18 Roadsides	3	0.16	37		43.8	0.000000
19 Grasslands	3	0.16	47		42.5	0.000000
20 Plant	7	0.38	6,910		41.8	0.000000
21 Drained	4	0.22	436		41.3	0.000000
22 Arable	3	0.16	126		36.7	0.000000
23 Tim	6	0.33	5,726		36.2	0.000000
24 Common	7	0.38	12,626	0.01	33.6	0.000000
25 Pastures	3	0.16	221		33.3	0.000000
26 Becoming	6	0.33	7,948		32.4	0.000000
27 Flora	3	0.16	391		29.9	0.000000

Sample Text C4: 960728

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Ants	17	1.06	275		244.2	0.000000
2 Caterpillars	15	0.94	108		238.7	0.000000
3 Butterfly	15	0.94	485		195.2	0.000000
4 Ant	13	0.81	184		190.1	0.000000
5 Blue	20	1.25	9,065		155.5	0.000000
6 Large	24	1.5	22,609	0.02	152	0.000000
7 Colony	9	0.56	1,219		91.5	0.000000
8 Grubs	5	0.31	34		80.1	0.000000
9 Caterpillar	5	0.31	133		66.9	0.000000
10 Thomas	9	0.56	8,186		57.5	0.000000
11 Butterflies	5	0.31	384		56.5	0.000000
12 Rabbits	5	0.31	416		55.7	0.000000
13 Larvae	4	0.25	95		54.4	0.000000
14 Species	7	0.44	3,322		53.7	0.000000
15 Thyme	4	0.25	158		50.4	0.000000
16 Red	9	0.56	15,257	0.02	46.6	0.000000
17 Nest	4	0.25	793		37.6	0.000000
18 Mimic	3	0.19	166		35.8	0.000000
19 Extinct	3	0.19	341		31.5	0.000000
20 Smell	4	0.25	1,802		31.1	0.000000
21 Land	6	0.37	13,867	0.01	27.4	0.000000
22 Grass	4	0.25	3,114		26.8	0.000000
23 Creatures	3	0.19	901		25.7	0.000000
24 Britain	8	0.5	41,928	0.04	24.3	0.000001

Sample Text C5: 960804

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Gaia	11	0.48	109		160.6	0.000000
2 Lovelock	10	0.44	73		151.8	0.000000
3 Organisms	12	0.53	338		150.9	0.000000
4 Granite	11	0.48	289		139.9	0.000000
5 Gaian	6	0.26	1		122	0.000000
6 Biosphere	8	0.35	69		118.9	0.000000
7 Earth	11	0.48	6,524		72	0.000000
8 Charlson	3	0.13	0		63.9	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
9 Coccolithosporids	3	0.13	0		63.9	0.000000
10 Petford	3	0.13	0		63.9	0.000000
11 Water	13	0.57	20,146	0.02	60.7	0.000000
12 Superorganism	3	0.13	1		59.4	0.000000
13 Ants	5	0.22	275		56.3	0.000000
14 Oxygen	6	0.26	850		56.3	0.000000
15 Regulating	5	0.22	286		55.9	0.000000
16 Gcms	3	0.13	3		55.5	0.000000
17 Atmosphere	8	0.35	4,118		54.6	0.000000
18 Klinger	3	0.13	4		54.3	0.000000
19 Daisyworld	3	0.13	4		54.3	0.000000
20 Feedbacks	3	0.13	8		51	0.000000
21 Earth's	5	0.22	478		50.8	0.000000
22 Daisies	4	0.18	128		49.3	0.000000
23 Basalt	3	0.13	34		43	0.000000
24 Carnivore	3	0.13	40		42.1	0.000000
25 Sediments	3	0.13	43		41.7	0.000000
26 Phosphate	3	0.13	59		39.8	0.000000
27 Atmospheric	4	0.18	505		38.4	0.000000
28 Interactions	3	0.13	82		37.9	0.000000
29 Carbon	5	0.22	1,755		37.9	0.000000
30 Droplets	3	0.13	88		37.5	0.000000
31 Greenhouse	4	0.18	608		36.9	0.000000
32 Bogs	3	0.13	114		36	0.000000
33 Temperature	5	0.22	2,352		35	0.000000
34 Hotter	3	0.13	155		34.1	0.000000
35 Climate	5	0.22	3,259		31.8	0.000000
36 Hypothesis	3	0.13	262		31	0.000000
37 Continents	3	0.13	274		30.7	0.000000
38 Oceans	3	0.13	289		30.4	0.000000
39 Formed	5	0.22	4,154		29.4	0.000000
40 Says	13	0.57	76,773	0.08	28.6	0.000000
41 Regulate	3	0.13	491		27.3	0.000000
42 Life	11	0.48	56,382	0.06	26.9	0.000000
43 Model	5	0.22	6,255		25.4	0.000000
44 Regulation	4	0.18	2,754		25	0.000001
45 Solar	3	0.13	746		24.8	0.000001

Sample Text C6: 960811

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Jupiter	22	0.81	432		284.5	0.000000
2 Galileo	15	0.55	130		217.6	0.000000
3 Earth	27	0.99	6,524		215.3	0.000000
4 Probe	14	0.52	637		157.9	0.000000
5 Planet	14	0.52	1,860		128.2	0.000000
6 Ganymede	6	0.22	11		103.5	0.000000
7 Io	7	0.26	126		91.7	0.000000
8 Voyager	6	0.22	74		83	0.000000
9 Jovian	5	0.18	19		80.1	0.000000
10 Solar	8	0.29	746		78.8	0.000000
11 Jupiter's	5	0.18	45		72.1	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
12 Moon	9	0.33	2,573		68.7	0.000000
13 Asteroid	5	0.18	87		65.8	0.000000
14 Squyres	3	0.11	0		62.8	0.000000
15 Gravitational	5	0.18	153		60.3	0.000000
16 Orbit	6	0.22	511		60.2	0.000000
17 Antenna	4	0.15	60		53.8	0.000000
18 Hydrogen	5	0.18	306		53.4	0.000000
19 Moons	4	0.15	79		51.7	0.000000
20 Scientists	8	0.29	4,313		51	0.000000
21 Clouds	6	0.22	1,135		50.7	0.000000
22 Galileo's	3	0.11	14		46.9	0.000000
23 Slingshot	3	0.11	19		45.3	0.000000
24 Surface	7	0.26	4,088		43.6	0.000000
25 Atmosphere	7	0.26	4,118		43.5	0.000000
26 Pictures	7	0.26	4,831		41.3	0.000000
27 Sulphur	4	0.15	334		40.3	0.000000
28 Comets	3	0.11	75		37.4	0.000000
29 Venus	4	0.15	545		36.4	0.000000
30 Impacts	3	0.11	173		32.4	0.000000
31 Space	7	0.26	10,036	0.01	31.3	0.000000
32 Europa	3	0.11	228		30.8	0.000000
33 Astronomers	3	0.11	263		29.9	0.000000
34 Oceans	3	0.11	289		29.4	0.000000
35 Volcanic	3	0.11	296		29.2	0.000000
36 Planets	3	0.11	298		29.2	0.000000
37 Tidal	3	0.11	383		27.7	0.000000
38 Cloud	4	0.15	1,642		27.7	0.000000
39 Earth's	3	0.11	478		26.4	0.000000
40 Atmospheric	3	0.11	505		26	0.000000
41 Sun	6	0.22	9,602	0.01	25.6	0.000000
42 Ice	5	0.18	5,251		25.4	0.000000
43 Bombardment	3	0.11	589		25.1	0.000001
44 Launch	5	0.18	5,663		24.6	0.000001
45 Rocky	3	0.11	648		24.6	0.000001
46 Layers	3	0.11	650		24.5	0.000001

Sample Text C7: 960818

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Science	14	0.94	8,544		102.7	0.000000
2 Tulips	6	0.4	97		87.1	0.000000
3 Dutch	10	0.67	5,474		75.5	0.000000
4 Century	13	0.88	19,195	0.02	72.7	0.000000
5 Horticulturalists	3	0.2	0		66.4	0.000000
6 Breughel	4	0.27	22		66.2	0.000000
7 Mendel	3	0.2	19		48.9	0.000000
8 Genetics	4	0.27	278		46.6	0.000000
9 Philosophy	5	0.34	2,856		37.3	0.000000
10 Dulwich	3	0.2	286		33	0.000000
11 Painting	4	0.27	3,186		27.2	0.000000
12 Natural	5	0.34	8,049		27.1	0.000000
13 Breeding	3	0.2	1,045		25.3	0.000000

Sample Text C8: 960825

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Body	13	1.27	16,271	0.02	86.4	0.000000
2 Anatomists	3	0.29	4		59.1	0.000000
3 Images	6	0.58	3,565		48.7	0.000000
4 Spore	3	0.29	30		48.5	0.000000
5 Gastric	3	0.29	55		45	0.000000
6 Patient's	4	0.39	548		44.1	0.000000
7 Magnified	3	0.29	127		40.1	0.000000
8 Imaging	3	0.29	133		39.8	0.000000
9 Scans	3	0.29	154		38.9	0.000000
10 Physicians	3	0.29	228		36.6	0.000000
11 Cat	4	0.39	1,921		34.1	0.000000
12 Cell	4	0.39	2,267		32.8	0.000000
13 Cells	4	0.39	2,300		32.7	0.000000
14 Technicians	3	0.29	497		32	0.000000
15 Rays	3	0.29	596		30.9	0.000000
16 Organs	3	0.29	599		30.8	0.000000
17 Voyage	3	0.29	606		30.8	0.000000
18 X	4	0.39	3,002		30.6	0.000000
19 Imagery	3	0.29	643		30.4	0.000000
20 Blood	5	0.49	8,302		30.4	0.000000
21 Surgeons	3	0.29	737		29.6	0.000000
22 Renaissance	3	0.29	990		27.8	0.000000
23 Layer	3	0.29	1,148		27	0.000000

Sample Text C9: 960901

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Malaria	50	1.83	389		735.5	0.000000
2 Parasite	14	0.51	121		203	0.000000
3 Mosquito	14	0.51	150		197.3	0.000000
4 Plasmodium	7	0.26	3		134.2	0.000000
5 Eradication	8	0.29	86		112.6	0.000000
6 Disease	15	0.55	5,285		108.2	0.000000
7 Chloroquine	6	0.22	8		106.4	0.000000
8 Resistance	11	0.4	3,450		81.8	0.000000
9 Malarial	5	0.18	29		76.2	0.000000
10 Sinden	5	0.18	38		73.7	0.000000
11 Vaccine	7	0.26	485		73	0.000000
12 Bites	7	0.26	647		69	0.000000
13 Quinine	4	0.15	21		61.7	0.000000
14 Mosquitoes	5	0.18	142		61	0.000000
15 Gene	7	0.26	1,269		59.7	0.000000
16 Vivax	3	0.11	1		58.3	0.000000
17 Malariae	3	0.11	1		58.3	0.000000
18 Blood	10	0.37	8,302		55.2	0.000000
19 Salivary	3	0.11	4		53.2	0.000000
20 Professor	10	0.37	9,243		53.2	0.000000
21 Drug	10	0.37	9,590	0.01	52.4	0.000000
22 Insect	5	0.18	339		52.4	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
23 Tropical	6	0.22	1,041		51.7	0.000000
24 Insect's	3	0.11	11		48.2	0.000000
25 Antigen	3	0.11	12		47.7	0.000000
26 Protein	5	0.18	558		47.4	0.000000
27 Drugs	9	0.33	9,201		46.1	0.000000
28 Proteins	4	0.15	236		43	0.000000
29 Cases	9	0.33	13,328	0.01	39.6	0.000000
30 Ddt	3	0.11	65		38.2	0.000000
31 Resistant	4	0.15	525		36.7	0.000000
32 Effective	7	0.26	7,013		36.1	0.000000
33 Saliva	3	0.11	116		34.8	0.000000
34 Cells	5	0.18	2,300		33.4	0.000000
35 Spread	6	0.22	5,207		32.6	0.000000
36 Eradicated	3	0.11	169		32.5	0.000000
37 Fever	4	0.15	1,116		30.7	0.000000
38 p-value	7	0.26	10,594	0.01	30.5	0.000000
39 Mouse	4	0.15	1,196		30.1	0.000000
40 America	7	0.26	11,919	0.01	29	0.000000
41 Deaths	5	0.18	4,003		28	0.000000
42 Endemic	3	0.11	443		26.8	0.000000

Sample Text C10: 960908

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Names	26	1.46	9,379		208.6	0.000000
2 Iau	8	0.45	1		167.9	0.000000
3 Astronomers	9	0.51	263		117	0.000000
4 Minor	11	0.62	3,716		89.6	0.000000
5 Asteroid	6	0.34	87		86.2	0.000000
6 Planets	7	0.39	298		85.8	0.000000
7 Name	15	0.84	20,659	0.02	80.6	0.000000
8 Marsden	6	0.34	220		75.3	0.000000
9 Comet	6	0.34	389		68.5	0.000000
10 Naming	6	0.34	610		63.2	0.000000
11 Astronomical	5	0.28	210		61.4	0.000000
12 Planet	7	0.39	1,860		60.3	0.000000
13 Discovery	7	0.39	2,349		57.1	0.000000
14 Comets	4	0.22	75		55.4	0.000000
15 Celestial	4	0.22	163		49.3	0.000000
16 Asteroids	3	0.17	52		42	0.000000
17 Shoemaker	3	0.17	126		36.8	0.000000
18 Carolyn	3	0.17	163		35.3	0.000000
19 Named	6	0.34	6,772		34.6	0.000000
20 Austen	3	0.17	245		32.9	0.000000
21 Objects	4	0.22	1,856		30	0.000000
22 Venus	3	0.17	545		28.1	0.000000
23 Moon	4	0.22	2,573		27.5	0.000000
24 Honour	4	0.22	2,866		26.6	0.000000
25 Solar	3	0.17	746		26.2	0.000000

Sample Text C11: 960915

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Pneumoniae	13	0.69	1		274.5	0.000000
2 Heart	29	1.55	13,918	0.01	213.2	0.000000
3 Arteries	10	0.53	203		136	0.000000
4 Disease	16	0.85	5,285		129.3	0.000000
5 Artery	8	0.43	172		107.9	0.000000
6 Bug	9	0.48	478		105.4	0.000000
7 Blood	14	0.75	8,302		96.9	0.000000
8 Antibiotics	8	0.43	359		96.3	0.000000
9 C	14	0.75	12,973	0.01	84.6	0.000000
10 Plaque	7	0.37	357		82.5	0.000000
11 Infection	8	0.43	1,269		76.3	0.000000
12 Gupta	5	0.27	62		72.8	0.000000
13 Pylori	4	0.21	8		71.4	0.000000
14 Immune	7	0.37	908		69.6	0.000000
15 Atherosclerosis	3	0.16	0		65	0.000000
16 Factors	8	0.43	3,075		62.2	0.000000
17 Cholesterol	5	0.27	228		60	0.000000
18 Clotting	4	0.21	68		55.8	0.000000
19 Macrophages	3	0.16	8		52.1	0.000000
20 Chlamydia	3	0.16	16		48.4	0.000000
21 Scarring	3	0.16	60		40.9	0.000000
22 Researchers	5	0.27	2,432		36.6	0.000000
23 Attack	8	0.43	16,184	0.02	36.2	0.000000
24 System	10	0.53	34,124	0.04	35.3	0.000000
25 Proteins	3	0.16	236		32.8	0.000000
26 Respiratory	3	0.16	252		32.4	0.000000
27 Coronary	3	0.16	267		32	0.000000
28 Healthy	4	0.21	2,997		25.8	0.000000
29 Smoking	4	0.21	3,036		25.7	0.000000
30 Walls	4	0.21	3,399		24.8	0.000001
31 Key	6	0.32	14,978	0.02	24.7	0.000001

Sample Text C12: 960922

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Bracken	37	2.27	309		577.7	0.000000
2 Spores	13	0.8	95		206.1	0.000000
3 Ticks	5	0.31	87		70.9	0.000000
4 Bac	5	0.31	125		67.4	0.000000
5 Louping	3	0.18	0		65.9	0.000000
6 Cancers	5	0.31	294		59	0.000000
7 Uplands	4	0.25	99		54	0.000000
8 Reservoirs	4	0.25	153		50.6	0.000000
9 Carcinogens	3	0.18	25		46.8	0.000000
10 Autumn	7	0.43	5,849		45.7	0.000000
11 Water	9	0.55	20,146	0.02	41.5	0.000000
12 Digestive	3	0.18	138		36.8	0.000000
13 Tract	3	0.18	191		34.9	0.000000
14 Filters	3	0.18	204		34.5	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
15 Spreading	4	0.25	1,185		34.3	0.000000
16 Walking	5	0.31	3,632		34	0.000000
17 Cattle	4	0.25	1,284		33.7	0.000000
18 Lyme	3	0.18	247		33.4	0.000000
19 Animals	5	0.31	4,249		32.5	0.000000
20 Grants	4	0.25	2,478		28.5	0.000000
21 Areas	6	0.37	13,877	0.01	27.2	0.000000
22 Dairy	3	0.18	871		25.9	0.000000
23 Iii	3	0.18	905		25.6	0.000000
24 Humans	3	0.18	1,127		24.3	0.000001
25 Plants	4	0.25	4,347		24	0.000001

Sample Text C13: 960929

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Taylor	16	0.8	9,241		109.6	0.000000
2 Sexual	12	0.6	8,259		78	0.000000
3 Sex	13	0.65	12,654	0.01	75.7	0.000000
4 Females	6	0.3	533		63.3	0.000000
5 Female	9	0.45	6,141		58.6	0.000000
6 Ice	8	0.4	5,251		52.7	0.000000
7 Figurine	3	0.15	17		47.7	0.000000
8 Reproduction	4	0.2	373		41.8	0.000000
9 Male	7	0.35	7,596		39.2	0.000000
10 Evidence	9	0.45	21,968	0.02	36.4	0.000000
11 Humans	4	0.2	1,127		33.1	0.000000
12 Age	8	0.4	19,625	0.02	32.2	0.000000
13 Hormones	3	0.15	285		31.3	0.000000
14 Plants	5	0.25	4,347		30.2	0.000000
15 Induce	3	0.15	381		29.5	0.000000
16 Ancestors	3	0.15	485		28.1	0.000000
17 Urine	3	0.15	544		27.4	0.000000
18 Penis	3	0.15	691		26	0.000000
19 Breasts	3	0.15	720		25.7	0.000000
20 Graves	3	0.15	730		25.7	0.000000
21 Carved	3	0.15	756		25.5	0.000000
22 Species	4	0.2	3,322		24.5	0.000001

Sample Text C14: 961006

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Mithen	17	0.87	2		354.2	0.000000
2 Mind	23	1.17	14,268	0.02	155.3	0.000000
3 Evolutionary	9	0.46	298		113	0.000000
4 Tools	11	0.56	1,227		111.6	0.000000
5 Cognitive	8	0.41	135		111	0.000000
6 Mithen's	5	0.25	0		107.9	0.000000
7 Modules	6	0.31	165		77.5	0.000000
8 Domain	6	0.31	533		63.6	0.000000
9 Axes	5	0.25	164		62.8	0.000000
10 Modularity	3	0.15	2		58	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
11 Modern	10	0.51	12,705	0.01	53.4	0.000000
12 Hand	11	0.56	21,323	0.02	49.7	0.000000
13 Intelligence	8	0.41	6,719		49.2	0.000000
14 Axe	5	0.25	1,107		43.9	0.000000
15 Humans	5	0.25	1,127		43.7	0.000000
16 Archaeology	4	0.2	308		43.5	0.000000
17 Minds	6	0.31	3,159		42.4	0.000000
18 Integrate	4	0.2	385		41.8	0.000000
19 Fluidity	3	0.15	101		37.6	0.000000
20 Module	3	0.15	128		36.2	0.000000
21 Specialised	4	0.2	806		35.9	0.000000
22 Evolution	4	0.2	832		35.6	0.000000
23 Archaeologist	3	0.15	185		34	0.000000
24 Cathedral	4	0.2	1,622		30.3	0.000000
25 Archaeological	3	0.15	343		30.3	0.000000
26 Analogy	3	0.15	390		29.5	0.000000
27 Human	7	0.36	17,037	0.02	28.6	0.000000
28 Psychologists	3	0.15	524		27.8	0.000000
29 Stone	5	0.25	5,796		27.6	0.000000
30 Knowledge	5	0.25	6,294		26.8	0.000000
31 Populations	3	0.15	683		26.2	0.000000
32 Evolved	3	0.15	694		26.1	0.000000
33 Swiss	4	0.2	2,880		25.8	0.000000
34 Social	8	0.41	31,974	0.03	25.3	0.000000

Sample Text C15: 961013

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Protease	16	0.72	7		313.3	0.000000
2 Hiv	29	1.31	2,924		293.5	0.000000
3 Viral	13	0.59	173		183.2	0.000000
4 Virus	18	0.81	1,930		179.9	0.000000
5 Aids	21	0.95	5,637		171.6	0.000000
6 Inhibitors	9	0.41	35		147.5	0.000000
7 Cells	11	0.5	2,300		95.3	0.000000
8 Zidovudine	5	0.23	9		88.5	0.000000
9 Treatment	13	0.59	10,153	0.01	78.7	0.000000
10 Eradication	5	0.23	86		68	0.000000
11 Drugs	11	0.5	9,201		65.1	0.000000
12 Nucleoside	3	0.14	0		64	0.000000
13 Inhibitor	4	0.18	24		62.4	0.000000
14 Cancer	9	0.41	5,087		60.2	0.000000
15 Tissue	6	0.27	682		59.2	0.000000
16 Replication	4	0.18	39		58.7	0.000000
17 Therapy	7	0.32	1,943		56.7	0.000000
18 Levels	10	0.45	9,632	0.01	56.5	0.000000
19 Vancouver	5	0.23	295		55.9	0.000000
20 Infection	6	0.27	1,269		51.8	0.000000
21 Analogues	3	0.14	7		51.8	0.000000
22 Bloodstream	4	0.18	96		51.8	0.000000
23 Drug	9	0.41	9,590	0.01	49	0.000000
24 Ho	5	0.23	721		47	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
25 Undetectable	3	0.14	35		43	0.000000
26 Combination	6	0.27	3,288		40.5	0.000000
27 Waxman	3	0.14	58		40.1	0.000000
28 Load	5	0.23	1,470		39.9	0.000000
29 Scientists	6	0.27	4,313		37.3	0.000000
30 Immune	4	0.18	908		34	0.000000
31 Enzymes	3	0.14	195		32.9	0.000000
32 Proteins	3	0.14	236		31.8	0.000000
33 Triple	4	0.18	1,238		31.5	0.000000
34 Patient's	3	0.14	548		26.8	0.000000
35 Licensed	3	0.14	724		25.1	0.000001

Sample Text C16: 961020

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Cassava	20	1.25	55		353	0.000000
2 Genes	16	1	814		193.9	0.000000
3 Gene	12	0.75	1,269		128	0.000000
4 Crops	11	0.69	899		123	0.000000
5 Ishikazi	4	0.25	0		88	0.000000
6 Pests	6	0.38	166		79.9	0.000000
7 Fats	6	0.38	214		76.9	0.000000
8 Tropical	7	0.44	1,041		69.9	0.000000
9 Temperate	5	0.31	113		68.6	0.000000
10 Nishizawa	3	0.19	0		66	0.000000
11 Crop	7	0.44	1,452		65.3	0.000000
12 Tumefaciens	3	0.19	1		61.5	0.000000
13 Plants	8	0.5	4,347		59.3	0.000000
14 Genetic	6	0.38	1,466		54	0.000000
15 Resistant	5	0.31	525		53.4	0.000000
16 Saturated	4	0.25	304		45.3	0.000000
17 Resistance	6	0.38	3,450		43.8	0.000000
18 Plant	7	0.44	6,910		43.6	0.000000
19 Biodegradable	3	0.19	50		42.9	0.000000
20 Staple	4	0.25	442		42.3	0.000000
21 Engineering	6	0.38	4,893		39.7	0.000000
22 Varieties	4	0.25	760		38	0.000000
23 Tropics	3	0.19	123		37.6	0.000000
24 Bacterium	3	0.19	130		37.3	0.000000
25 Freezing	4	0.25	968		36.1	0.000000
26 Pesticide	3	0.19	163		35.9	0.000000
27 Cereals	3	0.19	186		35.2	0.000000
28 Natural	6	0.38	8,049		33.8	0.000000
29 Scientists	5	0.31	4,313		32.5	0.000000
30 Temperatures	4	0.25	1,533		32.4	0.000000
31 Grown	5	0.31	4,395		32.3	0.000000
32 Insect	3	0.19	339		31.6	0.000000
33 Plastics	3	0.19	353		31.3	0.000000
34 Food	7	0.44	17,450	0.02	31	0.000000
35 Tobacco	4	0.25	2,406		28.8	0.000000
36 Dna	3	0.19	990		25.2	0.000001

Sample Text C17: 961027

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Plankton	20	1.06	45		353.1	0.000000
2 Coccolithophores	8	0.42	0		173.3	0.000000
3 Ocean	15	0.79	1,505		156.6	0.000000
4 Oceans	10	0.53	289		129	0.000000
5 Dms	7	0.37	51		108.9	0.000000
6 Nutrients	8	0.42	173		107.7	0.000000
7 Phytoplankton	6	0.32	15		104.8	0.000000
8 Carbon	10	0.53	1,755		93.3	0.000000
9 Diatoms	5	0.26	6		93.1	0.000000
10 Drawdown	4	0.21	4		75.5	0.000000
11 Nutrient	5	0.26	47		75.4	0.000000
12 Oceanography	4	0.21	20		65	0.000000
13 Bloom	6	0.32	605		62.6	0.000000
14 Solutes	3	0.16	1		60.5	0.000000
15 Silicate	3	0.16	1		60.5	0.000000
16 Sea	10	0.53	9,940	0.01	58.9	0.000000
17 Water	12	0.64	20,146	0.02	58.5	0.000000
18 Osmo	3	0.16	4		55.4	0.000000
19 Species	7	0.37	3,322		51.5	0.000000
20 Surface	7	0.37	4,088		48.6	0.000000
21 Ecosystems	3	0.16	59		40.9	0.000000
22 Woods	5	0.26	1,979		38.5	0.000000
23 Global	6	0.32	5,921		35.4	0.000000
24 Dioxide	4	0.21	1,036		34.2	0.000000
25 Climate	5	0.26	3,259		33.6	0.000000
26 Warming	4	0.21	1,149		33.4	0.000000
27 Iron	5	0.26	3,642		32.5	0.000000
28 Atmosphere	5	0.26	4,118		31.3	0.000000
29 Viruses	3	0.16	378		29.9	0.000000
30 Abundance	3	0.16	489		28.4	0.000000
31 Dense	3	0.16	559		27.6	0.000000
32 Bacteria	3	0.16	753		25.8	0.000000

Sample Text C18: 961103

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Eye	29	2.82	8,525		276.5	0.000000
2 Tracking	8	0.78	505		100.6	0.000000
3 Drivers	8	0.78	3,590		69.4	0.000000
4 Information	8	0.78	18,647	0.02	43.4	0.000000
5 Frame	5	0.49	2,601		41.9	0.000000
6 Professor	6	0.58	9,243		37.4	0.000000
7 Providing	4	0.39	5,039		26.5	0.000000

Sample Text C19: 961110

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Mars	49	2.29	951		658.9	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
2 Martian	15	0.7	102		231.6	0.000000
3 Surface	25	1.17	4,088		230.6	0.000000
4 Surveyor	9	0.42	252		114.4	0.000000
5 Pathfinder	7	0.33	120		95.7	0.000000
6 Lander	6	0.28	213		73.5	0.000000
7 Orbiter	4	0.19	9		69.6	0.000000
8 Marsquakes	3	0.14	0		64.2	0.000000
9 Craft	7	0.33	1,261		63.2	0.000000
10 Orbit	6	0.28	511		63.1	0.000000
11 Spacecraft	5	0.23	182		61	0.000000
12 Missions	6	0.28	743		58.6	0.000000
13 Sojourner	3	0.14	2		57.5	0.000000
14 Earth	9	0.42	6,524		56.4	0.000000
15 Landers	3	0.14	3		55.9	0.000000
16 Rocks	6	0.28	1,090		54	0.000000
17 Launch	8	0.37	5,663		50.5	0.000000
18 Mission	7	0.33	3,765		48	0.000000
19 Planet	6	0.28	1,860		47.7	0.000000
20 Atmosphere	7	0.33	4,118		46.8	0.000000
21 Weather	7	0.33	6,510		40.5	0.000000
22 Nasa	4	0.19	418		40.4	0.000000
23 Solar	4	0.19	746		35.8	0.000000
24 Life	12	0.56	56,382	0.06	32.5	0.000000
25 Experiments	4	0.19	1,195		32.1	0.000000
26 Russian	7	0.33	13,498	0.01	30.6	0.000000
27 Robot	3	0.14	335		29.9	0.000000
28 Landing	4	0.19	1,612		29.7	0.000000
29 Global	5	0.23	5,921		26.5	0.000000
30 Shield	3	0.14	894		24.1	0.000001

Sample Text C20: 961117

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
1 Leptin	22	0.85	0		462.9	0.000000
2 Fat	24	0.93	3,407		219.1	0.000000
3 Obesity	12	0.47	130		170.2	0.000000
4 Ob	10	0.39	66		151.2	0.000000
5 Obese	9	0.35	135		122	0.000000
6 Weight	16	0.62	5,074		120.5	0.000000
7 Gene	12	0.47	1,269		116.5	0.000000
8 Genes	9	0.35	814		90.2	0.000000
9 Mice	8	0.31	443		87.9	0.000000
10 Bmi	5	0.19	8		87.8	0.000000
11 Energy	14	0.54	9,264		85.1	0.000000
12 Npy	4	0.16	0		84.1	0.000000
13 Calories	6	0.23	312		66.7	0.000000
14 Brain	9	0.35	3,591		63.7	0.000000
15 Receptor	4	0.16	29		59.8	0.000000
16 Levels	10	0.39	9,632	0.01	53.4	0.000000
17 Hypothalamus	3	0.12	7		50.9	0.000000
18 Diabetes	4	0.16	271		42.4	0.000000
19 Metabolic	3	0.12	67		38.3	0.000000

Keywords	*freq.*	*%*	*ref. freq.*	*%*	*keyness*	*p-value*
20 Body	9	0.35	16,271	0.02	37.2	0.000000
21 Protein	4	0.16	558		36.6	0.000000
22 Eat	6	0.23	4,106		36.1	0.000000
23 Intake	4	0.16	667		35.2	0.000000
24 Drug	7	0.27	9,590	0.01	32.6	0.000000
25 Appetite	4	0.16	1,053		31.6	0.000000
26 Proteins	3	0.12	236		30.9	0.000000
27 Genetic	4	0.16	1,466		29	0.000000
28 Burn	4	0.16	1,491		28.8	0.000000
29 Body's	3	0.12	361		28.3	0.000000
30 Blood	6	0.23	8,302		27.8	0.000000
31 Drugs	6	0.23	9,201		26.7	0.000000
32 Food	7	0.27	17,450	0.02	24.7	0.000001

Appendix 13. Ratios of key words to all words in the Theme and Rheme areas of Group B Sample Texts

Sample Text B1: 950319

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Intelligence	7	9	16	12.64	8.96	10.26
2. Apes	5	5	10	9.03	4.98	6.41
3. Byrne	9	0	9	16.25	0	5.77
4. Socia	3	6	9	5.42	5.97	5.77
5. Chimps	3	4	7	5.42	3.98	4.49
6. Gorillas	4	3	7	7.22	2.99	4.49
7. Ape	4	2	6	7.22	1.99	3.85
8. Complex	0	6	6	0	5.97	3.85
9. Groups	2	4	6	3.61	3.98	3.85
10. Human	4	2	6	7.22	1.99	3 85
11. Orangutans	4	2	6	7.22	1.99	3 85
12. Species	2	4	6	3.61	3.98	3 85
13. Ancestors	3	2	5	5.42	1.99	3.21
14. Behaviour	1	4	5	1.81	3.98	3.21
15. Humans	2	3	5	3.61	2.99	3.21
16. Primate	2	3	5	3.61	2.99	3.21
17. Primates	3	2	5	5.42	1.99	3.21
18. Understanding	0	5	5	0	4.98	3.21
19. Animals	1	3	4	1.81	2.99	2.57
20. Monkeys	2	2	4	3.61	1.99	2.57
21. Orangutan	3	1	4	5.42	1	2.57
22. Requires	0	4	4	0	3.98	2.57
23. Sheep	2	2	4	3.61	1.99	2.57
24. Borneo	2	1	3	3.61	1	1.92
25. Evolution	1	2	3	1.81	1.99	1.92
26. Evolved	0	3	3	0	2.99	1.92
27. Imitation	1	2	3	1.81	1.99	1.92
Total	*70*	*86*	*156*	*126.35*	*85.57*	*100.06*

Sample Text B2: 950514

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Asymmetry	3	1	4	5.43	1.2	2.89
2. Chemical	2	2	4	3.62	2.41	2.89
3. Chiral	3	5	8	5.43	6.02	5.79
4. Chirality	1	2	3	1.81	2.41	2.17
5. Company	5	3	8	9.06	3.61	5.79
6. Davies	8	4	12	14.49	4.82	8.68
7. Davies's	3	2	5	5.43	2.41	3.62
8. Drug	4	1	5	7.25	1.2	3.62
9. Enzymes	4	3	7	7.25	3.61	5.07
10. Forms	0	4	4	0	4.82	2.89
11. Handed	2	8	10	3.62	9.64	7.24
12. Handedness	3	2	5	5.43	2.41	3.62
13. Left	2	11	13	3.62	13.25	9.41
14. Molecular	0	3	3	0	3.61	2.17
15. Molecule	3	0	3	5.43	0	2.17
16. Molecules	8	6	14	14.49	7.23	10.13
17. Nature	4	1	5	7.25	1.2	3.62
18. Oxford	7	5	12	12.68	6.02	8.68
19. Palm	4	0	4	7.25	0	2.89
20. Pheromone	1	2	3	1.81	2.41	2.17
21. Reactions	3	0	3	5.43	0	2.17
22. Research	1	5	6	1.81	6.02	4.34
23. Right	2	7	9	3.62	8.43	6.51
24. Scaffolding	0	3	3	0	3.61	2.17
25. Synthetic	1	2	3	1.81	2.41	2.17
Total	*74*	*82*	*156*	*134.06*	*98.8*	*112.88*

Sample Text B3: 950709

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. America	6	4	10	11.83	4.15	6.8
2. Ancestors	1	2	3	1.97	2.08	2.04
3. Animals	5	9	14	9.86	9.35	9.52
4. Climate	5	2	7	9.86	2.08	4.76
5. Creatures	1	3	4	1.97	3.12	2.72
6. Elephants	1	2	3	1.97	2.08	2.04
7. Equator	1	2	3	1.97	2.08	2.04
8. Extinct	1	3	4	1.97	3.12	2.72
9. Extinctions	2	2	4	3.94	2.08	2.72
10. Hoofed	1	2	3	1.97	2.08	2.04
11. Mammals	2	3	5	3.94	3.12	3.4
12. Mass	1	5	6	1.97	5.19	4.08
13. Native	5	1	6	9.86	1.04	4.08
14. North	1	6	7	1.97	6.23	4.76
15. Northerners	3	0	3	5.92	0	2.04
16. Pleistocene	4	1	5	7.89	1.04	3.4
17. Pliocene	2	1	3	3.94	1.04	2.04
18. Sloths	5	4	9	9.86	4.15	6.12
19. South	4	9	13	7.89	9.35	8.84
20. Southerners	3	0	3	5.92	0	2.04
21. Species	9	3	12	17.75	3.12	8.16
Total	*63*	*64*	*127*	*124.26*	*66.46*	*86.39*

Sample Text B4: 950820

	Key words	*Key words (x1,000)/All words*

	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Ancient	0	4	4	0	4.83	3.33
2. Blackbird	2	1	3	5.33	1.21	2.49
3. Cooksonia	1	2	3	2.67	2.42	2.49
4. Cyanobacteria	3	0	3	8	0	2.49
5. Evolution	4	3	7	10.67	3.62	5.82
6. Evolved	1	3	4	2.67	3.62	3.33
7. Exhibition	2	2	4	5.33	2.42	3.33
8. Extinct	1	3	4	2.67	3.62	3.33
9. Ferns	3	0	3	8	0	2.49
10. Giant	0	4	4	0	4.83	3.33
11. Kew	2	3	5	5.33	3.62	4.16
12. Kew's	3	1	4	8	1.21	3.33
13. Landscape	2	3	5	5.33	3.62	4.16
14. Lonsdale	4	1	5	10.67	1.21	4.16
15. Mosses	1	3	4	2.67	3.62	3.33
16. Mud	0	3	3	0	3.62	2.49
17. Period	2	4	6	5.33	4.83	4.99
18. Plants	5	4	9	13.33	4.83	7.48
19. Primeval	1	2	3	2.67	2.42	2.49
20. Rocks	5	1	6	13.33	1.21	4.99
21. Stems	0	3	3	0	3.62	2.49
22. Visitors	1	4	5	2.67	4.83	4.16
Total	*43*	*54*	*97*	*114.67*	*65.22*	*80.63*

Sample Text B5: 950903

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Accident	1	5	6	1.26	3.77	2.83
2. Accidents	2	5	7	2.51	3.77	3.3
3. Affect	0	4	4	0	3.02	1.89
4. Chemicals	5	4	9	6.28	3.02	4.24
5. Death	2	10	12	2.51	7.54	5.66
6. Hse	5	2	7	6.28	1.51	3.3
7. Industry	2	7	9	2.51	5.28	4.24
8. Nuclear	8	14	22	10.05	10.56	10.37
9. Paling	2	1	3	2.51	0.75	1.41
10. Plant	5	1	6	6.28	0.75	2.83
11. Public	13	4	17	16.33	3.02	8.01
12. Public's	2	1	3	2.51	0.75	1.41
13. Radiation	1	3	4	1.26	2.26	1.89
14. Risk	12	25	37	15.08	18.85	17.44
15. Risks	5	18	23	6.28	13.57	10.84
16. Smoking	4	3	7	5.03	2.26	3.3
Total	*73*	*109*	*182*	*91.71*	*82.2*	*85.77*

Sample Text B6: 951210

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Agent	3	3	6	4.69	3.25	3.84
2. Beef	2	3	5	3.13	3.25	3.2
3. Brain	0	4	4	0	4.33	2.56
4. Bruce's	1	2	3	1.56	2.16	1.92
5. Bse	9	7	16	14.08	7.58	10.24
6. Cjd	4	2	6	6.26	2.16	3.84
7. Cows	1	3	4	1.56	3.25	2.56
8. Disease	1	4	5	1.56	4.33	3.2
9. Experiment	3	2	5	4.69	2.16	3.2
10. Farmers	2	3	5	3.13	3.25	3.2

11. Humans	1	3	4	1.56	3.25	2.56
12. Infect	2	1	3	3.13	1.08	1.92
13. Positive	2	3	5	3.13	3.25	3.2
14. Problem	4	3	7	6.26	3.25	4.48
15. Protein	1	2	3	1.56	2.16	1.92
16. Scientific	2	7	9	3.13	7.58	5.76
17. Scientists	5	1	6	7.82	1.08	3.84
18. Scrapie	4	1	5	6.26	1.08	3.2
19. Species	2	2	4	3.13	2.16	2.56
20. Transmitted	0	3	3	0	3.25	1.92
Total	*49*	*59*	*108*	*76.68*	*63.85*	*69.1*

Sample Text B7: 960407

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Blood	3	2	5	6.88	2.62	4.17
2. Cells	4	2	6	9.17	2.62	5.01
3. Coat	3	0	3	6.88	0	2.5
4. Coating	2	1	3	4.59	1.31	2.5
5. Enzyme	1	2	3	2.29	2.62	2.5
6. Fold	1	2	3	2.29	2.62	2.5
7. Glucose	8	5	13	18.35	6.56	10.85
8. Hbv	3	0	3	6.88	0	2.5
9. Hepatitis	1	2	3	2.29	2.62	2.5
10. Mannose	1	2	3	2.29	2.62	2.5
11. Molecules	2	2	4	4.59	2.62	3.34
12. Proteins	2	1	3	4.59	1.31	2.5
13. Sugar	2	2	4	4.59	2.62	3.34
14. Sugars	7	11	18	16.06	14.44	15.03
15. Sugoid	6	1	7	13.76	1.31	5.84
16. Sugoids	3	6	9	6.88	7.87	7.51
17. Toxic	1	4	5	2.29	5.25	4.17
18. Virus	5	2	7	11.47	2.62	5.84
19. Viruses	2	1	3	4.59	1.31	2.5
Total	*57*	*48*	*105*	*130.73*	*62.99*	*87.65*

Sample Text B8: 960519

	Key words			*Key words (x1,000)/All words*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Astronomers	2	1	3	3.93	1.12	2.14
2. Cerro	7	3	10	13.75	3.37	7.14
3. Chile	4	0	4	7.86	0	2.86
4. Chilean	2	2	4	3.93	2.24	2.86
5. Instrument	2	1	3	3.93	1.12	2.14
6. Light	2	7	9	3.93	7.86	6.43
7. Mirror	3	2	5	5.89	2.24	3.57
8. Mirrors	1	2	3	1.96	2.24	2.14
9. Observation	2	1	3	3.93	1.12	2.14
10. Observatory	4	2	6	7.86	2.24	4.29
11. Paranal	5	2	7	9.82	2.24	5
12. Space	2	4	6	3.93	4.49	4.29
13. Telescope	4	4	8	7.86	4.49	5.71
14. Tolerance	0	3	3	0	3.37	2.14
15. Vlt	2	3	5	3.93	3.37	3.57
Total	*42*	*37*	*79*	*82.51*	*41.53*	*56.43*

Sample Text B9: 960526

	Key words			Key words (x1,000)/All words		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Believes	0	5	5	0	6.83	4.15
2. Card	0	10	10	0	13.66	8.29
3. Customers	3	2	5	6.33	2.73	4.15
4. Digital	5	1	6	10.55	1.37	4.98
5. Electronic	2	3	5	4.22	4.1	4.15
6. Icl	3	0	3	6.33	0	2.49
7. Internet	0	4	4	0	5.46	3.32
8. List	3	3	6	6.33	4.1	4.98
9. Loyalty	1	4	5	2.11	5.46	4.15
10. Pc	1	3	4	2.11	4.1	3.32
11. Retailers	3	1	4	6.33	1.37	3.32
12. Screen	1	3	4	2.11	4.1	3.32
13. Shopping	10	8	18	21.1	10.93	14.93
14. Smart	0	4	4	0	5.46	3.32
15. Store	4	6	10	8.44	8.2	8.29
16. Stored	0	4	4	0	5.46	3.32
17. Stores	3	1	4	6.33	1.37	3.32
18. Supermarket	3	5	8	6.33	6.83	6.63
19. Technology	4	3	7	8.44	4.1	5.8
Total	*46*	*70*	*116*	*97.05*	*95.63*	*96.19*

Sample Text B10: 960623

	Key words			Key words (x1,000)/All words		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Countryside	3	5	8	4	4.9	4.52
2. Example	2	6	8	2.67	5.88	4.52
3. Experiences	1	3	4	1.33	2.94	2.26
4. Humans	3	1	4	4	0.98	2.26
5. Landscapes	1	4	5	1.33	3.92	2.82
6. Natural	1	5	6	1.33	4.9	3.39
7. Nature	15	4	19	20	3.92	10.73
8. Sensory	2	2	4	2.67	1.96	2.26
9. Study	9	0	9	12	0	5.08
10. Urban	3	2	5	4	1.96	2.82
11. Wilderness	3	0	3	4	0	1.69
Total	*43*	*32*	*75*	*57.33*	*31.37*	*42.37*

Sample Text B11: 960630

	Key words			Key words (x1,000)/All words		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1.Ape	0	3	3	0	2.55	1.65
2.Apes	13	10	23	20.31	8.51	12.67
3.Booee	3	1	4	4.69	0.85	2.2
4.Chimp	3	5	8	4.69	4.26	4.41
5.Chimpanzee	2	1	3	3.13	0.85	1.65
6.Chimpanzees	4	5	9	6.25	4.26	4.96
7.Chimps	7	1	8	10.94	0.85	4.41
8.Communicate	0	3	3	0	2.55	1.65
9.Communicating	0	3	3	0	2.55	1.65
10.Food	3	9	12	4.69	7.66	6.61
11.Fouts	4	0	4	6.25	0	2.2
12.Human	1	7	8	1.56	5.96	4.41
13.Kanzi	2	2	4	3.13	1.7	2.2
14.Katharine	3	1	4	4.69	0.85	2.2
15.Language	12	10	22	18.75	8.51	12.12
16.Panbanisha	3	0	3	4.69	0	1.65
17.Portions	0	4	4	0	3.4	2.2

18.Primate	2	1	3	3.13	0.85	1.65
19.Professor	12	1	13	18.75	0.85	7.16
20.Research	3	4	7	4.69	3.4	3.86
21.Rivas	3	0	3	4.69	0	1.65
22.Rudimentary	0	3	3	0	2.55	1.65
23.Rumbaugh	5	1	6	7.81	0.85	3.31
24.Savage	6	1	7	9.38	0.85	3.86
25.Sign	6	6	12	9.38	5.11	6.61
26.Signing	3	2	5	4.69	1.7	2.75
27.Use	4	10	14	6.25	8.51	7.71
28.Washoe	3	0	3	4.69	0	1.65
29.Wild	5	2	7	7.81	1.7	3.86
30.Words	2	5	7	3.13	4.26	3.86
Total	*114*	*101*	*215*	*178.13*	*85.96*	*118.46*

Appendix 14. Keyness in the Theme and Rheme areas of Group B Sample Texts

Sample Text B1: 950319

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Intelligence	7	9	16	55.87	71.83	127.7
2. Apes	5	5	10	73.15	73.15	146.3
3. Byrne	9	0	9	104.60	0.00	104.6
4. Social	3	6	9	11.43	22.87	34.3
5. Chimps	3	4	7	43.89	58.51	102.4
6. Gorillas	4	3	7	59.89	44.91	104.8
7. Ape	4	2	6	51.80	25.90	77.7
8. Complex	0	6	6	0.00	37.70	37.7
9. Groups	2	4	6	9.17	18.33	27.5
10. Human	4	2	6	16.93	8.47	25.4
11. Orangutans	4	2	6	80.53	40.27	120.8
12. Species	2	4	6	14.83	29.67	44.5
13. Ancestors	3	2	5	32.64	21.76	54.4
14. Behaviour	1	4	5	5.78	23.12	28.9
15. Humans	2	3	5	18.44	27.66	46.1
16. Primate	2	3	5	26.84	40.26	67.1
17. Primates	3	2	5	44.16	29.44	73.6
18. Understanding	0	5	5	0.00	31.60	31.6
19. Animals	1	3	4	6.13	18.38	24.5
20. Monkeys	2	2	4	23.20	23.20	46.4
21. Orangutan	3	1	4	57.83	19.28	77.1
22. Requires	0	4	4	0.00	27.70	27.7
23. Sheep	2	2	4	15.45	15.45	30.9
24. Borneo	2	1	3	25.27	12.63	37.9
25. Evolution	1	2	3	8.80	17.60	26.4
26. Evolved	0	3	3	0.00	27.50	27.5
27. Imitation	1	2	3	10.00	20.00	30.0
Total	*70*	*86*	*156*	*796.62*	*787.18*	*1,583.8*

Sample Text B2: 950514

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*

1. Asymmetry	3	1	4	47.48	15.83	63.3
2. Chemical	2	2	4	13.70	13.70	27.4
3. Chiral	3	5	8	66.86	111.44	178.3
4. Chirality	1	2	3	22.27	44.53	66.8
5. Company	5	3	8	15.75	9.45	25.2
6. Davies	8	4	12	60.47	30.23	90.7
7. Davies's	3	2	5	34.74	23.16	57.9
8. Drug	4	1	5	20.88	5.22	26.1
9. Enzymes	4	3	7	54.40	40.80	95.2
10. Forms	0	4	4	0.00	25.10	25.1
11. Handed	2	8	10	16.02	64.08	80.1
12. Handedness	3	2	5	41.64	27.76	69.4
13. Left	2	11	13	7.98	43.92	51.9
14. Molecular	0	3	3	0.00	32.40	32.4
15. Molecule	3	0	3	36.60	0.00	36.6
16. Molecules	8	6	14	107.66	80.74	188.4
17. Nature	4	1	5	21.60	5.40	27
18. Oxford	7	5	12	51.16	36.54	87.7
19. Palm	4	0	4	38.20	0.00	38.2
20. Pheromone	1	2	3	15.07	30.13	45.2
21. Reactions	3	0	3	25.70	0.00	25.7
22. Research	1	5	6	4.48	22.42	26.9
23. Right	2	7	9	6.33	22.17	28.5
24. Scaffolding	0	3	3	0.00	33.90	33.9
25. Synthetic	1	2	3	10.37	20.73	31.1
Total	*74*	*82*	*156*	*719.36*	*739.65*	*1,459*

Sample Text B3: 950709

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. America	6	4	10	36.18	24.12	60.3
2. Ancestors	1	2	3	10.00	20.00	30
3. Animals	5	9	14	43.68	78.62	122.3
4. Climate	5	2	7	39.43	15.77	55.2
5. Creatures	1	3	4	9.33	27.98	37.3
6. Elephants	1	2	3	9.97	19.93	29.9
7. Equator	1	2	3	12.60	25.20	37.8
8. Extinct	1	3	4	11.25	33.75	45
9. Extinctions	2	2	4	32.15	32.15	64.3
10. Hoofed	1	2	3	15.87	31.73	47.6
11. Mammals	2	3	5	23.80	35.70	59.5
12. Mass	1	5	6	6.25	31.25	37.5
13. Native	5	1	6	41.33	8.27	49.6
14. North	1	6	7	3.80	22.80	26.6
15. Northerners	3	0	3	38.50	0.00	38.5
16. Pleistocene	4	1	5	77.52	19.38	96.9
17. Pliocene	2	1	3	39.80	19.90	59.7
18. Sloths	5	4	9	94.89	75.91	170.8
19. South	4	9	13	16.58	37.32	53.9
20. Southerners	3	0	3	36.90	0.00	36.9
21. Species	9	3	12	80.25	26.75	107
Total	*63*	*64*	*127*	*680.08*	*586.53*	*1,266.6*

Sample Text B4: 950820

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Ancient	0	4	4	0.00	27.30	27.3
2. Blackbird	2	1	3	26.53	13.27	39.8
3. Cooksonia	1	2	3	22.57	45.13	67.7
4. Cyanobacteria	3	0	3	57.10	0.00	57.1

5. Evolution	4	3	7	44.00	33.00	77
6. Evolved	1	3	4	10.25	30.75	41
7. Exhibition	2	2	4	13.25	13.25	26.5
8. Extinct	1	3	4	11.65	34.95	46.6
9. Ferns	3	0	3	40.30	0.00	40.3
10. Giant	0	4	4	0.00	25.80	25.8
11. Kew	2	3	5	25.96	38.94	64.9
12. Kew's	3	1	4	60.53	20.18	80.7
13. Landscape	2	3	5	16.48	24.72	41.2
14. Lonsdale	4	1	5	58.16	14.54	72.7
15. Mosses	1	3	4	15.48	46.43	61.9
16. Mud	0	3	3	0.00	24.40	24.4
17. Period	2	4	6	10.20	20.40	30.6
18. Plants	5	4	9	41.06	32.84	73.9
19. Primeval	1	2	3	13.87	27.73	41.6
20. Rocks	5	1	6	50.75	10.15	60.9
21. Stems	0	3	3	0.00	27.90	27.9
22. Visitors	1	4	5	6.80	27.20	34
Total	*43*	*54*	*97*	*524.94*	*538.88*	*1,063.8*

Sample Text B5: 950903

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
Accident	1	5	6	6.08	30.42	36.5
Accidents	2	5	7	17.09	42.71	59.8
Affect	0	4	4	0.00	25.30	25.3
Chemicals	5	4	9	43.67	34.93	78.6
Death	2	10	12	8.60	43.00	51.6
Hse	5	2	7	61.64	24.66	86.3
Industry	2	7	9	6.78	23.72	30.5
Nuclear	8	14	22	54.11	94.69	148.8
Paling	2	1	3	31.33	15.67	47
Plant	5	1	6	26.92	5.38	32.3
Public	13	4	17	44.89	13.81	58.7
Public's	2	1	3	17.47	8.73	26.2
Radiation	1	3	4	8.20	24.60	32.8
Risk	12	25	37	92.85	193.45	286.3
Risks	5	18	23	47.67	171.63	219.3
Smoking	4	3	7	29.20	21.90	51.1
Total	*73*	*109*	*182*	*496.5*	*774.6*	*1,271.1*

Sample Text B6: 951210

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Agent	3	3	6	21.45	21.45	42.9
2. Beef	2	3	5	17.36	26.04	43.4
3. Brain	0	4	4	0.00	25.90	25.9
4. Bruce's	1	2	3	12.83	25.67	38.5
5. Bse	9	7	16	119.98	93.32	213.3
6. Cjd	4	2	6	58.67	29.33	88
7. Cows	1	3	4	9.53	28.58	38.1
8. Disease	1	4	5	6.14	24.56	30.7
9. Experiment	3	2	5	24.06	16.04	40.1
10. Farmers	2	3	5	13.60	20.40	34
11. Humans	1	3	4	8.75	26.25	35
12. Infect	2	1	3	25.13	12.57	37.7
13. Positive	2	3	5	11.24	16.86	28.1
14. Problem	4	3	7	15.54	11.66	27.2
15. Protein	1	2	3	9.57	19.13	28.7
16. Scientific	2	7	9	16.20	56.70	72.9

17. Scientists	5	1	6	34.50	6.90	41.4
18. Scrapie	4	1	5	60.32	15.08	75.4
19. Species	2	2	4	13.25	13.25	26.5
20. Transmitted	0	3	3	0.00	26.90	26.9
Total	*49*	*59*	*108*	*478.12*	*516.59*	*994.7*

Sample Text B7: 960407

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Blood	3	2	5	17.34	11.56	28.9
2. Cells	4	2	6	34.73	17.37	52.1
3. Coat	3	0	3	26.00	0.00	26
4. Coating	2	1	3	25.13	12.57	37.7
5. Enzyme	1	2	3	12.50	25.00	37.5
6. Fold	1	2	3	9.00	18.00	27
7. Glucose	8	5	13	130.09	81.31	211.4
8. Hbv	3	0	3	59.40	0.00	59.4
9. Hepatitis	1	2	3	11.53	23.07	34.6
10. Mannose	1	2	3	22.57	45.13	67.7
11. Molecules	2	2	4	22.50	22.50	45
12. Proteins	2	1	3	23.67	11.83	35.5
13. Sugar	2	2	4	14.50	14.50	29
14. Sugars	7	11	18	116.82	183.58	300.4
15. Sugoid	6	1	7	135.43	22.57	158
16. Sugoids	3	6	9	67.70	135.40	203.1
17. Toxic	1	4	5	10.28	41.12	51.4
18. Virus	5	2	7	46.64	18.66	65.3
19. Viruses	2	1	3	21.80	10.90	32.7
Total	*57*	*48*	*105*	*807.63*	*695.07*	*1,502.7*

Sample Text B8: 960519

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Astronomers	2	1	3	22.60	11.30	33.9
2. Cerro	7	3	10	133.00	57.00	190
3. Chile	4	0	4	39.30	0.00	39.3
4. Chilean	2	2	4	21.75	21.75	43.5
5. Instrument	2	1	3	16.00	8.00	24
6. Light	2	7	9	11.18	39.12	50.3
7. Mirror	3	2	5	19.38	12.92	32.3
8. Mirrors	1	2	3	9.50	19.00	28.5
9. Observation	2	1	3	16.47	8.23	24.7
10. Observatory	4	2	6	52.80	26.40	79.2
11. Paranal	5	2	7	99.64	39.86	139.5
12. Space	2	4	6	10.93	21.87	32.8
13. Telescope	4	4	8	50.70	50.70	101.4
14. Tolerance	0	3	3	0.00	27.00	27
15. Vlt	2	3	5	40.28	60.42	100.7
Total	*42*	*37*	*79*	*543.53*	*403 57*	*947.1*

Sample Text B9: 960526

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Believes	0	5	5	0.00	26.60	26.6
2. Card	0	10	10	0.00	74.30	74.3
3. Customers	3	2	5	16.92	11.28	28.2
4. Digital	5	1	6	46.83	9.37	56.2

5. Electronic	2	3	5	15.08	22.62	37.7
6. Icl	3	0	3	31.40	0.00	31.4
7. Internet	0	4	4	0.00	34.60	34.6
8. List	3	3	6	16.45	16.45	32.9
9. Loyalty	1	4	5	8.62	34.48	43.1
10. Pc	1	3	4	6.73	20.18	26.9
11. Retailers	3	1	4	25.43	8.48	33.9
12. Screen	1	3	4	6.08	18.23	24.3
13. Shopping	10	8	18	98.11	78.49	176.6
14. Smart	0	4	4	0.00	30.20	30.2
15. Store	4	6	10	35.40	53.10	88.5
16. Stored	0	4	4	0.00	39.50	39.5
17. Stores	3	1	4	21.53	7.18	28.7
18. Supermarket	3	5	8	28.39	47.31	75.7
19. Technology	4	3	7	24.63	18.47	43.1
Total	*46*	*70*	*116*	*381.6*	*550.84*	*932.4*

Sample Text B10: 960623

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Countryside	3	5	8	25.61	42.69	68.3
2. Example	2	6	8	9.55	28.65	38.2
3. Experiences	1	3	4	7.30	21.90	29.2
4. Humans	3	1	4	25.58	8.53	34.1
5. Landscapes	1	4	5	10.62	42.48	53.1
6. Natural	1	5	6	5.43	27.17	32.6
7. Nature	15	4	19	113.29	30.21	143.5
8. Sensory	2	2	4	25.40	25.40	50.8
9. Study	9	0	9	51.80	0.00	51.8
10. Urban	3	2	5	19.50	13.00	32.5
11. Wilderness	3	0	3	26.80	0.00	26.8
Total	*43*	*32*	*75*	*320.88*	*240.03*	*560.9*

Sample Text B11: 960630

	Key words			*keyness*		
	Th	*Rh*	*Overall*	*Th*	*Rh*	*Overall*
1. Ape	0	3	3	0.00	33.80	33.8
2. Apes	13	10	23	206.87	159.13	366
3. Booee	3	1	4	65.18	21.73	86.9
4. Chimp	3	5	8	46.91	78.19	125.1
5. Chimpanzee	2	1	3	25.53	12.77	38.3
6. Chimpanzees	4	5	9	59.82	74.78	134.6
7. Chimps	7	1	8	102.11	14.59	116.7
8. Communicate	0	3	3	0.00	25.50	25.5
9. Communicating	0	3	3	0.00	31.90	31.9
10. Food	3	9	12	15.68	47.03	62.7
11. Fouts	4	0	4	86.90	0.00	86.9
12. Human	1	7	8	4.49	31.41	35.9
13. Kanzi	2	2	4	36.75	36.75	73.5
14. Katharine	3	1	4	32.48	10.83	43.3
15. Language	12	10	22	89.51	74.59	164.1
16. Panbanisha	3	0	3	65.20	0.00	65.2
17. Portions	0	4	4	0.00	47.60	47.6
18. Primate	2	1	3	24.20	12.10	36.3
19. Professor	12	1	13	79.57	6.63	86.2
20. Research	3	4	7	12.77	17.03	29.8
21. Rivas	3	0	3	49.40	0.00	49.4
22. Rudimentary	0	3	3	0.00	34.20	34.2
23. Rumbaugh	5	1	6	97.50	19.50	117
24. Savage	6	1	7	56.14	9.36	65.5

25. Sign	6	6	12	38.35	38.35	76.7
26. Signing	3	2	5	21.54	14.36	35.9
27. Use	4	10	14	17.40	43.50	60.9
28. Washoe	3	0	3	56.90	0.00	56.9
29. Wild	5	2	7	33.36	13.34	46.7
30. Words	2	5	7	9.09	22.71	31.8
Total	*114*	*101*	*215*	*1,333.65*	*931.68*	*2,265.3*

Appendix 15. Ratios of keyness to all words in the Theme and Rheme areas of Group B Sample Texts

Sample Text B1: 950319

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Intelligence	10.08	7.15	8.19
2. Apes	13.20	7.28	9.38
3. Byrne	18.88	0.00	6.71
4. Social	2.06	2.28	2.20
5. Gorillas	10.81	4.47	6.72
6. Chimps	7.92	5.82	6.57
7. Orangutans	14.54	4.01	7.75
8. Ape	9.35	2.58	4.98
9. Species	2.68	2.95	2.85
10. Complex	0.00	3.75	2.42
11. Groups	1.66	1.82	1.76
12. Human	3.06	0.84	1.63
13. Primates	7.97	2.93	4.72
14. Primate	4.85	4.01	4.30
15. Ancestors	5.89	2.17	3.49
16. Humans	3.33	2.75	2.96
17. Understanding	0.00	3.14	2.03
18. Behaviour	1.04	2.30	1.85
19. Orangutan	10.44	1.92	4.95
20. Monkeys	4.19	2.31	2.98
21. Sheep	2.79	1.54	1.98
22. Requires	0.00	2.76	1.78
23. Animals	1.11	1.83	1.57
24. Borneo	4.56	1.26	2.43
25. Imitation	1.81	1.99	1.92
26. Evolved	0.00	2.74	1.76
27. Evolution	1.59	1.75	1.69
Total	*143.79*	*78.30*	*101.59*

Sample Text B2: 950514

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Asymmetry	8.60	1.91	4.58
2. Chemical	2.48	1.65	1.98
3. Chiral	12.11	13.43	12.90
4. Chirality	4.03	5.37	4.83
5. Company	2.85	1.14	1.82
6. Davies	10.95	3.64	6.56
7. Davies's	6.29	2.79	4.19
8. Drug	3.78	0.63	1.89
9. Enzymes	9.86	4.92	6.89
10. Forms	0.00	3.02	1.82
11. Handed	2.90	7.72	5.80
12. Handedness	7.54	3.34	5.02
13. Left	1.45	5.29	3.76
14. Molecular	0.00	3.90	2.34
15. Molecule	6.63	0.00	2.65
16. Molecules	19.50	9.73	13.63
17. Nature	3.91	0.65	1.95
18. Oxford	9.27	4.40	6.35
19. Palm	6.92	0.00	2.76
20. Pheromone	2.73	3.63	3.27
21. Reactions	4.66	0.00	1.86
22. Research	0.81	2.70	1.95
23. Right	1.15	2.67	2.06
24. Scaffolding	0.00	4.08	2.45
25. Synthetic	1.88	2.50	2.25
Total	*130.32*	*89.11*	*105.57*

Sample Text B3: 950709

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. America	7.14	2.50	4.10
2. Ancestors	1.97	2.08	2.04
3. Animals	8.62	8.16	8.32
4. Climate	7.78	1.64	3.76
5. Creatures	1.84	2.91	2.54
6. Elephants	1.97	2.07	2.03
7. Equator	2.49	2.62	2.57
8. Extinct	2.22	3.50	3.06
9. Extinctions	6.34	3.34	4.37
10. Hoofed	3.13	3.29	3.24
11. Mammals	4.69	3.71	4.05
12. Mass	1.23	3.25	2.55
13. Native	8.15	0.86	3.37
14. North	0.75	2.37	1.81
15. Northerners	7.59	0.00	2.62
16. Pleistocene	15.29	2.01	6.59
17. Pliocene	7.85	2.07	4.06
18. Sloths	18.72	7.88	11.62
19. South	3.27	3.88	3.67
20. Southerners	7.28	0.00	2.51
21. Species	15.83	2.78	7.28
Total	*134.14*	*60.91*	*86.16*

Sample Text B4: 950820

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Ancient	0.00	3.30	2.27
2. Blackbird	7.07	1.60	3.31
3. Cooksonia	6.02	5.45	5.63
4. Cyanobacteria	15.23	0.00	4.75
5. Evolution	11.73	3.99	6.40
6. Evolved	2.73	3.71	3.41
7. Exhibition	3.53	1.60	2.20
8. Extinct	3.11	4.22	3.87
9. Ferns	10.75	0.00	3.35
10. Giant	0.00	3.12	2.14
11. Kew	6.92	4.70	5.39
12. Kew's	16.14	2.44	6.71
13. Landscape	4.39	2.99	3.42
14. Lonsdale	15.51	1.76	6.04
15. Mosses	4.13	5.61	5.15
16. Mud	0.00	2.95	2.03
17. Period	2.72	2.46	2.54
18. Plants	10.95	3.97	6.14
19. Primeval	3.70	3.35	3.46
20. Rocks	13.53	1.23	5.06
21. Stems	0.00	3.37	2.32
22. Visitors	1.81	3.29	2.83
Total	*139.98*	*65.08*	*88.43*

Sample Text B5: 950903

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Accident	0.76	2.29	1.72
2. Accidents	2.15	3.22	2.82
3. Affect	0.00	1.91	1.19
4. Chemicals	5.49	2.63	3.70
5. Death	1.08	3.24	2.43
6. Hse	7.74	1.86	4.07
7. Industry	0.85	1.79	1.44
8. Nuclear	6.80	7.14	7.01
9. Paling	3.94	1.18	2.21
10. Plant	3.38	0.41	1.52
11. Public	5.64	1.04	2.77
12. Public's	2.19	0.66	1.23
13. Radiation	1.03	1.86	1.55
14. Risk	11.66	14.59	13.49
15. Risks	5.99	12.94	10.33
16. Smoking	3.67	1.65	2.41
Total	*62.37*	*58.42*	*59.90*

Sample Text B6: 951210

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Agent	3.36	2.32	2.74
2. Beef	2.72	2.82	2.78
3. Brain	0.00	2.80	1.66
4. Bruce's	2.01	2.78	2.46
5. Bse	18.78	10.10	13.65
6. Cjd	9.18	3.17	5.63
7. Cows	1.49	3.09	2.44

8. Disease	0.96	2.66	1.96
9. Experiment	3.77	1.74	2.57
10. Farmers	2.13	2.21	2.18
11. Humans	1.37	2.84	2.24
12. Infect	3.93	1.36	2.41
13. Positive	1.76	1.82	1.80
14. Problem	2.43	1.26	1.74
15. Protein	1.50	2.07	1.84
16. Scientific	2.54	6.14	4.66
17. Scientists	5.40	0.75	2.65
18. Scrapie	9.44	1.63	4.82
19. Species	2.07	1.43	1.70
20. Transmitted	0.00	2.91	1.72
Total	*74.82*	*55.91*	*63.64*

Sample Text B7: 960407

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Blood	3.98	1.52	2.41
2. Cells	7.97	2.28	4.35
3. Coat	5.96	0.00	2.17
4. Coating	5.76	1.65	3.15
5. Enzyme	2.87	3.28	3.13
6. Fold	2.06	2.36	2.25
7. Glucose	29.84	10.67	17.65
8. Hbv	13.62	0.00	4.96
9. Hepatitis	2.64	3.03	2.89
10. Mannose	5.18	5.92	5.65
11. Molecules	5.16	2.95	3.76
12. Proteins	5.43	1.55	2.96
13. Sugar	3.33	1.90	2.42
14. Sugars	26.79	24.09	25.08
15. Sugoid	31.06	2.96	13.19
16. Sugoids	15.53	17.77	16.95
17. Toxic	2.36	5.40	4.29
18. Virus	10.70	2.45	5.45
19. Viruses	5.00	1.43	2.73
Total	*185.24*	*91.22*	*125.43*

Sample Text B8: 960519

Key words	*Keyness to all words (x100)*		
	Theme	*Rheme*	*Overall*
1. Astronomers	4.44	1.27	2.42
2. Cerro	26.13	6.40	13.57
3. Chile	7.72	0.00	2.81
4. Chilean	4.27	2.44	3.11
5. Instrument	3.14	0.90	1.71
6. Light	2.20	4.39	3.59
7. Mirror	3.81	1.45	2.31
8. Mirrors	1.87	2.13	2.04
9. Observation	3.24	0.92	1.76
10. Observatory	10.37	2.96	5.66
11. Paranal	19.58	4.47	9.96
12. Space	2.15	2.45	2.34
13. Telescope	9.96	5.69	7.24
14. Tolerance	0.00	3.03	1.93
15. Vlt	7.91	6.78	7.19
Total	*106.78*	*45.29*	*67.65*

Sample Text B9: 960526

Key words	*Keyness to all words (×100)*		
	Theme	*Rheme*	*Overall*
1. Believes	0.00	3.63	2.21
2. Card	0.00	10.15	6.16
3. Customers	3.57	1.54	2.34
4. Digital	9.88	1.28	4.66
5. Electronic	3.18	3.09	3.13
6. Icl	6.62	0.00	2.60
7. Internet	0.00	4.73	2.87
8. List	3.47	2.25	2.73
9. Loyalty	1.82	4.71	3.57
10. Pc	1.42	2.76	2.23
11. Retailers	5.36	1.16	2.81
12. Screen	1.28	2.49	2.01
13. Shopping	20.70	10.72	14.64
14. Smart	0.00	4.13	2.50
15. Store	7.47	7.25	7.34
16. Stored	0.00	5.40	3.28
17. Stores	4.54	0.98	2.38
18. Supermarket	5.99	6.46	6.28
19. Technology	5.20	2.52	3.57
Total	*80.51*	*75.25*	*77.31*

Sample Text B10: 960623

Key words	*Keyness to all words (×100)*		
	Theme	*Rheme*	*Overall*
1. Countryside	3.41	4.19	3.86
2. Example	1.27	2.81	2.16
3. Experiences	0.97	2.15	1.65
4. Humans	3.41	0.84	1.93
5. Landscapes	1.42	4.16	3.00
6. Natural	0.72	2.66	1.84
7. Nature	15.11	2.96	8.11
8. Sensory	3.39	2.49	2.87
9. Study	6.91	0.00	2.93
10. Urban	2.60	1.27	1.84
11. Wilderness	3.57	0.00	1.51
Total	*42.78*	*23.53*	*31.69*

Sample Text B11: 960630

Key words	*Keyness to all words (×100)*		
	Theme	*Rheme*	*Overall*
1. Ape	0.00	2.88	1.86
2. Apes	32.32	13.54	20.17
3. Booee	10.18	1.85	4.79
4. Chimp	7.33	6.65	6.89
5. Chimpanzee	3.99	1.09	2.11
6. Chimpanzees	9.35	6.36	7.42
7. Chimps	15.95	1.24	6.43
8. Communicate	0.00	2.17	1.40
9. Communicating	0.00	2.71	1.76
10. Food	2.45	4.00	3.45
11. Fouts	13.58	0.00	4.79
12. Human	0.70	2.67	1.98
13. Kanzi	5.74	3.13	4.05
14. Katharine	5.08	0.92	2.39
15. Language	13.99	6.35	9.04

16. Panbanisha	10.19	0.00	3.59
17. Portions	0.00	4.05	2.62
18. Primate	3.78	1.03	2.00
19. Professor	12.43	0.56	4.75
20. Research	2.00	1.45	1.64
21. Rivas	7.72	0.00	2.72
22. Rudimentary	0.00	2.91	1.88
23. Rumbaugh	15.23	1.66	6.45
24. Savage	8.77	0.80	3.61
25. Sign	5.99	3.26	4.23
26. Signing	3.37	1.22	1.98
27. Use	2.72	3.70	3.36
28. Washoe	8.89	0.00	3.13
29. Wild	5.21	1.14	2.57
30. Words	1.42	1.93	1.75
Total	*208.38*	*79.29*	*124.81*