THE UNIVERSITY
of LIVERPOOL

Automatic identification of segments in written texts

A P Berber Sardinha

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Antonio Paulo Berber Sardinha.

September 1997

Declaration

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

Date: Signature:

Acknowledgements

Thanks to CNPq (Conselho Nacional de Pesquisa, Brasília, Brazil) I have had the chance to pursue the degree of PhD in English at the University of Liverpool.

My deepest gratitude goes to Prof Michael Hoey, my supervisor, for his brilliant insights, his assistance and encouragement during these four years.

Dr Michael Scott has been an inspiration to me long before I came to Liverpool; I have always looked up to him as a model. Thanks, Mike!

Nell Scott's assistance throughout these years has been so great I cannot thank her enough.

I want to thank Geoff Thompson for reading and commenting on a preliminary version of this thesis. I also want to thank both the internal and the external examiners for their criticisms. Any inadequacies that still remain are my responsibility.

Without Maureen Molloy's, Karin Alecock's, Gill Lester's, and Jill Yates's prompt and kind help I would not have been able to complete this work.

The staff at AELSU have always been very supportive; my thanks go to Dr Sue Thompson, Sarah Waite, Celia Shalom, and Keith Stuart.

My colleagues at AELSU have always provided much needed support. I want to thank them all (in alphabetical order): Alex, Alfred, Ali, Ching, Cigdem, Connie, David, Eman, Maria Stella, Mohammad, Paul, Puleng, and Zargham.

I thank Dr Rob Birch for taking a lot of his time to help me sort out the computer software used in the thesis. I also thank Kevin O'Donovan for actually writing the software for me. A big thanks goes to Peter Kulawec for making the program work when I needed it most.

Bob Gallop, David Alderton, Ellen Hertz, Dr Pralay Senchaudhuri, Sue Byrne, Suchada Supattathum, Tim Borryhill, and especially Matthew Zack, gave me good SAS advice and programming tips. Thanks!

I am grateful to Dr Tony Hak for teaching me about Pêcheux.

Most of all, I want to thank my wife, Marilisa Shimazumi; her unending support and dedication made this thesis possible. I thank my parents, Antonio and Leonor, my brother, Carlos, and my father and mother-in-law, Jorge and Cecília, for taking care of me and for always being there when I needed them most.

Abstract

Although computers are being used for language analysis more often, the majority of studies employing computers for language analysis are concerned with the analysis of corpora, where the interest lies not in individual texts but in collocation and word frequency. The use of computers enables the investigation of greater quantities of data; the analyses themselves are also more reliable. Using computers for the analysis of central issues in text research would allow for a better understanding of major features of texts. One particular issue which would benefit from computer-assisted analyses is text organisation. Typically, text organization is investigated in discourse analysis by means of the application of models which are aimed at uncovering the regularities in the constitution of the text. Models are essentially designed for hand analysis of single texts or short text fragments. An aspect which bears centrally on text organisation is segmentation, or the principled division of texts into constituents. Segmentation is also a fundamental aspect underlying models of discourse. The research reported in this thesis is aimed at developing a computer-assisted procedure for segmenting texts. A major problem with the implementation of computer-assisted procedures for text analysis is the choice of which linguistic feature to compute, since not all linguistic features are relevant to text analysis, and fewer still are amenable to computer treatment. A review of the relevant research indicated that a feature which is closely related to how texts function is lexical cohesion; therefore it was selected to serve as the basis for the computation of the segments. Another problem relates to how segmentation is to be evaluated. In order to achieve a more reliable evaluation, it was decided to compare the segmentation with an independent criterion. The best solution was to check to what extent segment boundaries matched typographical section divisions in the texts. The main segmentation procedure developed for the present investigation is called LSM ('Link Set Median'). It was applied to a corpus of 300 texts from three different genres. The results obtained by application of the LSM procedure on the corpus were then compared to segmentation carried out at random. Statistical analyses suggested that LSM significantly outperformed random segmentation, thus indicating that the segmentation was meaningful. Two other analyses focused on explaining why the segmentation worked. Multiple regression techniques were used to identify the textual characteristics which were significantly associated with better segmentation. Finally, an analysis based on logistic regression indicated that an important reason why the segmentation procedure worked was that it managed to predict section boundaries. The analysis revealed specific patterns of lexical cohesion which were significantly associated with the probability of sentences being section boundaries. The main contributions of the present study to discourse analysis, research in lexical cohesion, and computer-based models of segmentation are discussed. These include the suggestions that texts are segmentable by computer, that corpus analysis of individual texts is possible, and that segmentation can be achieved using meaningful linguistic criteria. Importantly, the present investigation provides further evidence that lexical cohesion relates to text organization, and original evidence that typographical sections are not arbitrary units.

To
Marilisa Shimazumi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many of the structural features of discourse are large scale and highly variable. As the units of language description get larger, the identification of meaningful units becomes more problematic. The computer is now available to help in this work.
(Sinclair, 1994, p.15)

The last ten years have seen an increase in the number of publications dealing with the use of computers in linguistic research (e.g. Barnbrook, 1996; Butler, 1992a; Hockey and Ide, 1994b,a; Lancashire, 1991; Landow and Delany, 1993; Stubbs, 1996). If the now vast body of research on corpus linguistics is added to that (e.g. Aarts and Meijs, 1990; Aijmer and Altenberg, 1991; Kyto et al., 1988; McEnery and Wilson, 1996; Sinclair, 1991; Svartvik, 1992), the list of publications which report the use of computers for analysis of language will reach thousands. However, a quick look at the titles of most of these works will reveal a worrying shortage of investigations dealing

with text organization. The literature on corpus linguistics is geared to the investigation of linguistic phenomena occurring in large bodies of naturally occurring textual data ('corpora'), but not necessarily in individual texts. And research in linguistic computing concerns various aspects of individual texts but not necessarily text organization. There is, therefore, a need for large-scale computer-aided linguistic research on text organization.

In what follows, the argument will be presented that despite the fact that there are several disciplines which use computers for the analysis of language data, there has been very little interest in the analysis of individual texts from a discourse analysis perspective. A discourse analysis perspective is one in which 'the basic unit of analysis is text' (Scott, 1997; Georgakopoulou and Goutsos, 1997, p.5), and which is geared to answering the question of how texts are organized. It will also be argued (on page 15) that the kind of computer-based analysis which can address questions relating to text organization is *segmentation*, or the division of texts into discrete units.

1.1 Computers and language analysis

Although computers were originally designed to carry out complex numerical calculations, nowadays they are commonly used for handling non-numerical data, including natural language texts (Butler, 1992b, p.viii). A major discipline which makes use of computers for the analysis of language data is corpus linguistics. It is generally recognized that corpus linguistics as it is known today was introduced in Britain by John Sinclair (Hoey, 1993, p.v) and Geoffrey Leech (Svartvik, 1996, pp.4-5). Their research agendas have shaped the way computer-held corpora have been created and explored. Owing to the limitations of computers at the time when corpora were first being held in machine-readable form, both the size of corpora and the way they could be

explored were affected. For one thing, corpora had to be restricted to what could be accommodated on media storage devices at the time; for another, the kinds of analyses that were actually carried out on those data were also dictated by the ability of computers to cope with text processing. Over the more than 30 years which separate us from the ground-breaking work of Sinclair on lexis in corpora (Sinclair, 1966), the basic paradigm of analysis of computer-held corpora has not changed substantially: the focus still is on either the study of lexical co-occurrence in narrow contexts (collocation) or word frequencies *per se* (cf. Leech and Fligelstone, 1992). Writing about collocations, Scott (1997, p.235) argues that:

> By analysing words within a narrow immediate context of a few words to left and right (chiefly to the left), a very large number of valuable observations have been made ... about the English language in general, and its characteristic lexico-grammatical patterns. A quite different perspective arises if one starts from the category *text*.

In other words, as Scott (1997, p.235) goes on to explain, 'the starting-point to a considerable degree determines the Corpus Linguistics tools; these determine the kinds of patterns which can be found'. Kirk (1994, p.19) agrees stating that 'the methodology of corpus linguistics as a branch of linguistic enquiry is inseparable from the computer's resources not only to store the data but to sort, manipulate, calculate and transform them'. The study of collocation as a co-occurrence pattern within a nine-word stretch of text (four words on either side of a node) was clearly determined by the limited processing capacity of computers in the 1960's, since, according to Sinclair (1966, p.413), taking into account wider contexts 'would be difficult indeed'. Scott (1997) argues that with today's personal computers it becomes possible to investigate wider contexts and therefore to consider investigating co-occurrence at the level of whole texts. Despite its now being possible, few

researchers have actually taken on the challenge.

The most obvious reason why the study of collocation remains alive is that collocation, narrow or wide, is a truly fascinating research area in its own right; in addition, it continues to pose new questions to researchers as different and large corpora are created not only of English but of other languages as well. Still, the side-effect of the predominance of frequency or collocation research is that other important areas of research which need answers from corpora have stayed in the background.

Furthermore, the fact that collocation has been widely researched across various corpora does not imply that its role in organizing individual texts has been also investigated. On the contrary, the role of collocation in text constitution has received very little attention (cf. Benson and Greaves, 1992; Berber Sardinha, 1995c,b; Phillips, 1985). Most attention is focused on the relationship between collocation and idiomaticity (e.g. Smadja, 1992), and between collocation and language in general (Sinclair, 1991). In short, neither has text-wide collocation as seen by Firth (1957) and Scott (1997) become common in corpus analysis nor has narrow-span collocation *à la* Sinclair (1966) had an impact in text analysis. This is symptomatic of the difficulty methodologies face in crossing discipline boundaries.

A research interest which developed out of the greater availability of computers and machine-readable texts is what has become known as the field of 'Humanities Computing' (e.g. Hockey and Ide, 1994a,b). The discipline can now be traced back over 30 years ago (Raben, 1991); therefore it has developed alongside corpus linguistics. Researchers working in this field are concerned with the application of computer-processing to the analysis of individual texts or specific text collections. A particular interest is in describing literary work by computational means (e.g. Harris, 1989; Miall, 1992; Thury, 1988). Regular conferences have been held for many years which

have gathered specialists in humanities computing working in various parts of the world and in a variety of subjects, and two journals serve the research community ('Literary and Linguistic Computing' and 'Computers and the Humanities').

However, by far the main discipline devoted to the analysis of language by computational means is computational linguistics. The branch of computational linguistics devoted to the analysis of naturally-occurring language, including discourse, is known as 'Natural Language Processing' (NLP). This is a vast field which includes empirical investigation of issues dealt with at a theoretical level only in non-computational linguistics, such as computer assessment of incoherence (Donaldson et al., 1996), identification of rhetorical relations (Knott and Dale, 1993), and recognizing puns and riddles (Binsted, 1994), as well as research into issues that are avoided by other linguists such as topic identification in discourse (Chen, 1995; Fisher, 1994).

A major interest of researchers investigating discourse in NLP is 'document analysis and retrieval' (Grosz, 1995, p.227), which concerns the analysis of texts in order to help users retrieve texts from databases. This is also a vast field, and the number of studies which would be potentially informative to the non-computational linguist are many. The key figure in the field is Gerard Salton, who, together with associates, has experimented widely with different retrieval techniques (e.g. Salton, 1988; Salton and Buckley, 1991; Salton et al., 1994, 1990).

1.2 Segmentation

The term segmentation is used in the research presented in this thesis to refer to the division of written texts into discrete units. Segmentation is not restricted to the division of discourse, though. The general meaning of the

term is 'division', and so it can be used to mean the division of language into phonemes, morphemes, and syntactic groups. As such, segmentation could be seen as a central activity in disciplines such as phonology, morphology, and syntax (Crystal, 1991, pp.308-309).

The segmentation of discourse has been addressed by a number of disciplines, including psychology, semiotics, business management, sociolinguistics, computational linguistics, text linguistics, and discourse analysis. For instance, in psychology, Thorndyke (1977) proposes a *schema* for story comprehension comprising setting, theme, plot and resolution, and further subdivisions of each schematic component. Working from a semiotic perspective, Barthes (1977, p.101) suggests that certain discoursal activities are arranged in *sequences*; for example, a telephone call could be viewed as comprising the following sequences: telephone ringing, picking up the receiver, speaking, and putting down the receiver. In management, Lewis (1996, p.116) proposes a division of Japanese business meetings into the following phases: platitudinous preamble, outline of subjects to be discussed, airing of views, replies of each party to each other's views, and summary by both sides. In sociolinguistics, Labov and Waletzky (1967) describe narratives of personal experience as comprising an abstract, an orientation, a complication, an evaluation, a resolution, and a coda. In computational linguistics, Hearst (1994a) and Beeferman et al. (1997), among others, devise *mathematical algorithms* for programming computers to identify segments.

In discourse analysis and text linguistics, discourse has been commonly described by means of *structures*, *models*, or *patterns*. Discourse analysis and text linguistics are by far the disciplines which have given more attention to the organisation of discourse. To mention just a few of the studies available in the literature on discourse models and patterns, Hasan (1977) describes medical appointment making as comprising an identification, an application,

an offer, and a confirmation, and service encounters as consisting of a sale request, a sale compliance, a sale, a purchase, a purchase closure (Hasan, 1989). Hoey (1983) and Jordan (1984) discuss various kinds of texts as being organised in terms of a Situation – Problem – Solution – Evaluation pattern. Van Dijk (1980) argues that narratives can be adequately accounted for by a *superstructure* consisting of a plot (in turn made up of a setting and episodes), and a moral. Bhatia (1993) describes job application letters as having the following *moves*: (1) establishing credentials, (2) introducing the offer, (3) offering incentives, (4) enclosing documents, (5) soliciting response, (6) using pressure tactics, and (7) ending politely. And Swales (1990) suggests research article introductions comprise the following moves: (1) establishing a territory, (2) establishing a niche, and (3) occupying the niche.

Although many disciplines address discourse segmentation, few make use of the actual expression 'segment' (or 'segmentation'). In computational linguistics the term has been used frequently (e.g. Grosz and Sidner, 1986; Kozima and Furugori, 1993; Hearst, 1994b; Beeferman et al., 1997). Some cognitive psychologists have made use of the term 'segmentation marker' (Bestgen and Costermans, 1997; Bestgen and Vonk, 1995; Ehrich and Koster, 1983); In other disciplines mention of 'segments' or 'segmentation' has been rarer (e.g. Burke, 1991; Cloran, 1995; Fries, 1990; Giora, 1983; Goutsos, 1996a; Hinds, 1979; Lamprecht, 1988).

Several discourse analysts and text linguists have expressed their views on the existence of discourse segments without referring to 'segments' or 'discourse models'. Kukharenko (1979, p.235) observes that long texts are constituted by sentence clusters, or 'semantic topical and lexico-grammatical unities of two or more sentences'. Langleben (1979) shares the same view as Kukharenko (1979), and Scinto (1986, pp.113–114) proposes that sentence clusters are common to texts of all lengths, and describes what he

calls 'combinatoric text modules', or multi-sentential units sharing a particular thematic progression pattern. Grimes (1975, p.91) refers to 'spans', or 'stretches of text within which there is some kind of uniformity'. Two basic factors which contribute to uniformity are theme and setting (Grimes, 1975, p.102–103). The segmentation of texts, in Grimes's (1975, p.334) view, seems to be most readily accounted for by the property of staging, or 'the dimension of prose structure which identifies the relative prominence given to various segments of prose discourse' (Clements, 1979, p.287). According to Hoey (1985, p.105), 'a paragraph might well be part of a larger sub-grouping or be involved in relations with one or more non-adjacent paragraphs', which suggests that paragraphs may form groupings spanning larger stretches of text. García-Berrio and Albadejo Mayordomo (1987) argue that segments are manifested on the surface of texts typographically as chapters, for instance. They believe chapters to be 'subtexts, smaller texts integrated into a greater one' (Garcia-Berrio and Mayordomo, 1987, p.198).

1.3 Discourse segments

Computational linguists have addressed the problem of finding discourse segments for many years; there is a general recognition that texts are segmentable:

> discourses divide into *discourse segments* much like sentences divide into phrases. Utterances group into segments, with the meaning of a segment encompassing more than the meaning of the individual parts. (Grosz, 1995, pp.227-228)

The problem of finding segments in computational linguistics has been tackled with very little input from discourse analysis. Discourse analysis has also received little input from information retrieval and NLP/computational linguistics (Sparck Jones, 1996). Perhaps one of the reasons is that much

work in NLP follows a syntactic paradigm (as even the previous quotation hints), which is strange to discourse analysis.

Importantly, a similar interest in discourse segments characterizes discourse analysis as a discipline:

> [a] basic tool of linguistic analysis is [to] segment text into sections, labels those sections as part of a structure, and assign functions to those sections. (Schiffrin, 1994, p.11)

Models of discourse provide a principled method for segmenting spoken and written texts. In other words, discourse models are concerned with showing 'regularities in the linguistic realizations used by people to communicate' (Brown and Yule, 1983, p.26) and how these regularities are expressed sequentially within texts. Sequentiality is central to discourse analysis according to Labov (1972, p.252), for whom 'the fundamental problem of discourse analysis is to show how one utterance follows another in a rational, rule-governed manner'. Schiffrin (1994) also places sequentiality at the top of the discourse analyst's list of priorities because:

> discourse (by definition) is comprised of sequentially arranged units, and because sequential regularities (sequences that fulfil our expectations) are a key ingredient in our identification of something as a text. (Schiffrin, 1994, p.63)

Segments embody these key characteristics, namely individual uniformity (Grimes, 1975, p.91) and sequentiality (Schiffrin, 1994, p.63). These two characteristics relate to the traditional concerns of discourse analysis, one of which is 'to discover how it is that discourse differs from random sequences' (Harris (1951), in Schiffrin (1994, p.18)). By showing that texts segment, one is also showing that the sequences in it are not random, but motivated by internal consistency and sequential arrangement.

The segmental view of discourse is only one of the possible perspectives on discourse organisation. Halliday (1978, p.188) distinguishes four types

of structure: constituent, recursive, prosodic, and culminative. Matthiessen (1988) also proposes four types of structure: constituency, interdependency, prosody, and prominence. The segmental view proposed here is associated by Martin (1992, p.549) with the constituent type of structure. Constituency can be understood as a type of representation 'with teleologically driven stages, working their way towards a goal' (Martin, 1992, p.550). In this sense, the segmental view of discourse draws on the notion of *stage*, as proposed by Mitchell (1957/1975, p.43), as an 'abstract category' which is employed to describe the order of textual elements.

The segmental view, to the extent that it relates to text constituency (Martin, 1992, p.549), also ties in with the *particulate* perspective proposed by Pike (1972), according to whom one can see discourse organisation from three complementary perspectives: particle, wave, and field. The particulate perspective is defined as that which sees texts as having 'bricks juxtaposed' in structure (Pike, 1972, p.130); the wave perspective views language as 'waves smearing into some kind of continuum whose prominent parts make up nuclei' (Pike and Pike, 1977, p.30); and in the field perspective 'we turn to sets of relationships which occur when units are linked to one another by their presence in some larger system' (Pike and Pike, 1977, p.30).

The principles of exhaustiveness and contiguity as discussed by Fries (1990) apply to the segmental view of text organisation. Fries (1990) identifies three different ways in which text organisation can be viewed: as 'large-scale strings of grammatical or semantic functions' (Gregory, 1985b; Hasan, 1977; Hoey, 1983; Jordan, 1984; Ventola, 1979), as 'large-scale immediate constituent tree structures' (e.g. Grimes, 1975; Mann and Thompson, 1986a; Pike, 1982); or as 'a non-exhaustive patterning of widely interspersed realizations of principled choices' (Fries, 1990, pp.363, 377). The first two perspectives have in common the notion of exhaustiveness, that is, 'once one

has finished assigning an appropriate structure to a text, then the entire text should be accounted for. No part of the text should be "left over" ' (Fries, 1990, p.363). In addition, they have in common the principle of contiguity. In string-based descriptions, contiguity is manifested by having 'all the portions of a text which realize a given function' appear contiguously in the description (Fries, 1990, p.363). And in tree-structure approaches, contiguity is revealed by keeping 'all immediate constituents of the same larger construction' contiguous to one another (Fries, 1990, p.363). By contrast, the third perspective is componential, that is, it assumes that certain aspects of text organisation are neither contiguous nor exhaustive; rather, certain 'phenomena do not affect ALL portions of a given text segment, and the phenomena may affect the language used at various locations distributed throughout a text segment.' (Fries, 1990, p.364). The componential approach to text organisation is useful to reveal how authors make 'principled choices at a number of disconnected points' in the text (Fries, 1990, p.377). For instance, Fries (1990) notes that the main argument in a particular letter was not expressed 'in a simple, explicit logical order as in a mathematical proof'; rather, the argument was evoked discontinuously across the text. The componential approach to text organisation is not favoured by the segmental view of discourse.

Segments have not been extracted computationally within a discourse analytical perspective. Discourse models are primarily meant to be used in manual analysis of single texts. The number of models for discourse description is enormous; the variety is so great that there have been volumes dedicated to comparing how different models account for the same data (Grimshaw, 1991; Mann and Thompson, 1992; van Dijk and Petöfi, 1977), The irony is that in discourse analysis very little data is actually analysed (Stubbs, 1996, p.129); as Phillips (1989) puts it:

> Linguistics has traditionally been restricted to the investigation of the extent of language that can comfortably be accommodated on the average blackboard. (Phillips, 1989, p.8)

The automatic extraction of segments would translate into an increase in the amount of data which can be handled. This would be desirable since it would enable discourse analysis to make objective statements 'based on language as it really is rather than statements which are subjective and based upon the individual's own internalised cognitive perception of the language' (McEnery and Wilson, 1996, p.87). Further, it would allow for an improvement in the 'quality of evidence' (Sinclair, 1991, p.4) presented for discourse phenomena. Extracting segments automatically would also enable the researcher to analyse texts without forcing too many *a priori* assumptions on the data. As Sinclair (1991, p.29) argues:

> Linguistics usually operates with ...abstract categories ...it is good policy to defer the use of them for as long as possible, to refrain from imposing analytical categories from the outside.

This is in accordance with Gregory (1985a), according to whom the imposition of structural categories such as discoursal schemes treats discourse as if it were rule-governed instead of patterned (cf. Di Pietro, 1983; Hoey, 1991b).

The automatic extraction of segments would also have two other important advantages. The first advantage has to do with comparability and evaluation, or 'how analyses of different texts can be replicated and compared' (Stubbs, 1996, p.129). The second is the possibility of analysing whole texts, instead of text fragments. As Biber (1995b, p.344) observes,

> Before the use of computers, empirical discourse analyses were typically based on a few thousand words of text; an analysis of 10,000 words was regarded as a major undertaking that required a long research period.

Standard references on discourse analysis normally restrict themselves to the presentation of fragments of texts or spoken interaction (e.g. Brown and Yule, 1983; Stubbs, 1983). Stubbs himself has come to realize that, and in a more recent book argues that not being able to handle whole texts 'poses problems of evidence and generalization' (Stubbs, 1996, p.129).

1.4 Computers and discourse analysis

In discourse analysis, only a few scholars have addressed the issue of analysis of individual texts by computer. Hoey (1995a) argues that important aspects of text organisation such as paragraphing can be better understood if collocation is taken into account. He observed the probability of certain phrases being paragraph-initial and then used those to predict the paragraph breaks of written texts. He found that the paragraph divisions in his target texts could be well accounted for by the presence of the paragraph-initial collocations drawn from the corpus. Hoey's (1995a) study suggests that information from a corpus can be fed back into the analysis of individual texts to aid in explaining certain aspects of text organisation.

Phillips (1985) used collocation to derive patterns of organisation of science texts. He argued that contemporary approaches to text organisation were inadequate because they relied on grammar to explain text. He set out to develop a methodology of text analysis which did not depend on grammatical categories, and which could be modelled on the computer. This would allow him to handle larger quantities of text without having to resort to manual analysis of the texts. Unlike Hoey (1995a), Phillips extracted collocations from within each text and not from a corpus. He then observed how collocations intercollocated, that is, how they connected to each other to form networks which he then mapped onto the existing chapter divisions

of the textbooks. He also observed that chapters shared collocations and hence formed groupings, or 'segments'. Phillips (1985) concluded that the network of collocations and the multi-chapter segments revealed the underlying macrostructure of the text. Importantly, Phillips (1985) argues that his findings would not be possible without the aid of the computer. This is a major feature of truly computational text analysis – the computer is not a tool for merely doing hand analysis faster; the computer is used as a tool for doing an analysis which would not be possible without it.

Another scholar who has carried out research of individual texts by computer is Stubbs (1996). According to him, 'the most powerful interpretation emerges if comparisons of texts across corpora are combined with the analysis of the organisation of individual texts' (Stubbs, 1996, p.34). He presents computer-assisted analyses of individual texts and their comparison to corpora. He argues that the use of computers in text analysis has the advantage of being replicable and comparable (p.131). Further, computer methodology allows for the discovery of patterns which are directly observable:

> Such patterns may be discernible, in a rough way, via intuition. But in order to describe such distributions systematically, significant amounts of text must be stored in a computer and searched, and quantitative methods must be used to describe the patterns. (Stubbs, 1996, p.131)

As the quotation above indicates, Stubbs (1996) also argues in favour of quantitative methods. According to him:

> When new quantitative methods are applied to very large amounts of data, they always do more than provide a mere summary. By transforming the data, they generate insight. (Stubbs, 1996, p.232)

In other words, quantitative treatment of textual data is part of the analysis and not an after-fact. This is because certain patterns are not 'cat-

egorical but probabilistic' (Stubbs, 1996, p.131), that is, they can only be perceived if cumulative evidence for them is found over large bodies of data.

In short, the situation sketched so far is that while computer-assisted analysis of texts is routine in information retrieval and humanities computing, in discourse analysis it is very rare. What has been argued is that it would be desirable to apply discourse analysis insights in computer-aided text analysis. This leads to the question of which textual features amenable to computer analysis are relevant for discourse analysis. As argued above, as far written texts are concerned, the central concern in discourse analysis is the sequential organisation of texts, which has typically been described in terms of discourse models. The problem with discourse models, as with most linguistic theory (Mann and Thompson, 1987a, p.42), is that they have not been designed with computational applications in mind. Hence, discourse models are not *a priori* adaptable for computer applications. The identification of segments, on the other hand, is a typical task of discourse analysis. As Hrebicek and Altmann (1993, p.1) put it, 'linguists intending to investigate texts always feel the need to solve questions like "In which parts is it to be segmented?"'. As previous research in information retrieval suggests, segmentation of texts can be carried out by computer, thus the automatic extraction of segments presents itself as a good candidate for computer-assisted analysis of discourse. Since segments are by definition discourse units, a computer-assisted description of texts based on segments would fit Sinclair's (1994, pp.14-15) requirement for a 'special model for discourse':

> We should build a model which emphasizes the distinctive features of discourse. A special model for discourse will offer an explanation of those features of discourse that are unique to it, or characteristic of it, or prominent in discourse but not elsewhere.

1.5 Aims

The major aim of the present study is the development of a computer-assisted segmentation procedure. The fundamental characteristic of such a procedure is that it should primarily be based on insights derived from research in discourse analysis and text linguistics, and only secondarily on considerations of computational feasibility.

The specific aims of the present investigation are:

1. The specification of discourse characteristics which can be used for analysing texts on the computer;

2. Experimentation with a variety of segmentation techniques;

3. The development of specific computer software which will aid in the analysis of the texts;

These aims derive from the need to investigate patterns within whole *texts* (Scott, 1997) instead of within a corpus with no regard for text boundaries. Since computer-aided investigation of whole texts is a novel enterprise in discourse analysis, several techniques will be explored, including statistical analyses; as Sinclair (1991, p.3) admits, in large-scale text analysis 'the numerical and statistical side has scarcely begun', which suggests that there are no established quantitative methodologies for computer-assisted text analysis.

1.6 Working definition of segment

Initially, for the purposes of this investigation, a segment is defined as a contiguous portion of written text consisting of at least two sentences. This definition reflects a position put forward by Kukharenko (1979) and Scinto

(1986) who observe that texts are constituted by sentence clusters, or 'semantic topical and lexico-grammatical unities of two or more sentences' (Kukharenko, 1979, p.235). It also ties in with a definition of text segment proposed by Fries (1995, p.54), according to whom, 'the term "text segment" is intended to apply to any chunk of text (presumably larger than one sentence in length) that is perceived as a unit'. It is assumed that segments are motivated, that is, there is a reason for considering a group of sentences part of the same segment. The motivation behind segments must be discourse-based: the linguistic characteristics holding these portions together must have been shown to be relevant for the analysis of discourse. In the chapters that follow a range of linguistic features will be discussed which could serve as the basis for segmentation. From these, one particular feature will be chosen to be used in the development of the computer-assisted segmentation procedure.

1.7 Organisation of the thesis

The thesis is organised as follows. Three theoretical chapters follow, each one focusing on a major area relating to the development of a segmentation procedure. Chapter 2 is concerned with the analysis of segmentation in discourse analysis. Key studies in discourse analysis are discussed from a segmentational point of view. Their potential contribution to the development of a computational methodology is assessed. Chapter 3 reviews the most important computer-assisted approaches to segmentation. Attention is focused on the potential contribution that each approach makes for an understanding of discourse organisation. Chapter 4 surveys studies which have dealt with lexical cohesion, with a special interest in looking for the most adequate analytical model for computer implementation. Chapter 5 presents a series

of pilot studies which have put into practice the various insights gathered from reviewing the literature. The chapter reports on the experimentation with three original segmentation procedures. Chapter 6 describes the development of the Link Set Median (LSM) procedure, and reports on the initial application of the procedure to a small corpus of texts. Because of the promising results, LSM is chosen as the procedure to be used in the large-scale investigation of larger bodies of data. Chapter 7 reports on the main large-scale study of segmentation using the LSM procedure. Several aspects of the analysis of a corpus of hundreds of texts are presented and interpreted. Chapter 8 presents a discussion of the results of the main study and relates the main findings to the literature reviewed earlier in chapters 2, 3, and 4. Finally, chapter 9 concludes with a general assessment of the achievement of the aims set out in the present chapter, and considers how future research can tackle some of the issues which have not been adequately dealt with in this thesis.

Chapter 2

Discourse Analysis and segmentation

In this chapter major models for analysis of discourse will be reviewed and interpreted in terms of what contribution they can make for the task of segmenting texts. Approaches designed with the specific aim of segmenting texts are included, namely Giora (1983), Goutsos (1996a), and Cloran (1995). The chapter also looks at a sample of the work of major exponents in discourse analysis, concentrating mainly on works dealing with written text, the exception being Sinclair and Coulthard (1992) in spoken interaction, whose work is seminal and representative of a school of discourse analysis (the 'Birmingham school'), Cloran (1995), whose model has been argued to be applicable to written text as well, and Hasan (1989), whose work on spoken interaction has been used elsewhere on written discourse (see Parsons (1990), reviewed in section 4.3.8, p.146 ff.). A number of other approaches reflecting a range of interests and perspectives on the organisation of written discourse are reviewed, namely Hasan (1989), Van Dijk (1980), Hoey (1983), Longacre (1983), Davies (1994), Mann and Thompson (1986b), and Pitkin (1969). Genre analysis, an influential perspective in discourse modelling, is also rep-

resented by the works of Bhatia (1993) (itself a review of genre analysis), and Paltridge (1994), a critic of this tradition.

2.1 Linguistic analysis and segmentation

Discourse Analysis has been described as that branch of language inquiry which is concerned with 'discovering linguistic regularities in discourse' (Crystal, 1991), p.106). As its name implies, it does so by analysis, that is, through 'the separation of a whole into its parts for study' (*American Heritage Dictionary*, 1994, p.30). Seen in this light, it is clear that the goals of Discourse Analysis are not in essence different from that which this thesis concerns itself with, namely segmentation (see section 1.2, p.5 in the Introduction). As a result, there are a large number of approaches in Discourse Analysis which could be read as being proposals for segmenting texts.

One reason why segmentation is not foreign to most approaches to the analysis of discourse is that segmentation is in a sense also inherent to linguistic analysis in general. As Schiffrin correctly states, 'a basic tool of linguistic analysis [is to] segment text into sections, label those sections as part of a structure, and assign function to those sections' (1994, p.11). Hence, segmentation of language is also a common activity in non-discourse linguistics, for instance, phonology and syntax.

At least in one particular theory of discourse, segmentation has a central place, namely the discourse theory proposed by Grosz and Sidner (1986). Their theory identifies three high-level constituents in discourse: linguistic structure, intentional structure, and attentional state. Segments are part of linguistic structure: 'just as the words in a single sentence form constituent phrases, the utterances in a discourse are naturally aggregated into *discourse segments*' (Grosz and Sidner, 1986, p.177, original emphasis). Their theory

as a whole has had an important influence in a variety of computational approaches to discourse (e.g. Hearst, 1993; Mann and Thompson, 1986b; Morris, 1988). The importance of Grosz and Sidner's (1986) theory to computational linguistics lies in the perception that 'it is one of those relatively rare efforts whose serious linguistic claims about discourse also have clear computational consequences' (Mann and Thompson, 1987a, p.42).

Segmentation is also a central activity in disciplines other than Discourse Analysis (see previous discussion in section 1.2, p.5 ff. in the Introduction). One such discipline is Conversation Analysis (cf. Glass, 1983), whose main aims consist of presenting a division of dialogic discourse in discrete parts such as 'turns' or 'adjacency pairs' (combinations of turns in an expected sequence, such as question–answer). Group Dynamics, an area of inquiry devoted to studying how groups of people interact in specialized contexts, has had an interest in segmentation for many decades. For instance, Bales and Strodtbeck (1968, p.389) note that the idea that problem-solving sessions can be divided into *phases*, or discrete sequential units, can be seen to date back to the 1910's. Cognitive psychologists investigating reading comprehension have also shown an interest in text analysis. For example, Meyer and Rice (1984) present and discuss several possible structures of prose and how they relate to cognitive processing during reading, and Clements (1979) discusses the influence of 'information chunks' on recall from written text. Story grammarians do not normally consider themselves discourse analysts. Nonetheless, one of their major concerns is how to represent narratives in terms of schematic constituents (e.g. Rumelhart, 1975). Contrastive rhetoricians have also made use of segmentation in analysing texts (Connor, 1996). For example, Ostler (1987) borrows the notion of discourse unit from Pitkin (1969) in order to analyse the organisation of EFL texts produced by Arabic writers.

2.2 Scope of the chapter

Given the large number of approaches to Discourse Analysis available in the literature, it would be unrealistic to try to cover them all in a single chapter. Such an enterprise would require a whole volume at least. A more realistic goal for a chapter in a thesis about segmentation would be then to provide a detailed look at a selection of discourse analytical approaches from the point of view of the solutions that they present to the issue of how to segment discourse.

However, because of the large number of approaches which fall within the boundaries of discourse analysis, criteria for selection had to be adopted. Otherwise, only a brief mention of each approach could be made, which would be inadequate since it is necessary to present both the tools that each approach makes use of and how they make use of such tools in order that the contribution of each approach to segmentation can be appreciated.

The first and most obvious criterion was to include those studies whose concern was explicitly with segmentation. There are but a few discourse studies which proclaim to be investigating segmentation. While it would be possible to restrict the chapter to those studies, this strategy would have excluded other discourse studies which, as was argued above, can also be seen as dealing with segmentation.

Apart from studies dealing directly with segmentation, there was a large body of contributions which were dealing with segmentation in so far as they were analysing discourse, that is, dividing a spoken or written text into discrete parts. In the event, these studies made up the vast majority of the literature, and therefore other criteria had to be adopted. The first criterion was to consider inclusion of a sample of those contributions which declared themselves to be part of discourse analysis. The sampling criteria for these studies are explained below. The next criterion was to include a

sample of studies in rhetoric, since rhetoric is in a sense the forerunner of discourse analysis. It has been a discipline devoted to analysis of discourse for centuries.

Criteria for exclusion also had to be adopted. It was decided to exclude those studies which proclaimed themselves to be part of Conversation Analysis. The major reason is obvious: Conversation Analysis is a vast discipline and would deserve at least a chapter of its own. Another reason is that the concerns of Conversation Analysts differ from those of discourse linguists in general. Conversation Analysts are more concerned with the description of the conversation in terms of how the participants interact than with the careful description of the linguistic base on which the conversation is realized. As Coulthard and Brazil (1992, p.51) rightly observe, Conversation Analysts 'do not attempt to define their descriptive categories but instead use "transparent" labels like *misapprehension sequence*, *clarification*, *complaint*, *continuation*, *pre-closing*.'

Admittedly, the boundaries between discourse disciplines are blurring, as attested by the inclusion of several independent well-established disciplines in the four volumes of the Handbook of Discourse Analysis (van Dijk, 1985). Hence, reasons could be adduced to support the inclusion of more discourse-related studies. Nevertheless, it would be more difficult to argue against the inclusion criteria presented so far. That is why it is hoped that the sample included in this chapter will be considered an acceptable compromise between depth and coverage.

It is argued here that approaches to discourse analysis, to the extent that they are relevant for a discussion about segmentation, can be divided into two groups: those which assign greater importance to content in proposing segment boundaries, and those which give greater prominence to surface markers in identifying segments. The segments identified by the former approaches

are not accounted for by linguistic signals but rather by other aspects such as speaker's goals, communicative purpose, and cultural familiarity. The segments identified by the latter approaches rely more heavily on the presence of linguistic signals of various kinds as indicators of segments, such as cohesion, discourse markers, and theme.

2.3 Content-oriented segmentation

In this section, contributions which make mostly use of content in defining segments will be discussed.

2.3.1 Cloran

The work of Cloran (1995) is an attempt to identify segments within a systemic functional grammar framework. The main question posed by the investigation is to what extent there are linguistic criteria for the identification of perceived 'chunks' of text. The results are based on an analysis of a dialogue between a pre-school child and his mother. Cloran (1995, p.401) argues, however, that her model is suitable for analysing written texts as well.

Rhetorical units and 'chunks'

The central concept in Cloran's (1995) proposal is the rhetorical unit, which is conceptualized as the linguistic realization of the pre-theoretical notion of (text) *chunk*. A rhetorical unit is understood as a semantic unit, but is realized lexicogrammatically by specific linguistic devices.

The starting point for the analysis of rhetorical units is the division of a text (or text extract) into segments based on the unit of *message*, which is defined as 'the smallest unit which is capable of realising an element of the generic structure of a text' (Cloran, 1995, p.362). Cloran (1995) draws

on the work of Hasan for the notion of message (cf. Hasan, 1996b, p.171): messages are realized by clauses.

Once the text has been divided into messages, Cloran (1995, p.362) looks for relationships among messages using 'the normal speaker's ability to understand the meanings made by language'. The chunks of related messages are then interpreted in terms of rhetorical units. The analysis is discussed below.

Segmentation

The basic question which is asked at this stage is 'what are the criteria by which these rhetorical activities may be recognized?' (Cloran, 1995, p.362). The answer is provided on the basis of the analysis of one text extract, namely a conversation between a child and his mother. The extract is divided into two segments in advance of the analysis, and the rhetorical units within each segment are discussed at length. Several rhetorical units were identified in the segments, and they were given labels such as 'situational observation' (or simply 'observation'), 'textual observation' (or 't-observation' for short), 'generalization', 'account', and 'commentary'. Some units were subdivided into smaller overlapping componential units.

The definition of each rhetorical unit is discussed in full in Cloran (1995, pp.364-365). For example, a generalization consists of 'making class exhaustive reference to whatever class of entity is mentioned', normally by describing entities in terms of its 'timeless attributes' (Cloran, 1995, p.365). An instance of generalization is the message 'A hydroplane is a plane that can land on the water' (Cloran, 1995, p.363). An account is constituted by a linguistic account of 'the existence and habitual functions of an entity' (Cloran, 1995, p.365), as in 'There's a helicopter that goes up and down the beaches in summer watching out for people' (Cloran, 1995, p.363). And a commentary is 'a

rhetorical unit where a speaker comments on an event of state or affairs in which a co-present entity is engaged at the time of speaking' (Cloran, 1995, p.365). The only commentary in the extract is an undeveloped one: 'Mother: Where's the pilot?; Child: Um this man'.

Formalization

The working definitions of individual rhetorical units offered by Cloran (1995) are considered unsatisfactory because they rely on the language user's intuitive interpretation of the interaction. Cloran offers a formalization of these definitions in terms of the systemic categories of Subject and Finite. Subject realizes the semantic function of 'entity' while Finite expresses 'event orientation'. Rhetorical units are determined by the nature of the entity and by the temporal orientation of the event. As a result, the classification of rhetorical units can be based on the analysis of the mood structure in terms of Subject and Finite.

In her thesis, Cloran (1994, p.133) provides a fuller explanation of the relationship between rhetorical units and mood elements. A rhetorical unit is realized by a central entity ('CE') and by an event orientation ('EO'). A CE in turn is typically realized by the entity functioning as Thing in the Subject role, and an EO is typically realized by the Finite verbal operator.

Conclusions

The model presented by Cloran (1995) offers a detailed account of the place of intermediate text units in relation to a systemic theory of text. The model is based on a rank scale consisting of three units: text, rhetorical unit, and message: 'messages are constituents of "rhetorical activity" and "rhetorical units" are constituents of text' (Cloran, 1995, p.399). The formalization of rhetorical units in terms of their lexicogrammatical features helps to make the

classification of rhetorical units more consistent, since it is the way in which rhetorical units are worded which is taken as the basis for their classification and not intuition on the part of the analyst.

Implications

The fact that rhetorical units are offered as an intermediate unit of text suggests that they can be treated as segments, even though the term 'segment' is reserved by Cloran (1995) to mean a collection of rhetorical units. The implication for the present investigation is that viewing rhetorical units as possible segments lends support to the notion that texts are segmentable. Furthermore, it suggests that segments are not arbitrary notions, rather they are realized linguistically.

A limitation of Cloran's study is that segments (not rhetorical units) are defined prior to the analysis. As such, Cloran's approach is designed for a particular kind of (limited) segmentation: further segmenting pre-defined segments. Having said this, the two segments turned out to have different rhetorical activity make-ups, and therefore in a sense the original division of segments was justified.

2.3.2 Bhatia and Swales

The works of Bhatia and Swales are representative of genre analysis, an analytical tradition which concerns itself with describing the typical structure of genres, or highly-structured communicative events 'with constraints on allowable contributions in terms of their intent, positioning, form and functional value' (Swales, 1990). Several studies have been published which looked at a range of different genres (e.g. Hopkins and Dudley-Evans, 1988; Hyland, 1990; Marshall, 1991; Nwogu, 1991; Salager-Meyer, 1989, 1990; Swales, 1981, 1990; Tinberg, 1988). In his book, Bhatia (1993) reviews the literature and

findings from more than a decade of genre analysis. For this reason, his book was chosen to serve as the representative of genre analysis in this chapter.

Interpretation and communicative purpose

Genre analysis is centred on the proposition that discourse is created 'as a result of the reader's interpretation of the text' (Bhatia, 1993, p.8). It follows that meaning is not inherent in the text, rather it is interactive. Ultimately, discourse is 'reader's discourse' since it is the result of the reader's interpretation of the text. Thus, subjectivity is at the heart of genre analysis.

Central to genre analysis is the notion of *communicative purpose*, which refers to the function genres are meant to fulfil in the world. The analyst defines the communicative purpose of the genre and then uses this information to guide him/her in describing the *cognitive structuring* of the genre. The cognitive structuring 'represents the typical regularities of organisation' (Bhatia, 1993, p.21). Such regularities are considered cognitive because 'they reflect the strategies that members of a particular discourse or professional community typically use in the construction and understanding of that genre' (Bhatia, 1993, p.21).

Moves

Another central notion in genre analysis is that of *moves*, or 'discriminative elements of generic structure' (Bhatia, 1993, p.30). Moves are the kinds of segments that genres are divided into. The set of moves forms the centrepiece of a genre analysis since it represents the typical cognitive structure of the genre, or the 'preferred ways of communicating intention' (Bhatia, 1993, p.29-30).

Moves are key elements in genre analysis because they represent the typical cognitive structure of a genre. Cognitive structure is defined as no less

than the 'conventionalized and standardized organisation used by almost all the members of the professional community' (Bhatia, 1993, p.32). Hence, a move analysis is said to represent the regularities of organisation adhered to by a professional community.

Research article abstract

As an example of genre analysis, the research article abstract is said to have the communicative purpose of 'telling all the important aspects of the very much lengthier research report' (Bhatia, 1993, p.82). Its typical structure contains the following four moves: (1) introducing purpose, (2) describing methodology, (3) summarizing results, and (4) presenting conclusions (Bhatia, 1993, p.78). These are illustrated in figure 2.1 on the next page.

The identification of these moves is based on the principle that moves are a representation of the typical cognitive structure perceived by a reader (Bhatia, 1993, p.30). Accordingly, move 1 'introduces purpose' because it is perceived as 'giving a precise indication of the author's intention, thesis or hypothesis' (Bhatia, 1993, p.79). In turn, move 2 'describes methodology' by being perceived to 'give a good indication of the experimental design' (Bhatia, 1993, p.79). Move 3 'summarizes results' by being perceived as the place where 'the author mentions his observations and findings' (Bhatia, 1993, p.79). And finally, move 4 'presents conclusions' because it is perceived as being 'meant to interpret results and draw inferences' (Bhatia, 1993, p.79).

The example discussed above illustrates the fundamental procedure of move assignment in genre analysis. The rationale behind the assignment of moves to other genres follows the same principle of perceived recurrent structure as that detailed above for abstracts. As a result, moves for other genres described below will not need to be exemplified. Move labels, in addition, are mostly self-explanatory.

Move	Text
1	This paper sets out to examine two findings reported in the literature: one, that during the one-word stage a child's word productions are highly phonetically variable, and two, that the one-word stage is qualitatively distinct from subsequent phonological development
2	The complete set of word forms produced by a child at the one-word stage were collected and analysed both cross-sectionally (month by month) and longitudinally (looking for changes over time).
3	It was found that the data showed very little variability, and that phonological development during the period studied was qualitatively continuous with subsequent development
4	It is suggested that the phonologically principled development of this child's first words is related to his late onset of speech

Figure 2.1: Move analysis of an abstract (Bhatia, 1993, p.79)

Conclusion

The large number of genres described using the methodology of genre analysis (e.g. Berber Sardinha, 1991; Hopkins and Dudley-Evans, 1988; Hyland, 1990; Marshall, 1991; Nwogu, 1991; Salager-Meyer, 1989; Swales, 1981; Tinberg, 1988) is proof that its methods are popular with a large number of discourse analysts. The strengths of the methodology lie in the power given to the subjective judgement of the analyst. In genre analysis no excuses are made for providing a description based on one's knowledge as a reader of the genre being described. Admittedly, Bhatia suggests consulting with members of the target discourse community, but the final decision lies with the genre analyst.

Implications

Accepting that discourse is 'reader's discourse' is fundamental to accepting the divisions of moves in genre analysis. Moves are intuitive categories and

their adequacy, placement, and labelling depend on the interpretation of the analyst, namely a reader who is in a position to observe recurrent cognitive structures across exemplars of the same genre. Move analysis is therefore an essentially subjective kind of analysis.

As was mentioned in the introduction to this section, genre analysis is concerned with segmenting texts. Genre analysis is devoted to segmenting prototypical texts (genres) into prototypical segments (moves), and as such it is relevant to the study presented in this thesis. However, as may have become clear during the preceding presentation, genre analysis approaches segmentation from a distinctively different perspective from that capable of being pursued in this thesis. The units of structural description in genre analysis are reader-based: moves are assigned by the analyst based on his/her interpretation of the contents of the parts of the text.

The implication of genre analysis to the present investigation is that it recognizes the central role played by segments ('moves') in discourse organisation. Moves are viewed not simply as a possible stylistic constituent but as a representation of the underlying cognitive structuring of the text. It is also believed that moves are representations of strategies employed by members of the discourse community, and are therefore considered to have a role that goes beyond that of descriptive categories. In this way, it is possible to think of certain kinds of segments as valid in the real world, and not simply as part of linguistic description.

2.3.3 Hasan

The model presented by Hasan (1989) is based on the notion that language is functional, which implies that there is a tight relationship between text and context. The approach is centred on showing how the contextual variables of field, mode and tenor map onto generic conventions. In practical terms,

context is seen in the model as the source of information which assists the analyst in identifying the elements of the text structure.

Contextual configuration

The central concept in Hasan's approach to text structure is 'contextual configuration' or CC, which is defined as 'a specific set of values that realises the register variables of field, tenor, and mode' (Hasan, 1989, p.55). A possible CC would be, for example, 'parent praising child in speech', in which field is expressed as 'praising', tenor as 'parent to child', and mode is specified as 'speech'.

The specification of the values of the register variables in a CC assists in making predictions about the structure of text. By structure is meant 'what elements must occur, what elements can occur, where must they occur, where can they occur, how often can they occur.' (Hasan, 1989, p.56). A CC can be used for predicting two kinds of elements: obligatory and optional. Moreover, a CC should include information about both the sequence and the iteration (i.e. recursion) of elements.

Service encounters

The analyses presented by Hasan (1989) refer to service encounters, more specifically interactions between a shopkeeper and a customer in a grocer's. The longest text analysed by Hasan (1989) is reproduced in figure 2.2 on the following page. Speakers are identified as V for 'vendor' and C for 'customer'. The encounter is segmented into structural elements, which are labelled down the left-hand column of the figure. Accordingly, SI stands for 'sale initiation', SC for 'sale compliance', S for 'sale', PC for 'purchase closure', SR for 'sale request', SE for 'sale enquiry', P for 'purchase', and F for 'finis'.

SI	V: Who's next? (1) C: I think I am. (2)
SR	C: I'll have ten oranges and a kilo of bananas please.(3)
SC	V: Yes, anything else? (4) C: Yes
SE	C: I wanted some strawberries (5) but these don't look very ripe. (6) V: O they're ripe all right. (7) They're just that colour kind a' greeny pink. (8) C: Mm I see (9)
SE	C: Will they be OK for this evening. (10) V: O yeah, they'll be fine; (11) I had some yesterday (12) and they're good very sweet and fresh. (13)
SR	C: O all right then, I'll take two. (14)
SE	V: You'll like them (15) cos they're good. (16)
SC	V: Will that be all? (17) C: Yeah, thank you. (18)
S	V: That'll be two dollars sixty-nine please. (19)
P	C: I can give you nine cents (20)
PC	V: Yeah OK thanks (21) eighty, three dollars (22) and two is five. (23) Thank you. (24)
F	V: Have a nice day. (25) C: See ya'. (26)

Contextual configuration of service encounter above:

field Economic transaction: purchase of retail goods: perishable food ...

tenor Agents of transaction: hierarchic: customer superordinate and vendor subordinate; social distance: near-maximum ...

mode Language role: ancillary; channel: phonic; medium: spoken with visual contact

Figure 2.2: Structure of service encounter (Hasan, 1989, p.61)

The CC in which the example text is embedded appears in figure 2.2. The CC is used for predicting the obligatory and optional elements in the example text. The obligatory elements are: sale request (SR), sale compliance (SC), sale (S), purchase (P), and purchase closure (PC); the remaining elements (i.e. sale initiation (SI), sale compliance (SC), sale enquiry (SE), and finis (F)) are considered optional.

The basis for deciding whether elements are obligatory or not lies in several factors. For instance, sales request is considered obligatory because 'the purchase of goods presupposes prior selection, and in a store with retail goods service, this selection must be made to the vendor' (Hasan, 1989, p.60). Similarly, the obligatory status of sales is justified as the need for the vendor to inform 'the customer what the exchange value of the goods is' (Hasan, 1989, p.60).

Underlying the segmentation and labelling decisions is a consideration for the key role played by ideology in service encounters. For instance, Hasan (1989, p.60) believes that the motivation for sales compliance can be found in the ideology of 'free enterprise' which 'raises the expectation of [the vendor's] readiness to serve as long as required' (Hasan, 1989, p.60) thus encouraging the customer to buy more.

Generic Structure Potential

The segmentation of an encounter into structural elements can be formalized in terms of *Generic Structure Potential* (GSP), which is a 'condensed statement of the conditions under which a text will be seen as one that is appropriate to [a particular] CC' (Hasan, 1989, p.64). The GSP for the example text (see figure 2.2 on the preceding page) is modeled after the contextual configuration in the same figure, and is written as:

$$[(G) \cdot (SI)\hat{}][(S\ \overset{\frown}{E}\cdot)\ \{SR\hat{}SC\ \overset{\frown}{\hat{}}\}\ \hat{}S\hat{}]P\hat{}PC(\hat{}F)$$

The conventions applying to the formalization of a GSP are as follows. A caret sign (^) represents sequence, the braces with a curved arrow signify that 'the degree of iteration for elements within the braces is equal' (Hasan, 1989, p.64), a dot indicates more than one option in sequence, round brackets signal optionality, and square brackets mean restricted optionality of sequence. Hence, the first square bracket must be read as 'G and/or SI may not occur; if they both occur, then either G may precede SI, or follow it; neither G nor SI can follow the elements to the right of SI' (Hasan, 1989, p.64).

Conclusion

The approach to text structure developed by Hasan (1989) is dedicated to showing how text structure can be derived from context. By highlighting the central role of context Hasan is able to describe each text in terms of a genre rather than as an individual text (Hasan, 1989, p.68). In theory, the description of text exemplars is informed by the conventions which are expected to operate in the genre which the exemplar belongs to.

Implications

One can envisage problems in applying Hasan's model. The most serious is that the analysis seems to depend on the initial description of contextual configuration of the text. Presumably, a badly-formulated CC would lead to a badly-partitioned text. However, there is a jump from the CC to the actual partitioning of the text which is not well explained. How does one translate CC into text constituents? In fact, it could be argued that the jump could equally be made in the other direction, that is, from the finished description to the formulation of the CC. This is because the link between CC and structural description is weak.

In fact, the description itself is achieved by drawing on the analyst's

intuitive knowledge of the genre. While there is nothing inherently undesirable about using intuition, there are no provisions built in the model which encourage the analyst to check his/her intuition against many instances of authentic data. As a result, the generalizability of descriptions derived from model is weakened.

The existence of these problems has implications for the present study. They suggest that a better model would be one which placed less emphasis on intuitive decisions to segment. At the same time, a better model would preferably be tested and testable on a large number of texts. All of this seems to point in the direction of a model which can be operationalized on the computer.

2.3.4 Paltridge

There are several criticisms which can be made of current models of discourse analysis. Some of these are put forward by Paltridge (1994), who focuses on genre analysis. The major criticism is that models of genre imply that the motivation for structural elements (textual boundaries) is linguistic, when in fact it is psychological. Paltridge (1994) argues that the rationale behind structural elements proposed by genre analysts is actually based on content rather than on linguistic form.

Structural elements vs lexical cohesion

The role of lexical cohesion in indicating structural elements is put into question. Paltridge (1994) reviews the fit between lexical chains and structural elements in service encounters as presented in Hasan (1989), and notes that chains normally go beyond structural boundaries. For instance, in figure 2.3, the lexical chain formed by the repetition of 'dollar(s)' cuts across the boundaries of sale, purchase, and purchase closure. Paltridge (1994) concludes that

lexical cohesion cannot account for the structural elements in the encounter.

Structural elements vs semantic attributes

The hypothesis put forward by Hasan (1996a) argues that two types of semantic attribute can account for segments ('structurally important units') of *any* text type: nuclear and elaborative. Nuclear attributes are of particular importance since they are obligatory elements. A review of the analysis of the nuclear elements in service encounters and nursery tales suggests that nuclear attributes do in fact seem to be related unequivocally to individual structural elements (Paltridge, 1994, pp.292-293). But, according to Paltridge (1994), the reason is that semantic attributes relate to content and not to linguistic realization, since there are no linguistic signals that correlate with structural elements.

Structural elements and content

The hypothesis that structural elements correlate with content is further pursued by Paltridge (1994). He presents a content-based analysis of structural elements in research articles and observes that a content category prevails in certain structural elements: 'quantity' (Paltridge, 1994, p.294).

Other work in genre analysis is also said to have followed content-based criteria for drawing textual boundaries. For instance, the moves suggested by Swales (1990) in research articles are said to be based on 'broad content-based terms'. Similarly, the rationale behind Bhatia's (1993) analyses of moves is attributed to considerations of content rather than linguistic form. Significantly, Paltridge argued that genre analysts have not been aware of the central role of content in their models.

Structural Element	Participant	Text	Lexical chain
Sale request	Customer	Can I have ten oranges and a kilo of bananas please?	
Sale compliance	Vendor	Yes, anything else?	
	Customer	No, thanks.	
Sale	Vendor	That'll be dollar forty.	dollar
			↕
Purchase	Customer	Two dollars.	dollars
			↕
Purchase closure	Vendor	Sixty, eighty, two dollars. Thank you.	dollars

Figure 2.3: Lexical chains in service encounter (adapted from (Paltridge, 1994, p.290)

Conclusions

The main conclusion in Paltridge's (1994, p.295) paper is that most work in genre analysis 'draws essentially on categories based on *content* to determine textual boundaries, rather than on the way in which the content is expressed *linguistically*.' According to him, this is a mistake since content is psychological. This deficiency has already been acknowledged even by genre analysts such as Bhatia (1993, p.19).

Implications

Of the points raised by Paltridge (1994, p.295) the most central to the present investigation is that which concerns the alleged absence of linguistic correlates of textual boundaries. However, since Paltridge (1994) did not provide an account of all possible linguistic signals (which would have been impossible), his position must be interpreted as meaning that no single linguistic signal can account for all segment boundaries. Indeed, some signals such as cohesion may not contribute at all to segmentation. This possibility has implications for the investigation presented in this thesis in that it suggests that lexical

cohesion cannot be expected to account for all segment boundaries. Such a position would be supported by other studies which investigated lexical cohesion and found lexical cohesion to play not an absolute but a relative role in texts. For instance, Parsons (1990) observed that lexical cohesion accounted not for the entirety but for about 30% of the coherence in his texts.

Limitations

The data in figure 2.3 on the preceding page can be revisited. The presence of a lexical chain across 'sale', 'purchase' and 'purchase closure' can be interpreted as showing that these three structural elements are connected. These three elements form a set which can be seen as distinct from the remainder of the encounter: they represent the portion of the encounter where the transaction actually takes place, that is, where money is exchanged for goods. Therefore it is not surprising that there is repetition of 'dollars' because the word is used to indicate that the customer should hand out money and the shopkeeper should hand out the groceries.

The separation between cognitive and linguistic perspectives on language can only be sustained if the linguistic is understood as grammatical, and not as discoursal. As Paltridge himself admits, it depends on how 'one defines the domain of linguistics' (p.297). He concedes that there is a place for studying boundary divisions in terms of linguistic content if linguistics is redefined as the investigation of 'how human beings process and use language' (p.297). The problem is that for discourse analysis this is exactly what language study stands for. It is only in a view of language study which excludes language use that the objections raised by Paltridge can be seen to hold. The lesson to be learned from Paltridge is that there is a greater need for explicitness in discourse description.

2.3.5 Burke

To close the current section on approaches to discourse description which rely on content, we introduce an investigation by Burke (1991). Burke's work is content-driven but he made an effort to bring in greater explicitness by making use of linguistic clues in analysing discourse. The interest of his investigation is in the segmentation of a dissertation defence ('viva') which took place at an American University. The dissertation itself was in sociology, and the defence was considered a typical event in American universities. Hence, Burke (1991) hopes to describe not only a single exemplar but a more or less generic form of organisation of dissertation defences.

Plans

Plans are central to development of interaction according to Burke (1991). He argues that speakers must have a mental plan of the desired structuring of the event in which they are engaging. Speakers signal their contribution to the development of the interaction by linking their turns to other turns in the conversation according to how they see the place of their individual contribution in the overall plan of the event. As Burke (1991, p.98) puts it, there seems to be 'some sort of consensual marking or identification of the unit components of the plan'.

Importantly, even though speakers have a mental plan of the overall structure of the interaction, their contribution to the realization of the plan is in terms of turns. If segments are to be identified at all, there must be ways in which turns are linked together to form segments. The description of such features which enable speakers and listeners to organise the interaction and perceive such organisation is the central concern of Burke's (1991) paper, and these points will be taken up below.

Criteria for segmentation

The unit which Burke uses as the starting point for the segmentation is the 'turn' (Sacks et al., 1974). Turns are realised serially in time, and Burke argues that they cluster together to form segments. In order for segments to be indentified, the progression of turns in time must include not only the means for speakers to recognize how each turn is to follow another in the conversation, but also the means by which speakers signal how they move from segment to segment. Burke further argues that segments cluster as well, forming a hierarchical structure. The organisation of turns can therefore be seen from two perspectives: vertically and horizontally. The vertical principles involves 'hierarchical relationships between parts and wholes (content and context)', while the horizontal principle involves 'linkages between parts over time' (Burke, 1991, p.98). The interaction of the two principles enables the conversation to shape up as an event familiar to the participants.

The interaction between the horizontal and vertical principles is described by Burke (1991) in terms of the conditions under which each one takes precedence. The horizontal principle seems to take precedence 'behaviorally' (Burke, 1991, p.99), or when the individual contributions are taken into account. In contrast, the vertical principle of organisation takes precedence when one considers the initial plans and expectations which interactants bring to the conversation. The vertical organisation is crucial because without it 'segmentation features could not be recognized or identified' (Burke, 1991, p.99). In this sense, segmentation is not treated by Burke (1991) as an after-the-fact phenomenon; rather, segmentation is an inherent property of interaction.

Several linguistic criteria for segmentation are identified by Burke (1991, p.98): (1) forward-backward referencing of turns, (2) use of metacommunication, (3) repetition, (4) key words and markers (e.g. 'now then' and 'okay'),

(5) joking and humor, (6) speaker continuity across episode boundaries, and (7) kinesic markers. He acknowledges that the list is not exhaustive, and that other markers may contribute in different ways to the segmentation of interactions.

Segmentation of defence

A typical dissertation defence is described as having five major segments: (1) Introductory background, (2) Questions [by each one of the examiners], (3) Assessments, (4) Interlude, and (5) Wrap-up. These segments are identified by Burke (1991, p.100) because he considers himself to be 'a member of a culture which recognizes and produces dissertation defences'. Hence, the ability to identify segments would seem to depend largely on top-down processes, that is, how familiar one is with the cultural context in which a particular interaction took place.

However, segmentation also depends on bottom-up processing. Major transitions are recognized as being 'usually marked in particular ways, and one not familiar with the specific cultural context can find his or her way about the discourse by knowing these particular markings' (Burke, 1991, p.101). Given that familiarity with the cultural context is not indispensable for recognizing the segments of the defence, both top-down and bottom-up processes are at play in Burke's (1991) approach to segmentation.

There are both major and minor transitions in Burke's (1991) approach to the segmentation of the dissertation defence. Major transitions are defined as being those which occur at higher levels of the hierarchy, that is, between major segments. By contrast, minor transitions are seen as occurring between 'units lower down in the hierarchy' (Burke, 1991, p.101). Burke (1991, p.101) warns that the difference between major and minor transitions is relative, and refer to the 'nature and type of units between which the transition is made'.

Two major segmentation features are identified by Burke (1991). The first is that one interactant takes up the role of key speaker by marking and controlling the segmentation. In the dissertation defence, this role was played by the chairman, who seemed to be in charge of marking the major transitions in the defence.

The second feature is the presence of 'metacommunication which directly provides a map of what is happening to all participants' (Burke, 1991, p.108). In the dissertation defence, this was presented in the form of a map by the chairman. The map comprised only the first two segments (Introductory background, and Questions). This seemed to influence the kind of metacommunication needed to signal the transition between the major segments in that there was no need for the chairman to mark the transition between the 'Background' and 'Questions' segments, unlike between the remaining segments.

Linguistically, the main clue which was used to signal the major transitions between segments was the marker 'so'. For instance, the Background segment was initiated by the phrase 'so why don't you tell us what you've been doing ...'; the Assessments segment was introduced in a similar way by 'so': 'okay so how do you do you wanta make some uh you got some reactions James'; and the Wrap-up segment was marked by the phrase 'so I can tell her that I ...'.

The most important types of transitions between minor segments are signalled by forward- and back-referencing linkages. It is argued that through these links 'the stream of discourse is segmented into relatively coherent 'chunks', each separated to some extent from the other' (Burke, 1991, p.117). Forward-referencing linkages are created by questions, commands, and summons. Burke (1991, p.114) claims that the importance of forward-referencing linkages lies mostly in their ability to interrupt the flow of conversation and

generate new segments. By contrast, back-referencing linkages enable the speaker to remain in the same segment. The most frequent devices for referencing backward are indexicals (expression such as 'this', 'that', and 'it'), repetitions, and extensions (when 'the current turn is constructed in such a way that it could have been uttered by the prior speaker as a continuation of that turn', Burke (1991, p.111)).

Control

An important issue entailed by segmentation is control. Burke (1991, p.124) states that 'control operates when one segment serves as context in which other segments are interpreted'. On a more abstract level, this means that major segments control the minor segments contained within them since they 'set the tone' for these minor segments. Thus, questions which appear in the 'Questions' segment will tend to have a different status from questions asked during the 'Wrap-Up', for instance. That is because the major segment, being 'Questions', exerts control over the individual turns within which questions are asked by providing the overall context which will guide how questions are interpreted.

On a more interpersonal level, Burke (1991, p.124) noticed that the speaker who initiates major segments 'tends in fact to control the context of the segments contained within'. Control seemed to be exerted by that speaker who initiated a major segment. That speaker seemed to be in control of the other turns which followed on from the initiation.

Control has two implications: 'first, that the context is set by the person in control; second, that the floor returns to the person in control for the next opportunity to initiate a segment at the same level' (Burke, 1991, p.124). This perhaps explains why the chairman exerted control during the defence, since it was he who initiated most major segments. By initiating major

segments, the chairman seemed to have set the context for the segment, which in turn later gave him the right to initiate another major segment, and so on. As Burke (1991, p.124) concludes, 'control perpetuates itself'.

Conclusion

There are two main points in the study into the segmentation of the dissertation defence presented in Burke (1991). Firstly, segmentation is a natural phenomenon which is part of the interactants' expectations when they engage in the interaction. Finally, segmentation is related to control in two ways, firstly in the form of the context provided by larger segments which controls how smaller segments are interpreted, and secondly, in the ways in which speakers become dominant in major segments by creating major segment transitions.

Implications

An important insight provided by Burke's (1991) study is that segmentation is not an analytical artefact; rather, it has psychological validity in that it represents a mental map on which interactants lay out the major organisational components of the interaction. Although his study was concerned with speech, there is no reason to suppose the same insight would not apply to writing. It is perfectly possible that writers and readers have a mental map of the organisation of a piece of writing to help them create or understand a text, and that parts of this map are represented graphically by section headings. If this is the case, then the study of segmentation in writing is not a simple analytical exercise, and it may contribute to some extent to understanding the way in which both composition and text comprehension processes are seen to take place.

A related point is provided by Teresa Labov's discussion of Burke's (1991)

study which appears at the end of the article. She makes the point that 'people are ordinarily well equipped to do segmenting' (Burke, 1991, p.125), referring to how people are part of 'nested collections of people': 'people who are examining a candidate, people signing papers, and of course people who are or who once were candidates' (Burke, 1991, p.125). In other words, people seem able to follow the segmentation of the dissertation defence because they are part of a culture which is familiar with dissertation defences, or in the words of Burke (1991, p.100), a culture which 'produces dissertation defences'. In summary, contextual factors allow speakers to recognize and reproduce the segmentation typical of dissertation defences; the segments therein are a part of the processes which give rise to the production of defences.

However, if the cultural context is so important in determining the segmentation of the defence, the question remains of whether the segmentation is entirely defined beforehand. The question is important because if the answer is affirmative, the logical conclusion would be that there is no need for a linguistic analysis of segmentation given that one would be able to deduce the segmentation by knowing cultural and contextual factors, or more precisely, by using the intuitive knowledge that comes from being part of 'a culture which produces dissertation defences'. The answer provided by Burke (1991) is that speakers will leave signals which indicate when segment boundaries are crossed. Further, the linguistic signals enable them to communicate in which segment they are. Although these points are valid from the point-of-view of those actually involved in the interaction, to the analyst they are less helpful. The question still remains of whether familiarity with a genre is a necessary prerequisite for the analyst, or, put in another way, whether familiarity would prove a hindrance in that it might make it more difficult for the analyst to take into account evidence which would run counter to

his/her expectations.

2.4 Surface markers as a basis of segmentation

The previous section drew attention to contributions which focused mostly on content as a means of identifying segments, and ended with a study which relied to some extent on explicit linguistic signals (Burke, 1991). In this section, attention will be turned to those contributions in which surface makers play a central role in segmenting texts.

2.4.1 Pitkin

The work of Pitkin (1969) is aimed at proposing an analytical scheme which lends itself to use in the composition classroom. His approach is devoted to finding *discourse blocs*, or rhetorical units which organise written discourse hierarchically. His approach expands on previous work in the paragraph by Christensen (1965), and was influential on later studies in the area of Contrastive Rhetoric (Connor, 1996; Ostler, 1987).

Discourse structure

Written discourse structure is viewed by Pitkin (1969) as being essentially hierarchical. He argues that discourse hierarchy is typically seen in terms of sentences being organised in clauses, and clauses in phrases, and phrases in words and so on down to phonemes. Additionally, hierarchy has been seen as applying to the way sentences form whole discourses, so sentences are also seen as forming paragraphs, paragraphs are generally regarded as forming chapters, chapters forming books and so on up to 'the Library of Congress

and beyond' (Pitkin, 1969, p.139).

The problem with this view of hierarchy is that it implies that written discourse structure is static, that is, discourse is a series of discrete units. Pitkin argues that a static view cannot account for the way discourse operates. Instead, he proposes to look at discourse as an 'operation' in which units 'are units because of what they do, not merely what they look like' (Pitkin, 1969, p.139). In this view, the discourse continuum 'would be segmented by junctures in space-time, not merely by joints in space' (Pitkin, 1969, p.139). Examples of such junctures would be points in discourse where we can say 'To this point we have been doing X; now we begin to do Y' (Pitkin, 1969, p.139).

Discourse blocs

The basic unit of analysis in Pitkin's model is the *discourse bloc*, a functional unit which describes the way in which smaller units are organised into larger units so that there are no gaps in the continuum. The fact that gaps are not allowed in the description is important, since it stipulates that all smaller units must be part of a larger unit. The number of units at any one level of detail can vary, but it can be no less than two.

The identification of discourse blocs starts with the division of the text into 'the smallest unit to have a discrete function in the discourse' (Pitkin, 1969, p.142). Normally such units correspond to rhetorical sentences, that is, a string starting with a capital letter and ending with a terminal punctuation mark. An important step in determining discourse blocs is the specification of the relations between them. These relations are summarized in what follows.

Vertical and horizontal relations

Discourse blocs are combined by means of the *vertical* and *horizontal relations*. The difference between vertical and horizontal relations rests on the level of generality expressed – vertical relations include 'a shift in generality' while horizontal relations do not. Shifts in generality are defined empirically, so that there is no shift in generality between 'question and answer' but there is a shift between 'genus and species'.

Vertical and horizontal relations are further subdivided. Vertical relations comprise 'subordination' and 'superordination', while horizontal relations incorporate 'coordination', and 'complementation'. The actual realization of subordination would include, for instance, a move from 'genus' to 'species', whereas superordination would work in the opposite direction, from 'species' to 'genus'. In turn, blocs would be identified as coordinated if they had mention of 'dog' and 'cat' but not 'pet' (in which case they would be subordinated/superordinated), and as complemented if one was seen as, say, 'cause' and another as 'effect'.

Teaching of composition

The motivation behind Pitkin's formulation of discourse blocs is essentially pedagogic. He believes that if discourse analysis is to have any impact at all on the teaching of writing then discourse analysis must abandon the notion that discourse is a series of individual sentences, or individual paragraphs. Rather, as was mentioned above, discourse is formed by relations between its parts. Pitkin claims that sentence-based approaches are useless because 'contemporary writers don't set out to write sentences, they set out to write discourse' (Pitkin, 1969, p.139).

A further consequence for understanding discourse as made up of individual units is that the learning of composition is then seen as learning to

classify units. According to Pitkin, teaching students to classify sentences grammatically as 'simple' or 'compound', or rhetorically as 'loose' or 'balanced', cannot show one how to write acceptable discourse. Instead, by showing students how units function together as 'a continuum of increasingly complex structures' (Pitkin, 1969, p.141) it becomes possible to help them write more successfully.

Analysis

An analysis of four passages is presented in Pitkin (1969). The first of these passages, and the one whose analysis is explained, is reproduced in figure 2.4 on the next page, a paragraph from an article published in *Scientific American* in 1951. In all, 13 discourse blocs are identified, one for each unit, though only the topmost blocs will be referred to here.

Individual units combine into larger units, the largest of which is referred to as 'definition of intelligence'. 'Definition of intelligence' is broken down into two blocs: 'definition' and 'refuted qualification of definition', the former comprising units 1, 2A and 2B, while the latter covers the rest of the text. 'Definition' is further divided into 'definition' (unit 1) and 'partial repetition with addition' (units 2A and 2B). In turn, 'refuted qualification of definition' subdivides into 'qualification and refutation' (units 3 through 10) and 'deduction' (unit 11).

The rationale for separating out some of the blocs is explained. For instance, the 'definition' bloc (units 1, 2A and 2B) is formed by means of the repetition of 'capacity' in unit 1 as 'ability' in unit 2A, and by the repetition of 'learn to adjust successfully to' (unit 1) and 'solve' (unit 2A). The break in the 'refuted qualification of definition' bloc at unit 11 is suggested by the bloc signal 'hence'. Accordingly, unit 11 forms a new bloc labelled 'deduction'.

(1) Intelligence may be defined as the capacity of an organism to learn to adjust successfully to novel and difficult situations. (2A) It involves the ability to solve new problems (2B) by drawing on past experience. (3) Those who consider that animals below man are mere mechanisms place great emphasis on a supposed distinction in methods of learning, as between human beings and animals (the word *animals* will be used here to mean the non-human ones). (4) In animals learning is mainly by trial and error. (5) When a cat, for example, is confined in a latched box that acts as an obstacle preventing it from reaching food outside the box, it tries all sorts of measures to solve the problem. (6) It claws at various parts of the box, attempts to shake the box to pieces, tries to push the door open, and so on. (7) After many such fumbling trials, it eventually learns to throw the latch and open the door without fumbling. (8) Man, on the other hand, can solve such simple problems almost immediately, with a minimum of trial and error. (9A) This 'insight learning' is commonly supposed to show a qualitative difference between men and animals, (9B) but actually some of the higher animals are capable of it. (10) Moreover, man often uses the trial-and-error method, particularly in forming new motor habits such as typing or playing golf and in solving mechanical puzzles. (11) Hence it would seem that the true measure of intelligence is the capacity to learn, regardless of the method involved.

Figure 2.4: Example text (Pitkin, 1969, p.143)

Conclusion

The approach presented by Pitkin (1969) is designed to show how units function as part of larger more complex hierarchical units. The motivation for his analysis is pedagogic, since he is interested in developing a model which can help teach people to write better. He believes that the concept of discourse bloc can assist in this task because discourse blocs are by definition hierarchical units.

Implications

The approach to discourse analysis presented by Pitkin (1969) must be interpreted within the context of discourse studies as they were conducted nearly thirty years ago. He is careful to frame his own work as independent of the current linguistic thinking of the time: 'I am asking to be tried here in the name of composition, not transformational-generative grammar' (Pitkin, 1969, p.138). To him, it was clear at the time that discourse could not be

adequately accounted for by sentence-based grammar.

His work contributes at least one interesting insight to discourse segmentation today, namely that existing orthographic divisions such as paragraphs do not necessarily match underlying functional divisions such as discourse blocs. This does not exclude the possibility of matches, though. In one of the texts analysed, paragraph breaks do correlate with middle-level discourse blocs (Pitkin, 1969, p.146-147). This insight gives support to the course of action adopted in the present study which consists of first finding segments independently of large-scale divisions, and then seeing whether segment and large-scale divisions match.

2.4.2 Hoey and Winter

The work of Hoey (1983) presents a framework for analysing texts in terms of cultural patterns such as 'Problem–Solution' and 'General–Particular'. Patterns are 'combination of relations organising (part of) a discourse' (Hoey, 1983, p.31). Underlying these patterns are clause relations such as 'Cause–Consequence' and 'Instrument–Achievement', which were originally introduced by Winter (1971, et seq.). Discourse patterns can be found in both short passages and long texts.

Clause relations

Clause relations play a central role in the way Hoey (1983) sees discourse as being organised. Clause relations are understood as the 'cognitive process whereby we interpret the meaning of a sentence or group of sentences in the light of its adjoining sentence or group of sentences' (p.18). Hoey (1983) draws on Winter's work (e.g. Winter, 1971, 1977). There are two kinds of clause relations: logical and matching. Logical (sequence) relations are relations 'between successive events or ideas', whereas in matching re-

lations 'statements are "matched" against each other in terms of degrees of identicality of description' (Hoey, 1983, p.19-20). Condition-Consequence, Instrument-Achievement, and Cause-Consequence are examples of logical sequence relations, whereas Contrast and Compatibility are examples of matching relations.

Clause relations are signalled by a number of elements that Winter (1977) normally referred to as three 'Vocabularies'. Vocabulary 1 comprises subordinators, Vocabulary 2 conjuncts, and Vocabulary 3 includes lexical signals. The same clause relation can be signalled by any of the three vocabularies. For instance, the Instrument-Achievement relation in the following sentence is signalled by Vocabulary 1 ('by + ing'):

'By appealing to scientists and technologists to support his party, Mr Wilson won many middle-class votes'.

In the following sentence, the same clause relation is expressed by a Vocabulary 2 item ('thereby'):

'Mr Wilson appealed to scientists and technologists to support his party. He thereby won many middle-class votes in the election'.

Finally, another version of the same sentence can be written in which Instrument-Achievement is signalled by Vocabulary 3 item (the lexical item 'instrumental'):

'Mr Wilson's appeals to scientists and technologists to support his party were instrumental in winning many middle-class votes in the election.' (all examples from Winter (1977), cited in Hoey (1983, p.23)).

Repetition

Relations within discourse are also signalled by repetition (Winter, 1974, 1979). Winter (1979) argues that sentences are selective, since no single sentence can hold all the information about a given subject. It follows that

sentences must relate to one another. Repetition functions in this context by virtue of ' "opening out" a sentence so that its lexical uniqueness may be used as the basis for providing further, related information' (Hoey, 1983, p.25).

Repetition also helps in the interpretation of sentences. The repeated information in a pair of sentences can be interpreted as a constant which allows 'the new information [to be] recognised and its importance to the context [to be] assessed' (Hoey, 1983, p.25). For instance, in the following sentence, repetition creates parallelism which aids in the identification of the relation between the two halves of the sentence: 'In spite of the hopes and promises of her new allies, Germany remains divided; in spite of strenuous efforts at international virtue, she feels herself morally reviled' (Hoey, 1983, p.24). The constant element is created by the repetition of 'In spite of' and 'Germany/she'. The compatibility between the variable elements is thus brought to the fore: 'remains v feels' and 'divided v reviled' (Hoey, 1983, p.25).

Discourse Patterns

The major aim of Hoey's (1983) work is to present and discuss popular patterns of discourse organisation. A pattern is defined as a 'combination of relations organising (part of) a discourse' (Hoey, 1983, p.31). A number of patterns exist, but the 'Problem–Solution' pattern is discussed in detail to show how other patterns can be handled using the same approach.

As a means of illustrating how the Problem–Solution pattern can be identified, Hoey (1983) presents a short narrative and analyses it in detail. The example text is reproduced in figure 2.5 on the following page.

While the example text can be intuitively recognized as presenting a problem and a solution, these categories are not explicitly marked. For instance,

the most obvious indicators of problems and solutions are not present, for instance the expressions 'the problem was ...' and 'the solution was ...'. Therefore, the analyst must uncover the relations between the sentences of the text before assigning Problem–Solution categories to them.

Methods of analysis

The assignment of the categories of the Problem–Solution pattern to the example text is accomplished by the application of one or more of four possible methods of analysis: (1) interpretation/introduction of subordination and conjuncts, (2) narrative interrogation, (3) elaborating interrogation, and (4) lexical signalling.

Relations may be uncovered by interrogating the text. In narrative interrogation, questions such as 'what happened?' and 'what was your response?' are asked; for instance: 'I was on sentry duty. Question: What happened? I saw the enemy approaching. Question: What was your response? I opened fire. Question: How successful was this?/What was the result of this? I beat off the attack' (Hoey, 1983, p.38). In elaborating interrogation, the questions that are asked are, for instance, 'how?' and 'why?'; the example text thus becomes: 'I beat off the attack. Question: How ⟨did you beat off the enemy attack⟩? I opened fire. Question: Why ⟨did you open fire⟩? I saw the enemy approaching. Question: In what situation ⟨did you see the enemy approaching⟩? I was on sentry duty.' (Hoey, 1983, p.38). An important difference between the two types of interrogation is that while narrative

Situation	(1)	I was on sentry duty.
Problem	(2)	I saw the enemy approaching.
Solution	(3)	I opened fire.
Evaluation	(4)	I beat off the attack.

Figure 2.5: Example text (Hoey, 1983, p.35)

interrogation 'is only complete when the last answer is given', elaborating interrogation 'is complete at each stage' (Hoey, 1983, p.38).

The final method is by exploring lexical signals. As mentioned above, if expressions such as 'the problem is (...)' had been part of the example text, these expressions alone would have been evidence of the status of the sentence as a problem. As Hoey (1983, p.63) puts it, 'lexical signals are the author's/speaker's explicit signalling of the intended organisation'. However, depending on genre lexical signals need not be present, and must in such cases be inferred. If the example text were rewritten including lexical signals, it could read, for instance, as: 'My *situation* was that I was on sentry duty. I saw the enemy approaching. I *solved* this *problem* by opening fire. This *achieved* the *desired result* of beating off the attack.' (Hoey, 1983, p.53, original emphasis).

Conclusion

The approach to the identification of discourse patterns presented by Hoey (1983) enables the analyst to locate and label discourse constituents based on the semantic relations between parts of the text. The rationale for the identification of discourse patterns is based on the notion of clause relation. Although the name implies that relations exist between clauses (Hoey, 1983, p.18), relations also exist between larger parts of discourse. This is crucial for the application of this framework to larger texts.

The relationship between clause relations and discourse patterning is expressed in great detail in Hoey (1983). Guidelines are provided for inferring clause relations as well as for relating clause relations to discourse patterns in terms of mapping conditions. These conditions are essential in allowing the model to be applied to a wide range of large texts.

Implications

Constituents of discourse patterns can be seen as one kind of semantic segments. The possibility of being able to segment texts in terms of a network of semantic relations has implications for the present investigation. It suggests that identifying large-scale patterns in whole texts is a complex yet feasible enterprise.

The framework proposed by Hoey (1983) is appealing in that it suggests that it can be implemented using only a layman's understanding of the concepts involved in the analysis. However, as Hoey (1983) himself admits, real-world categories are not necessarily similar to the linguistic counterparts, so that real-world 'problem' is not the same as the linguistic Problem. In truth, the successful identification of discourse patterns will depend on a considerable amount of knowledge about clause relations, in addition to a detailed understanding of the conditions which allow clause relations to map onto discourse patterns.

The recognition of the central role of repetition in underlying discourse patterns has implications for the present investigation in that it suggests that repetition is a major feature of discourse patterning. Unlike clause relations, repetition is amenable to identification by computer. The central role of repetition is all the more important because it is proposed within an approach which is not concerned with being implemented on the computer. Therefore, in a sense, the status of repetition has been 'independently' asserted because it was suggested without having computer-based applications in mind.

2.4.3 Mann and Thompson

Another approach to use semantic relations as a basis for segmentation is Rhetorical Structure Theory, or 'RST'. Rhetorical Structure Theory is a de-

scriptive theory of text organisation developed by William Mann and Sandra Thompson (Mann and Thompson, 1986b, 1987a; Mann et al., 1989). RST is based on the notion of rhetorical relations between units of texts (e.g. Hoey, 1983; Winter, 1977). A 'unit' for Mann and Thompson is typically an independent clause, and a 'span' is a combination of two or more units. The outcome of an RST analysis can be interpreted as a segmentation of text into parts defined functionally.

Relations, spans, and schemas

The central elements in RST are (rhetorical) relations and spans. The two concepts are co-dependent so that in order to understand relations it is necessary to understand what a span is. Relations 'identify particular relationships that can hold between two text spans' (Mann et al., 1989, p.11), while a text span is 'any portion of text that has an RST structure' (Mann et al., 1989, p.11).

Relations are functional, and as such their basis can be expressed in many ways. For instance, relations can express the 'purposes of the writer, the writer's assumptions about the reader' as well as the propositional content of the text (Mann et al., 1989, p.8). Relations are rhetorical because they represent the choices made by the writer in respect to how he/she presented and organised the text.

Most relations are asymmetrical, that is, they comprehend members of different degrees of centrality: a nucleus and satellite(s). As their names imply, the nucleus is the member of the span which is more central, whereas a satellite is a more peripheral member (Mann et al., 1989, p.8). Nuclearity is defined as that part of a schema which 'influences the way the reader assigns different roles to different parts of the text' (Mann et al., 1989, p.13).

There are a large number of relations in RST. Some of the most common

include evidence, concession, elaboration, motivation, and volitional result (Mann et al., 1989, p.18). The number of relations is not fixed, more can be added if necessary. Some of these relations are illustrated in the example below (see p. 60).

Notationally, individual relations are represented as a *schema*, a small pattern which 'indicates how a particular unit of text structure is decomposed into other units' (Mann and Thompson, 1986b, p.2). Figure 2.6 displays a generic schema. In a schema, lines on the horizontal axis show the spans, with a vertical line pointing to the nucleus and a curved line linking the nucleus to the satellite; the particular relation defined by the schema is specified over the curved line.

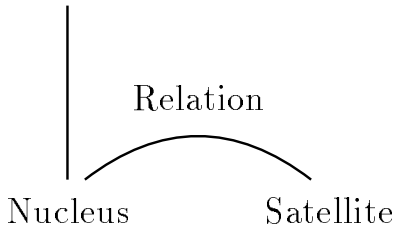


Figure 2.6: Generic schema (Mann and Thompson, 1986b, p.2)

Procedures

In analysing a text for its rhetorical structure, the text is first divided into units, usually independent clauses. The size of the unit can vary, though, from a lexical word to several paragraphs (1987a, p.16). The requirement is that units be 'relatively theory-neutral' and have 'functional integrity' (Mann and Thompson, 1987a, p.16).

After the text has been divided into units, the next step is to identify the spans and relations between spans. The identification can be top down, that is, by progressively refining a larger unit, or bottom up, by aggregating smaller units into larger ones, or both. The mandatory requirement is that

the analyst must ask 'at each point whether the relation definition plausibly applies' (Mann and Thompson, 1987a, p.16).

In RST the subjective nature of the analytical process is given support. Judgements are necessarily subjective because they depend on the analyst's knowledge of culture, society and language usage, which is also subjective. Therefore, in RST subjectivity is not a shortcoming; rather it is incorporated into the system as a natural feature of the interpretative process. In addition, RST analysts believe that any one RST analysis is only one of the possible analyses. For instance, Mann et al. (1989, pp.32ff) provide a number of alternative analyses of their texts. Alternative analyses are acceptable because analyses are made based on *plausibility judgements*, that is, 'each analytical statement should be read as 'it is plausible that the writer intended...' (Mann and Thompson, 1987a, p.24).

Example analysis

The literature on RST provides several examples of full texts analysed in detail. In Mann et al. (1989) it is claimed that over 400 texts had been analysed during the construction of the model. To illustrate a typical RST analysis, an example text is presented in figure 2.7; the units are numbered for convenience. The top levels of the rhetorical structure are displayed in the diagram in figure 2.8 on page 62.

(1) Farmington police had to help control traffic recently (2) when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriot Hotel. (3) The hotel's help-wanted announcement – for 300 openings – was a rare opportunity for many unemployed. (4) The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie. (5) Every rule has exceptions, (6) but the tragic and too-common tableaux of hundreds or even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs, (7) not laziness.

Figure 2.7: Example text (Mann and Thompson, 1987a, p.13)

In the text in figure 2.7 the top-most relation is that of background. The nucleus is from units 4 to 7, while units 1 to 3 are the satellite. The background relation stipulates that 'the reader will not comprehend the nucleus sufficiently before reading the text of the satellite' (Mann and Thompson, 1987a, p.54). This means that the rationale for considering units 4 to 7 to be the nucleus is that the reader would not understand why people were queuing unless they knew about the job opening and the many unemployed. At the next level down, two other relations are specified. The first, 'volitional result', holds between the nucleus in units 2 and 3 and the satellite in unit 1. The volitional result relation specifies that the 'satellite presents a volitional action or a situation that could have arisen from a volitional action' (Mann and Thompson, 1987a, p.62). In the text, the fact that the police had to control traffic is seen as the result of the people lining up. The second relation is 'evidence' which links the span from units 4 through to the end of the text, with unit 4 being the nucleus, and units 5 to 7 being the satellite. According to the evidence relation, 'the reader's belief of the nucleus is increased' (Mann et al., 1989, p.12), which in the text means that the reader's belief that there was a message being carried by the crowd lining up for jobs (nucleus) is likely to be increased if the information about the lack of jobs in the city is presented (satellite).

Conclusion

Rhetorical Structure Theory forms a comprehensive scheme for text analysis. A large set of relations are defined, which helps in the application of the scheme to a wide range of texts. At the same time, the objective description of each relation is not meant to eliminate the subjectivity inherent in the identification of the relations in texts, which is a realistic statement given the interpretative nature of the relations. Although the application of

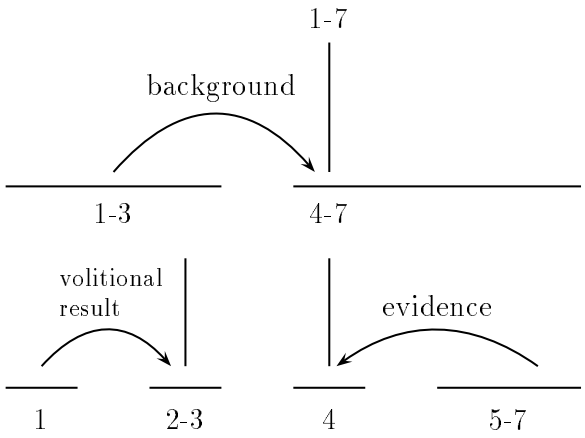


Figure 2.8: Top levels of RST for example text (Mann and Thompson, 1987a, p.14)

RST to text analysis is amply documented, the fact that RST was originally developed for text generation by computer must not be forgotten. By being applicable to both analysis and generation, RST is perhaps unique as a comprehensive model of discourse organisation.

Implications

The applicability of Rhetorical Structure Theory to segmentation is obvious in that implicit in the identification of relations is the division of the text into spans. Each level of the analysis presents the opportunity for the analyst to place boundaries across the text. Each unit may be taken as a single segment, and larger segments may be formed by combinations of units. The applicability of insights deriving from RST are not restricted to descriptive text analysis; the segmentation capability of rhetorical units has been shown to yield good good results in automatic summarisation tasks. Marcu (1997) showed that good summaries could be created by first segmenting texts into rhetorical units and then summarising segments individually. This suggests that RST has an inherent segmenting feature which could be explored in a number of tasks.

2.4.4 Goutsos

A more central place for segmentation is provided by Goutsos (1996a). The goal of Goutsos's (1996a) study is to investigate the linguistic features which indicate segmentation in a corpus of written texts. The study presents a detailed analysis of linguistic strategies for changing and maintaining segments in a corpus of 29,000 words. The corpus contains data from academic papers, extracts from non-fiction, popular science books, newspapers, and editorials from *The Guardian*.

Linear segmentation and macrostructure

There are two models of discourse organisation according to Goutsos (1996a). One is macrostructural and is based on the notion that discourse is organised at a 'deep' or 'schematic' level (Martin, 1989; Swales, 1990; van Dijk, 1980). Approaches which fall into this category are termed 'schematic' or 'propositional' since they see discourse as analysable in terms of 'the semantic relations between constitutive units of predications or propositions' (Goutsos 1996a, p.502).

By contrast, the model proposed by Goutsos (1996a) sees discourse as organised sequentially. As such it is based upon the frameworks proposed by Schegloff and Sacks (1973) for conversational analysis and by Sinclair and Coulthard (1975) for discourse analysis of classroom interaction. Goutsos (1996a) is especially interested in incorporating insights from sequential features of spoken and written discourse and applying them to the problem of explaining segments. As such, his proposal is aimed at accounting for the linear or sequential development of written texts.

Continuity and discontinuity

The notions of continuity and discontinuity are introduced to account for segmentation. Goutsos (1996a, p.504) observes that:

> an equally important task for the writer is to indicate discontinuity within the larger presupposed continuity of the text. In other words, the writer is faced with the tasks to manage the interaction through discourse in sequential terms and to segment discourse into chunks and indicate their boundaries, i.e. the discontinuity between one another.

Segmentation is therefore understood as a strategy which keeps discourse flowing. Nevertheless, Goutsos (1996a, p.504) warns that the segmentation is not absolute: 'sections cannot be presented as totally new or unexpected, but as more or less continuous or discontinuous with each other.' This is because discontinuity is counterbalanced by continuity in the interest of keeping the text flowing.

The areas of the text which correspond to either continuity and discontinuity are termed 'spans', so it is possible to talk about 'continuity spans' and 'discontinuity spans'. The former are characterized as areas of 'local continuity or stability' (Goutsos, 1996a, p.505) while the latter are areas where '(swift or abrupt) ruptures' occur (p.505). These spans are realized by specific strategies which in turn are signalled by specific linguistic devices (see p. 65).

Reader-writer interaction

Segments are a natural feature of texts because they come about as a result of the need for managing the interaction between reader and writer.

The interaction is maintained by providing elements in the text which signal both continuity and discontinuity. Typically, continuity is assumed to be the unmarked situation between adjacent sentences (Brown and Yule,

1983; Thompson, 1996). However, Goutsos notes that continuity must be signalled all the same, so that the reader is reassured that they are 'on the right way' (p.505).

Strategies and techniques

There are two fundamental types of sequential strategy in discourse: continuity and discontinuity (or shift). The continuity strategy is indicated by continuation techniques, which is is established by an utterance having a link to the preceding span. Discontinuity strategies are indicated by one or more of three techniques: introduction, framing (optional), and closure (optional). Introduction, as its name indicates, takes place as the 'opening of a continuation span' (Goutsos, 1996a, p.512). Framing involves the simultaneous ending of 'a continuation span and the starting of an ensuing transition span', and its function is to 'shift the scene by setting a new domain for the following text' (Goutsos, 1996a, p.508). Closure 'provides an advanced warning for the upcoming closing of the current continuation span' (Goutsos, 1996a, p.514). Goutsos (1996a, p.509) argues that each technique could be signalled by a metadiscourse comment. For introduction, the metadiscourse comment would be 'Now I am focusing on a specific aspect', for framing 'Now I am opening a new domain', for closure 'I am about to finish', and for continuation it would be 'I am continuing along the same lines'.

A sample analysis a newspaper editorial is presented in Goutsos (1996a, p.519ff). An excerpt is reproduced in figure 2.9 on page 67. In the example in the figure, framing is indicated by the question as well as by the paragraph break at the very beginning of the text (an orthographic signal). The introduction that follows is predicted by the previous question, although it is not an answer proper. The continuation is then realized by pronominalization ('they'), and local cohesion ('these'). Later on, closure occurs as a result

of an elliptical sentence ('a very small row of beans') and by the paragraph break.

Conclusions

The detailed proposal presented in Goutsos (1996a) can be summarized as follows. Discourse is subject to two textual or sequential strategies: continuity and discontinuity (or shift). These strategies are realized by four techniques: closure (optional), framing (optional), introduction, and continuation. The techniques of closure, framing and introduction are employed in order to create discontinuity, whereas continuation is used to create continuity. Each of these techniques is in turn realized by surface signals such as paragraph breaks, discourse markers, cohesion and tense.

Implications

The claim that segmentation is a surface phenomenon has relevance to the investigation presented in this thesis in that it corroborates the view that segments should be signalled linguistically. Furthermore, it implies that segments are not arbitrary; rather they are motivated by considerations of textual continuity. The presence of lexical cohesion among the possible types of linguistic signals which realize segments also lends support to the notion held in the present thesis that lexical cohesion is related to segmentation.

The formalization of the abstract notions of textual continuity and discontinuity is important since it puts into a broader context the problem of why there should be segments at all. The answer offered by Goutsos (1996a) is that segment breaks (or discontinuity) are not a different aspect of textuality from continuity. Continuity and discontinuity are two aspects of the same phenomenon, namely the sequential development of text.

Another contribution is the contextualization of segmentation in terms of

(**framing**)	¶1 (sent.1)	What do the words 'militarily insignificant' mean?
(**introduction**)	(sent.2)	They fall, repeatedly now, from the lips of General Schwarzkopf himself.
(**continuation**)	(sent.3)	They describe, first, the Scud attacks on Israel, then the capture and re-capture of Khafji.
		(.)
(**continuation**)	(sent.5)	These are diversions which don't affect the weight of weaponry or strategy.
(**continuation**)	(sent.6)	They're rows of beans.
		(.)
(**closure**)	¶2 (sent.10)	A very small row of beans (though a lot of breakfast TV).

Figure 2.9: Excerpt analysed for strategies and techniques (adapted from Goutsos (1996a, pp.519-520)

---

reader-writer interaction. Goutsos (1996a) argues that while it is true that continuation is assumed between adjacent sentences, it does not follow that continuation does not need to be signalled. The reason it must be signalled has to do with the necessary task of keeping the reader informed of whether he/she is interpreting the text as expected. In addition, continuity signals make discontinuity signals more noticeable.

A limitation of the approach presented by Goutsos (1996a) is that it cannot be fully automatized. As a consequence, there will be different interpretations of the role of the same linguistic signals by different readers. The fact that there may be different analyses is regarded as a deficiency by Goutsos himself (1996a, p.528) when he discusses the inadequacy of trying to determine discourse topic. Unfortunately, his criticism applies to his own approach. It cannot be argued here that subjectivity must (or can) be eliminated but the fact that Goutsos's (1996a) approach relies heavily on the reader's interpretation of linguistic devices implies that other readers may find different segments in the texts.

2.4.5 Davies

Another approach to use surface markers is the model of discourse organisation formulated by Davies (1994). The model, which is systemically-inspired, aims to be comprehensive in that it seeks to account for a range of strategies authors utilize for managing interaction with readers.

Primary elements

The model presented by Davies (1994) includes three primary elements of written discourse: topical, interactive, and organising. Each element performs specific communicative functions. Topical elements have the function of informing, and optionally of presenting 'writer Viewpoint'. The function of Interactive elements is mainly to contextualize topic, and 'establish goals and negotiate writer and reader roles' (Davies, 1994, p.172) in addition to informing. And organising elements perform the function of 'pointing forwards, backwards and sideways to the structure and progression of the discourse' (Davies, 1994, p.172).

The three elements realize different metafunctions (Halliday, 1985). Topical elements express the Ideational metafunction, Interactive elements the Interpersonal metafunction, and organising elements express the Textual metafunction. The elements are considered to map onto distinct metafunctions in view of the different resources which realize each element. These resources are discussed below.

Theme and writer's roles

The elements are identified primarily by the choice of theme. The initial categorization draws on previous work by Berry (1989) who devised a classification of theme into Interactional and Topical Themes. The former are typically realized by selection of personal pronouns, while the latter are iden-

tifiable by reference to 'writer's topic area'. Davies argues that this distinction assists in describing 'the way in which writers move, in their negotiations with their readers, from adopting an interactive, to an informing role' (Davies, 1994, p.172).

To the initial classification proposed by Berry, another type of theme is added: Discourse Theme, which is realized by including mention to the text itself (e.g. 'this paper'). Taken together, the three kinds of theme, viz. Topical (or informing), Interactive, and Organising, form the basis for the postulation of the three discourse function of informing, interacting, and organising. These three functions are said to be useful heuristics for 'tracking writer roles' in written texts.

Redefinition of Theme

The traditional definition of Theme in systemic linguistics is that theme consists of clause initial constituents up to and including the first ideational element. This definition leads to a major categorization of Theme as either marked or unmarked. Davies (1994) finds this definition too restrictive and stretches the boundary of Theme to include the grammatical Subject. Since grammatical Subject is identified with topic (Davies, 1994, p.174), this redefinition of theme allows for the identification of topical themes.

In this new definition, the elements preceding the grammatical Subject are treated as 'Contextualizing Frames'. More specifically, Contextualizing Frames include 'all pre-Subject Thematic elements, including dependent clauses in first position' (Davies, 1994, p.174). Contextualizing Frames present interactive and discourse themes.

Units and threads

The analysis focuses on how major segments of text relate to writer roles. Two types of segment are described: *threads* and *units* depending on how long they are. If 'one of the functions is consistent over three or more sentences or independent clauses' this constitutes a thread (Davies, 1994, p.174); if one of the functions is consistent over two sentences or independent clauses, then a unit is demarcated.

Consistency is achieved by recurrence of the same type of theme as well as other linguistic features. Interactive units or threads typically contain, in addition to interactive theme choices, features such as modality and evaluation, mental and verbal processes, and superordinate lexical items and short lexical chains relating to Topic. Among others, the following features commonly recur in organising units or threads: discourse themes, headings and sub-headings, and expressions of opposition. Finally, topical units or threads contain, for example, topical themes, declaratives, similarity and identity lexical chains (for a complete listing of features see Davies (1994, p.175).

Example

The data analysed by Davies (1994) consist of a booklet designed to promote the University of Liverpool. The goal of the writer is thus described as that of persuading the reader to consider studying at Liverpool University. One objective of the analysis is therefore to show how this goal is reflected in the choice and placement of units and threads across the text. Part of the data analysed is reproduced in figure 2.10 on page 72.

In the fragment in figure 2.10 three individual units are identified. The first unit is organising. It contains discourse themes ('This section' and 'The next few paragraphs') as well as a similarity chain created by the sub-topics of Liverpool (e.g. 'sport and entertainment', 'local attractions', 'shopping,

food and drink and the Liverpool people'). The second unit is considered interactive because of the presence of interactive themes ('If Liverpool does mean sport or entertainment to you, then' and 'What could be more alive than'), and also because it offers a range of roles to the reader ('part of the crowd', 'adherents of other sports'). Finally, the third unit is topical mainly because of the various topical themes occurring in succession (e.g. 'Liverpool Cricket Club', 'St Helen's RLFC, Waterloo RUFC and Liverpool St Helen's RUFC', etc.), but also because of the predominance of material processes, in contrast to the verbal and mental processes in the preceding units.

1	organising unit	**This section** sets out to show some of the features of Liverpool which make it an attractive place to spend your student days. **The next few paragraphs** will tell you a little about sport and entertainment, local attractions, enjoyment further afield, shopping, food and drink and the Liverpool people.
2	Interactive unit	*If Liverpool does mean sport or entertainment to you, then* **you** will not be disappointed. **You** can be part of the crowd at Anfield or Goodison Park, where you will see soccer of the highest quality; **no other city** has such a record of success in League or Cup competitions. **What** *could be more alive than* these grounds when the first team is playing at home? **Some** would say 'Aintree in April' for when the Grand National steeplechase takes place, the northern outskirsts of the city teem with cosmopolitan life. *But while these are well-known sporting images of Liverpool*, **adherents of other sports** will have little difficulty in satisfying their needs.
3	Topical Unit	*For example*, **Liverpool Cricket Club** hosts some of the Lancashire's County Championship matches each season; **St Helen's RLFC, Waterloo RUFC and Liverpool St Helen's RUFC** all play within Merseyside; **the Royal Birkdale Golf Club** has brought the British Open Championships to the region; **the Wirral International Tennis Tournament** attracts world-class players, **the city** claims international champions in archery and the martial arts and of course, **water sports** are well represented.

Key:
Bold type indicates Subject/Topical Themes
Italics indicates Themes functioning as Contextual Frames

Figure 2.10: Example of text fragment divided into threads and units (Davies, 1994, p.177-178)

Conclusion

The scheme for analysis of written texts developed by Davies (1994) allows for the identification of segments of texts based primarily on the identification of theme choices and process roles. The scheme takes into account the role of writers in organising the text and in presenting their viewpoint on the subject matter. By categorizing theme as topical, interactive, and organising, the scheme provides objective criteria for interpreting theme choices as means for topic and discourse management.

Implications

The system of analysis based on theme choice presented by Davies (1994) has implications for the analysis of segments in text because it is primarily a system which works from the data up to the segments. The only categories which are imposed on the data are those which come from the classification of the data in terms of systemic constituents. However, because these systemic constituents are based on clause boundaries, they do not predefine the boundaries to be assigned to the text.

Moreover, (marginally) built into the system is the notion that lexical cohesion helps define units of text. Although the precise way in which one makes the jump from identification of chains and definition of boundaries is not made explicit, it is important that the system recognises the role of lexical cohesion in providing a foundation for segments. In all, the implications of Davies's (1994) study is that it is possible to conduct a data-based segmentation of texts taking into account lexical cohesion.

2.4.6 Giora

A researcher who has a rather different stance on the relationship between segmentation and topic from that adopted by Davies is Giora (1983). The goal of Giora's (1983) study is to show that boundaries between segments such as chapters and paragraphs do not occur at the point where a new topic is introduced but later after the topic has been introduced. This view contrasts with the prevailing notion that a new topic is associated with the beginning of a new segment. The framework for the analysis is however Functional Sentence Perspective (Daneš, 1974), which shares common assumptions with the Theme–Rheme framework used by Davies.

Topic introduction and segmentation

A link is made by Giora (1983) between segmentation and topic introduction. She argues that the motivation for segmentation is 'the need to change or shift discourse topics' (p.156). As a result, she investigates the relationship between new topics and their placement in paragraphs and chapters.

A discussion is provided of the differences between introducing a new topic at the end of a segment (paragraph or chapter) or at the beginning. A topic is defined in terms of 'frames' or 'aboutness', that is, 'that which the segment can be interpreted as being about' (Giora, 1983, p.156).

Rhematic position and segmentation

The introduction of a new topic in rhematic (segment-final) position is discussed as being a strategy whereby a new topic is given foreground position. Foregrounding occurs because according to Daneš (1974) segment-final or *rhematic position* is where new information is normally placed, whereas segment-initial or *thematic* position is normally reserved for given information. In this manner, by placing new topics at the end of segments authors

give topics the status of new information. At the beginning of the next segment, the topic is reintroduced but this time it has the status of given information which is coherent with the expected pattern of given-new discourse development.

Analysis

The data analysed in Giora (1983) consist mostly of literary fiction and poetry. She shows that her data do not support the belief that introducing a topic at the beginning of a segment is the unmarked option. Various examples of introductions of new topics at the end of a segment are provided. For instance, chapter three of 'Alice in Wonderland' ends with the following sentence: 'In a little while, however, she heard *a little pattering of footsteps* in the distance and she looked up eagerly (...)', and chapter five begins with 'It was the *White Rabbit, trotting* slowly back again (...).' (p.175, original emphasis).

Conclusions and implications

Giora's major conclusion is that segments end with a new topic, while subsequent segments begin with the new topic introduced in the previous segment. This is taken to be the unmarked pattern of discourse development since it is coherent with the proposition that new information normally occurs at the end of a sentence (in the rheme).

One important aspect of Giora's (1983) study is that it is not aimed at providing segmentations of texts but rather concentrates on explaining one of the possible principles which seem to underlie segment divisions. In this manner her approach is valuable because it suggests that, for instance, chapter divisions are not arbitrary, instead they indicate major shifts of topic.

A text is considered to have many levels: the line, the sentence, the

paragraph, the chapter, and even the text as a whole (Giora, 1983, p.164). Each of these levels is formed by segmentation, and therefore a sentence, a paragraph, a chapter, etc. are segments. This definition has implications for the treatment of sections in this thesis as segments since it endorses the present view that sections are one of the many kinds of segments in a text.

One can point out several limitations of the study. The most serious has to do with the extension of Functional Sentence Perspective to account for all levels of discourse organisation. Although theoretically pleasing, the analysis relies on the premise that there is a regular pattern of discourse development which applies equally to sentences, paragraphs, or chapters, and that such pattern can adequately be accounted for by theme-rheme or given-new progression. While this may be true in the many examples provided throughout the paper, the question still remains of whether all levels are governed by the same principle. It appears this is assumed to be the case, and if it is then one has to believe that all sentences end with new information, and so does the paragraph which these sentences are in, and so do the chapters which they are in (if any) and so does the text which they are part of.

2.4.7 Longacre

The model of discourse presented by Longacre (1983) is intended to represent a *grammar* of discourse, within the larger framework provided by tagmemics. His goal is to account for the relationship between form and function of language in context, not the referential content of discourses. Being formulated as a grammar, the model allows him to make predictions about the structure of texts in general.

Episodes and structures

Central to Longacre's model of discourse is the notion of *episode*, which is a componential unit of narratives. Longacre (1983) describes four types of linguistic devices which may mark the onset of an episode: time horizons in succession, back-reference, conjunctions, and juxtaposition. In a later study, Darnton (1987) adds three more linguistic devices: introduction to locational reference, introduction of a new participant, and setting proposition. Importantly, Longacre admits that episode markers work when the episode boundaries have already been placed in advance. Thus, the identification of episodes depends largely upon intuition.

Longacre (1983) distinguishes two kinds of structure in narratives: a surface structure, and a notional or deep (Longacre, 1976) structure. The former can be described by its surface linguistic realization, whereas the latter is describable in terms of the semantics of story grammars (Darnton, 1987, p.29). The categories in the surface structure are nine: Title, Aperture, Stage, Pre-peak episode, Peak, Peak[1], Post-peak episode, Closure, and Finis. The categories in the notional structure are only 7: Exposition, Inciting moment, Developing conflict, Climax, Denoument, Final suspense, and Conclusion. As figure 2.11 on the following page indicates, there is not a one-to-one relationship between the elements in the two structures. The notional categories of Climax and Denoument can be realized as one of a series of surface categories, and Title, Aperture, and Finis have no counterparts in the notional structure.

Plot and Peak

The model of monologic discourse postulated by Longacre centres around the classical notion of plot, or a series of linked events. The elements of the plot

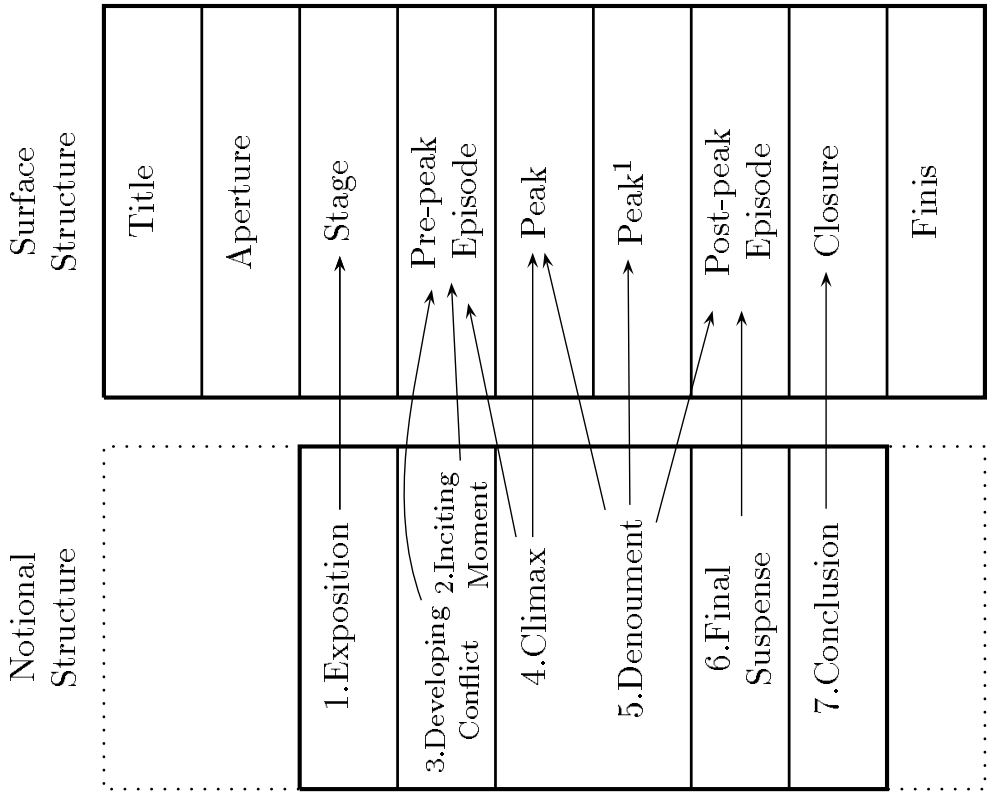


Figure 2.11: Longacre's notional and surface structures (adapted from Darnton (1987, p.28)

are represented in the notional structure (see previous paragraph). Plot is essentially characteristic of the narrative type of discourse. In classical times, plot was seen as applying to the structure of drama, but Longacre extends plot to account for narrative generally.

The surface structure category of 'Peak' is the central element in Longacre's model. In general terms, a peak is 'a zone of turbulence' in the flow of discourse (Longacre, 1983, p.25). More specifically, the term 'peak' is used to refer to 'any episode-like unit set apart by special surface structure features and corresponding to the Climax or Denoument' (Longacre, 1983, p.24). A narrative is then viewed as a sequence consisting essentially of pre-peak, peak, peak[1], and post-peak.

The identification of surface features is central to locating peaks. Five devices which function as indicators of Peak are provided by Longacre: rhetorical underlining, concentration of participants, heightened vividness, change of pace, and change of vantage point and/or orientation. There are certain linguistic characteristics associated with each of these devices. For example, rhetorical underlining is expressed by the employment of parallelism, paraphrase and tautologies; heightened vividness by a number of different kinds of shift: tense shift, nominal/verbal balance shift, shift to a more specific character; and change of pace involves shift in the length of sentences and paragraph or in the proportion of 'connective material'.

Plot and other discourse types

A plot-based structure is also posited to account for all discourse types, not only narratives. The basis for assignment plots to other discourse types lies in the folk notion of *struggle*. Longacre (1983) believes that just as narratives present a confrontation and a resolution of conflict, the same confrontation and resolution exist in other discourses in the form of a struggle. In pro-

cedural discourse, the struggle is 'to accomplish the goal of discourse, to carry through an activity, or to produce a product' (Longacre, 1983, p.38). In expository discourse, the struggle is to achieve clarity; and in hortatory discourse the struggle is to persuade or dissuade hearers.

Although Longacre (1983) argues that plot is a suitable framework for analysing other types of discourse, he suggests that the range of options for marking peak in non-narrative discourse is not as wide. In expository and hortatory discourse the most important device for marking peaks is rhetorical underlining. Thus, it is argued that in the course of explaining a subject (expository) or giving advice (hortatory), speakers make extensive use of parallelism and paraphrase.

Implications

The segmentation of narratives into episodes as proposed by Longacre is automatable since the list of linguistic devices originally provided by Longacre (1983) and later complemented by Darnton (1987) provides a starting point for the identification of episode boundaries by computer. A computer-assisted procedure could be designed to locate these devices in texts, and then use this information to place episode boundaries. A similar strategy has already been pursued in previous research in segmentation. Passonneau and Litman (1995, pp.14-15) looked at the role of 'cue words' (Cahn, 1996; Hirschberg and Litman, 1993) such as 'now' and 'and' in marking segment boundaries. Their research suggests that when combined with other linguistic devices (referential noun phrases and pauses), cue words can improve the performance of their segmentation algorithm (Passonneau and Litman, 1995, p.17). However, as mentioned above, in Longacre's model the linguistic devices are not unambiguous episode boundary markers (Darnton, 1987). This could be a drawback in the automatic identification of segments

using Longacre's model.

2.5 Conclusion

In this section, final comments will be made about the review presented in this chapter. The presentation of the individual studies was carried out in terms of whether contributions focused on content or surface markers for segmentation. According to this initial division, studies such as Cloran (1995), Bhatia (1993), Swales (1990), and Hasan (1989) all favour the demarcation of segments by judging the contents of discourse. The analyst working in these approaches normally makes use of his/her intuitive knowledge of the situation, his/her perception of the topic being addressed, and his/her familiarity with the culture in which the discourse is embedded. In contrast, research by Pitkin (1969), Hoey (1983), Mann and Thompson (1986b), Goutsos (1996a), Davies (1994), and Giora (1983) look at ways in which the analyst can draw on surface markers to identify segment boundaries. These models place greater emphasis on the presence of surface linguistic features such as repetition, shifts of tense, typographical signals, and thematic development in order to locate segment boundaries.

There are other characteristics shared by the discourse analytical approaches presented above which cut across the initial classification into content and surface. At least five major trends can be observed which represent major points of contact among discourse analytical approaches in respect to issues such as data handling, data coverage and orientation towards the data. These tendencies are therefore of a more practical nature, as they are related to the ways in which individual approaches to discourse analysis have been operationalized.

The first trend could be described as 'labelling'. Not only do most ap-

proaches segment discourses but they also provide labels for the segments. The functionality of labels lies in the fact that they provide a definition of the contents of the segment both in terms of the actual occurrence in a particular text and in future texts (if the label is part of a model; see the discussion on deductive models below). If there is insufficient evidence to support the choice of a particular label, the act of labelling loses its purpose.

The second trend is that most models are deductive, that is, they make *predictive* statements about the recurrence of segments in other texts. This is expected in so far as these approaches are formulated as models, that is, as abstract representations which are meant to explain all or most of the instances of a given discourse or text type.

A third trend refers to the lack of validation of segments. Invariably, models of discourse propose segmentations which are by definition correct and acceptable. Validation could be achieved by checking the results of the analysis with other possible analyses of the same data. However, because most approaches rely on subjective judgement for identification of segments, the problem then would be deciding which of the possible analyses would be the correct one.

A fourth trend which can be observed across most approaches is that which relates to the amount of data which most approaches have actually been applied to. With a few exceptions (RST for instance purports to have been tried on hundreds of texts), most discourse analytical approaches have restricted themselves to very few instances of actual discourses.

A final trend which is related to the number of discourse tokens is that most discourse analysis has been restricted to the investigation of short passages. Admittedly, there is no agreed definition of what counts as a short passage; while Renouf and Collier (1995) speak of a long text as being typically longer than 60 sentences, Passonneau and Litman (1993) consider a

passage to be a long one if it exceeds 200 words. In general, though it has been said that the typical amount of data in discourse analysis is that which can fit on the blackboard (Phillips, 1985).

These major trends suggest that while it can be argued that discourse analysts have typically occupied themselves with segmentation, they have done so with a particular kind of segmentation in mind: segments are labelled, located by inference, identified in short passages, and the whole model is tried on a few texts. While approaches which share all these characteristics are uncommon, at least one of these features can be found in each one of the approaches reviewed in this chapter. Arguably, these characteristics apply to the vast majority of discourse analytical approaches.

The problem with these restrictions is that they do not operate independently; rather they seem to influence each other. As mentioned above, labelling as such is not a restrictive feature of discourse analysis. Quite the opposite, labelling offers the opportunity for a model to explain the contents of a particular discourse as well as help find similar segments in other discourses. However, if the choice of label during the formulation of the original model did not carefully rest upon the examination of a large number of discourses, then it may be that the set of segment labels will prove inadequate for the segmentation of other discourses. And since labels are part of models, labels may be forced inadequately onto discourses which do not lend themselves to analysis by a certain model. Finally, there is the obvious observation that while models are finite, the number of texts in existence is not, therefore there will be texts whose segmentation cannot be adequately accounted for by any of the existing models available.

There is room for another perspective on segmentation, one which provides an alternative to the restrictions mentioned above. In this alternative perspective, a concern with extensive coverage should be central. The

alternative approach should be able to cope with both large quantities of exemplars and with large quantities of input tokens in each exemplar. In other words, an alternative approach would be considered suitable if it were designed in such a way that it made it possible to segment a corpus of long texts.

Another central concern should be with the orientation towards the data. An alternative approach should be inductive, that is, it should proceed by working from the data up to segments (i.e. a bottom-up orientation). Or, in the words of Salton (1988, p.387) a bottom-up, data-driven approach is one in which 'the individual text words are initially considered, and attempts are made to group them into successively larger, more comprehensive components'. By contrast, as pointed out above, discourse is typically analysed the other way round, that is, by applying a set of constituents (i.e. top-down orientation). Although at the time of formulation discourse models may be said to have followed a bottom-up orientation in that the category labels they propose are not invented, at the time of application models are deductive. During application, the role of the analyst is typically restricted to fitting his/her data into the labelled slots or categories proposed by the model.

Also central to this new perspective would be a concern with adequate validation of the results of the analysis. As mentioned above, approaches to discourse analysis normally evaluate their results internally if at all. A more satisfactory validation could be obtained by evaluating the results of the segmentation against an independent criterion. One such independent criterion could be 'an expert analysis'; however, expert analyses are available only for those data which have been investigated. New original data would not get the benefits of expert analyses. Another independent criterion would be not to evaluate the analysis against other analyses but against *the data itself.* In the case of written texts, this would take the form of checking whether the

segmentation matches existing divisions in the text such as paragraphs, sections or chapters. One advantage of these orthographic divisions is that they play an important role in signalling breaks in the sequential flow of written discourse (Goutsos, 1996b, p.82). Admittedly, no models of discourse organisation have been proposed which are aimed at providing an account of how texts are divided into sections or chapters, therefore it would be unfair to evaluate them against this criterion. Nevertheless, since major text divisions provide a window on the decisions taken by the writer(s) on how the text is organised into major discourse constituents, failure to provide at least a partial approximation of these constituent boundaries would indicate that a particular model fails to address the issue of how writers organise their texts, which is after all an underlying preoccupation of most discourse analysis models.

Adopting existing text divisions as the validation criterion is not without problems, though. First and foremost is the problem of which text division to adopt. When faced with the same decision, Phillips (1985) dismissed paragraphs on the grounds that they are 'unsystematic'. Sections were also dismissed because they are too flexible in length for comparative purposes, and they are also absent from many text types. He settled on the chapter because arguably the chapter has cognitive validity, that is, it 'is in the mind of the author' (Phillips, 1985, p.61). Another related problem is what to consider as a match, which in turn relates to how to carry out the matching. Although seemingly unproblematic, the issue of matching two segmentations is particularly complicated given that it is conceivable that many possible mappings can occur not just one to one boundary matches. The problem of how to compute matches is discussed in more detail on page 281 ff.

The criteria adopted by Phillips (1985) can be reinterpreted. First, cognitive validity can be claimed for sections as well. As Lorch and Lorch (1996)

indicate, headings seem to help comprehension of written texts, so sections as well as chapters can be claimed to be equal in this respect. Further, as Phillips (1985, p.124) himself acknowledges, sections provide a window onto text contents, so he himself does not take them to be arbitrary but related to the 'aboutness' of the text.

Other studies have also shown that sections are self-contained and that section divisions are not arbitrary. Gledhill (1995) showed different collocational patterning in different sections of corpora of cancer research articles. Swales (1990) identified several linguistic features which distinguish sections in research articles. For instance, in introductions '*that*-verb complements' are very frequent and passive voice is rare; by contrast, in methods sections the opposite picture applies: '*that*-verb complements' are very rare and passive voice is frequent. Similarly, Biber and Finegan (1994) investigated the occurrence patterns of selected linguistic features across sections of medical research articles and found that different sections have different linguistic profiles. Each section has substantially different mappings onto the five discourse dimensions posited by Biber (1988). For example, methods sections tend to be more 'informational' than discussion sections (dimension 1) and more 'impersonal' than results sections (dimension 5), whereas introductions are found to be both more 'narrative' (dimension 2) and more 'elaborated' (dimension 3) than results sections.

If sections are related to the content of texts, then adopting sections as the validation criterion would have the added advantage of allowing for the investigation of the topical organisation of text. Being able to investigate discourse topic would be a welcome possibility since topic is typically considered to be an intractable notion in discourse analysis:

> theme and especially its near-synonym *topic* are notoriously elusive concepts in linguistics and have been used to refer to a variety of phenomena (...) There is, consequently, no widely-accepted

> definition that could be useful to our purpose of identifying the text's internal structure. (Georgakopoulou and Goutsos, 1997, p.74, original emphasis)

Other related concerns have been voiced in the literature. For example, Sinclair (1991, p.29) argues that 'it is good policy ... to refrain from imposing analytical categories from the outside'. Sinclair (1991, p.29) believes that imposing categories is typical of linguistics as a whole since 'linguistics usually operates with ... abstract categories'

In similar vein, Phillips (1985) believes there are two principles which must be pursued in the analysis of discourse: the need for large-scale investigation without subjective judgement, and the use of computers in the investigation. He argues in favour of a distributional approach to discourse analysis which takes into account those properties of language data which are perceptible by examination of large quantities of data without *a priori* categorization.

Another concern which has been echoed in the literature is that which refers to the lack of explicit guidelines for segmentation. Kozima and Furugori (1993) complain that:

> Most studies on text structure assume that a text can be partitioned into units that have a hierarchical structure. Agreed commonly here is also that each unit plays its own role in the text (...) However, no clear discussion is ever given to the problem of how to partition a text into units'

With respect to validation, proponents of Rhetorical Structure Theory have showed their concern with the risk of considering any one individual analysis as 'the truth' about the text. Mann et al. (Mann et al., 1989) argue that the analysis should be construed as a series of 'plausibility judgements': 'though the analysis is presented as if it were the "truth", each analytical statement in it should be read as *It is plausible that the writer intended...*'

(Mann et al., 1989, p.24). They also recognize some of the dangers resulting from conceiving of the analysis as plausibility judgements (Mann et al., 1989, p.19):

- circularity
- divergence of analysis from actual function
- nonrestrictiveness of the theory
- vagueness
- indefiniteness of analytic outcome

The three main requirements for an alternative approach to segmentation are therefore extensive coverage, inductive orientation, and independent validation. Finding a methodology for segmentation which satisfies all of these requirements is a major concern of this thesis. The first requirement, that of coverage, certainly necessitates that computers be used in the analysis. By using computers, a larger amount of data can be examined. At the same time, the use of computers makes it possible to design a methodology which is inductive. The next logical step would then be to see how previous studies have used computers to segment texts. This is the subject of the next chapter.

Chapter 3

Computers and segmentation

In the previous chapter, it was argued that a first step in providing a methodology which could be applied to segmenting large quantities of data was to make use of computers. In addition, utilizing computers makes it possible to adopt an inductive orientation towards the data while minimizing the role of subjective judgement in the segmentation. The aim of this chapter is to review relevant studies which have dealt with segmentation of texts by computer. At the end of the chapter a summary will be offered of the main trends followed by the studies reviewed here.

3.1 Youmans

The work of Youmans (1991) is concerned with finding natural segment divisions in long texts. He develops a technique called 'Vocabulary Management Profile' (VMP) for segmenting narratives, and looks specifically at the performance of VMP in segmenting literary fiction. The basic mechanism in his technique is the computation of type-token ratios in 'windows' (short intervals) of text.

3.1.1 Vocabulary Management Profile

The Vocabulary Management Profile technique is devoted to the investigation of plot changes in written fiction. Youmans (1991) argues that the rate of introduction of new types corresponds to where major topics begin and end. He calls his type of analysis 'Vocabulary Management Profile' (VMP) and argues that in addition to showing the regular patterns of vocabulary introduction it can also show where the major divisions in the text occur.

An analysis for VMPs is based on measuring lexical density (type-token ratios) in short intervals. In practice, VMPs count the number of new words ('types') occurring in a 'window' of a certain number of running words ('tokens'). A text window is a 'group of words appearing in contiguous positions in text' (Haas and Losee, 1994, p.619). The type-token ratios are computed for 'moving windows', or a fixed portion of text read in one at a time and then moved on along the text one token at a time. The actual placement of boundaries is carried out by examining the plot of the type/token ratios for each window position. Moving windows are used because it is claimed that the traditional plot of type-token rations for the whole text is not sensitive to the kind of changes in vocabulary introduction which is indicative of boundaries.

VMPs had to be fine-tuned by experimenting with different window sizes. According to Youmans (1991), an interval of 35 words provided the best visualization of the 'rhythmicity' across the texts, and therefore it was chosen as the most adequate for the analysis. The fine-tuning is subjective and depends upon the analyst's judgement as to whether the VMP curve indicates what the analyst considers to be important segments in the text. According to Youmans (1991) it is also important that the VMP curve is not too flat or too peaky.

The VMPs for five texts are presented: Joyce's 'The Dead' (first 1189

words), 'Eveline' and 'Finnegans Wake', and Orwell's '1984' and 'Newspeak'. For all these texts, the VMPs indicated what Youmans considered to be major divisions in the narrative. For example, the VMP for 'The Dead' indicated a major division between the initial monologue and the subsequent dialogue. Mostly, the segments which the VMPs indicate seem to signal changes in the plot. It is worth noting that the divisions were not independently verified.

3.1.2 Influence of lemmatisation

It was hypothesized by Youmans that the performance of VMPs could be improved if texts were lemmatized and synonyms were treated as being the same type. Arguably, this would allow for the correct estimation of lexical density within each window since different forms of the same token would not be treated as different types. Youmans makes a comparison between two VMPs for two versions of the same text: one where inflections and synonyms were dealt with and another where the words were left as they appear in print. The results indicate that there is no discernible difference between the two VMPs, which suggests that lemmatisation and thesaurization are not essential.

3.1.3 Conclusions

The author concludes that the VMP technique seems a good tool for discourse analysis since it can help the discourse analyst gain insights into the regular alternations between new and repeated vocabulary which in turn help signal the major constituents of written texts. Youmans (1991) warns that the VMP works as a 'wind sock' for the major constituents in literary texts: it shows where the 'wind is blowing' but it also 'lags behind' or 'jumps ahead' of major structural shifts.

3.1.4 Implications

One of the problems of the VMP technique is that while the divisions it indicates seem to be accurate it does not indicate all of the text divisions in the texts. Therefore the technique cannot be relied upon for providing a full segmentation of the texts. Another possible problem is the arbitrary width of the window which is optimized by checking which value signals the most boundaries. It might be argued that the priorities are reversed: instead of using the technique to find the boundaries, it is as if the boundaries are found first and the VMP is fitted in later.

3.2 Kozima

Another computational approach to text segmentation is presented by Kozima (1993b; Kozima and Furugori, citeyear815). His technique is termed 'Lexical Cohesion Profile' (LCP), and it is designed to segment narratives.

3.2.1 Overview

LCP is specifically designed to deal with narratives, and in so doing to extract the 'scenes' in the narratives. This term derives from a metaphor employed by Kozima (1993a,1993b, and Kozima and Furugori, citeyear815) to refer to the scenes in a film. He uses this metaphor to explain the meaning of a segment in a story to the human readers who take part in the research in order to carry out the task of segmenting texts. The major criterion for locating scenes is that they should be 'contiguous and non-overlapping units' (Svartvik, 1990, p.16). It is also assumed that scenes exhibit coherence, which is measured by computing lexical cohesion. The texts reported to have been segmented by LCP are two: a short story by O. Henry ('Springtime à La Carte') and a biography of Mahatma Ghandi.

3.2.2 Relation to previous work

A basic premise of LCP is that text divisions are signalled by lexical cohesion rather than discourse markers. In this respect, LCP offered a new perspective on segmentation from that offered by previous work on cue phrases or clue words (Cahn, 1996; Hirschberg and Litman, 1993). The idea of looking for shifts in a measure of cohesion is partly inspired by Youmans' (1991) *Vocabulary Management Profile* (VMP), which detects major text divisions by observing changes in the number of new words being introduced in the text. Lexical Cohesion Profile sees it as a weakness that VMP relies on word repetition alone to compute vocabulary shifts, therefore LCP incorporates an annotated dictionary to aid in the identification of repetitions. Furthermore, Kozima (1993a) notes that VMP failed to detect major boundaries in so-called 'high-density texts', that is, texts with a high number of different word forms. In such texts VMP tended to report boundaries where in fact there were none. Again, Kozima (1993a) concludes that instead of repetition of word forms, it is necessary to compute repetition of word senses, which requires the introduction of thesaural information. Although LCP computes cohesion based on thesaural information it does so in such a way that cohesion is not computed in chains. It is argued that the success of segmentation by chain identification depends on text size, since in long texts one is more likely to find long chains which in turn are more likely to break for the simple reason that they are long and not because there are natural divisions in the text. In addition, a long text is more likely to have more chains which naturally leads to chain overlap and consequently to a less clear picture of where possible divisions might be.

3.2.3 How LCP works

LCP works by assessing *mutual lexical cohesiveness*, or the density of lexical cohesion. Density is computed by activation on a semantic network called *Paradigme* designed from a subset of the Longman Dictionary of Contemporary English. This subset consists of entries whose headwords are part of the *Longman Defining Vocabulary*, which in turn is based on an updated list of 2,851 words said to represent the core vocabulary of English for the purposes of foreign language teaching. This electronic reduced version of the Longman Dictionary of Contemporary English is termed *Glossème*.

The calculation of lexical cohesive density is carried out by estimating how close any group of words is in relation to the dictionary definitions in *Glossème*. As a result, the cohesiveness between the more coherent pair of words 'waiter' and 'restaurant' is estimated as 0.176, while that between the less coherent pair 'computer' and 'restaurant' is measured at 0.003. A higher cohesiveness coefficient is taken as an indicator of higher coherence. When plotted in a chart, a sequence of high coefficients should be indicative of a continuous segment, while a shift from high to low coefficients should be indicative of a discourse boundary.

LCP is set to operate with a fixed window size of 51 words. As discussed previously during the presentation of VMP (see section 3.1 on page 89), a window is a portion of text usually smaller than the text being analysed within which computations are carried out. Like Youmans (1991), Kozima also employs a shunting window which moves along the text. The 51-word window was arrived at after trying out other window sizes and adapting its size to capture the least amount of noise while at the same time tuning it to match human segmentation. The author concedes that the optimum window size will depend on each individual text, and also on its 'genre and style' (Kozima, 1993a, p.23); nevertheless the idea of employing a fixed unit per

text still remains. Also, the two texts whose analyses are reported had their LCP calculated in 51 word windows.

3.2.4 Analyses

The analysis of two texts is reported in Kozima (1993a). The first is an adapted version of the short story by O. Henry entitled 'Springtime à La Carte'. The segmentation by LCP was compared to segmentation provided by readers. The readers were told to view the story as if it were a movie and pretend they were directors so that they could insert 'cuts' in the story. The segmentation by LCP yielded 16 segments, and of these, $\frac{1}{3}$ matched the divisions proposed by the readers (roughly 33% precision). It was estimated that the human segmentation breaks agreed with the LCP breaks 60% of the time (60% recall; see section 3.3.4 on page 99 for an explanation of 'precision' and 'recall').

The results of the analysis of the other text, the Ghandi biography, are not discussed numerically in Kozima (1993a). This text was segmented manually by the author and its manual segmentation was later compared with LCP boundaries. The size of the window used for segmenting this text is also 51 words. The LCP values were plotted onto charts, and although an interpretation of the charts is not provided, it is possible by inspection to observe a medium level of correspondence between the breaks placed by the author and the shifts on the chart curve.

3.2.5 Implications

In general, Lexical Cohesive Profile is a valuable contribution to segmentation by computer. The two texts whose segmentation are discussed in Kozima (1993a) give us an indication that LCP can achieve good results. There are a number of limitations, though. One of them is the reliance on a thesaurus.

Comprehensive thesauri are difficult to create and they only work if used in restricted texts such as the adapted version of the short story by O. Henry. In the case of LCP, the thesaurus was tuned to the texts which it was applied to. The validity of the approach offered by LCP will therefore largely depend on the availability of a fine-tuned thesaurus. Another limitation is the use of fixed windows. The rationale for the adoption of a 51-word window is not explained, other than the supposition that this width is appropriate for 'most texts'. The adoption of a window is all the more strange in view of the fact that Kozima (1993a) criticizes previous approaches for utilizing windows. Despite these criticisms, Lexical Cohesion Profile constitutes a major contribution to segmentation by computer in that it has stressed the viability of using lexical cohesion in segmentation.

3.3 Beeferman

Another approach to use windows for computing segmentation is that introduced in Beeferman et al. (1997). Beeferman et al. (1997) present a segmentation algorithm based on the comparison of co-occurrence probabilities in short- and long-range contexts using statistical exponential models. They compare the probabilities of two words occurring together in a narrow co-text (a trigram, or 3-word interval) to their probability of occurring in a wide adaptive co-text (a 500-word interval of text). Although their method is adaptable to finding text-internal segments, their study only reports on the application of the method to finding boundaries between texts in a corpus.

3.3.1 Long- and short-range models

The system proposed by Beeferman et al. (1997) is based on the combination of a long- and a short-range model of co-occurrence probabilities. The

long-range model is the probabilities of words occurring within a moving window of 500 words running along the text. It is also called 'adaptive' because these probabilities are calculated and updated as the contents of the window change; for this reason, Beeferman et al. (1997, p.3) believe this model captures the 'nonstationary features of text'. The long-range model is operationalized in terms of 'trigger words', or words whose occurrence is triggered by the presence of another. For instance, Beeferman et al. (1997, p.3) report that the exponential probability of 'scab' being triggered by 'picket' is 103.1, the highest on their list.

The short-range model is formed by recurrent groups of three words found among the most frequent words in a corpus. It is also referred to as 'static' because it is based on the fixed frequencies of a whole corpus and is not updated for any one single text. Beeferman et al. (1997) report on the extraction of two sets of trigrams from two separate corpora. One is a 38-million-word Wall Street Journal corpus from which trigrams were extracted from the 20,000 most frequent words. The other is a 150-million-word corpus of broadcast news; details are not given of whether the trigrams were extracted from all words in the corpus or from a subset only.

The short-range trigram model is criticized by Beeferman et al. (1997, p.3) as being 'myopic'. The authors argue that the usage of a particular word in a text is conditioned by other words outside the trigram; however the cost of computing and storing clusters larger than 3 words is too high in computational terms and therefore in practical terms trigrams is the best one can get for large corpora.

3.3.2 Segment boundaries

Segment boundaries are inserted by Beeferman et al.'s (1997) algorithm on the basis of a 'relevance measure', or a quantitative indicator of the likelihood

of segment breaks in the corpus. In simple terms, this measure is obtained by comparing the long and short-range probabilities for a given word in the text. Beeferman et al. (1997, p.5) explain,

> one might be more inclined towards a partition when the long-range model suddenly shows a dip in performance – a lower assigned probability to the observed words – compared to the short-range model. Conversely, when the long-range model is consistently assigning higher probabilities to the observed words, a partition is less likely.

The comparison may reveal that the words expected to co-occur according to the corpus actually do appear near each other in the same sentence or in neighbouring sentences, as indicated by their appearance within the 500-word window. And two, that the words commonly occurring in the same sentence or in neighbouring sentences (500-word windows) seem more likely to appear together than their mutual occurrences in the trigrams would lead us to suppose. Beeferman et al. (1997) take both situations as not indicating a text boundary.

In contrast, the comparison may indicate that the words appearing near each other in trigrams cannot be found in the same sentence or in neighbouring sentences. In such cases there is a discrepancy between the static and the adaptive mutual co-occurrence expectancies. This discrepancy is indicative of a boundary, according to Beeferman et al. (1997).

The rationale for placing boundaries as described above is based on the topical organisation of texts. Beeferman et al. (1997) believe that words relating to a given topic normally appear near each other, and this is captured by their short-range model. Nevertheless, some words will appear near each other more or less often depending on where they are in the text, and this alternation is a reflection of the change of topics in the text.

3.3.3 Vocabulary features

The segment boundaries suggested by the 'relevance measure' described above are further aided by what Beeferman et al. (1997) call 'vocabulary features', or the induction by the algorithm of vocabulary occurring near segment boundaries. For example, the word 'incorporated' was found to appear consistently at the beginning of the financial texts in the Wall Street Journal corpus, since only at the onset of reports is the full name of the company mentioned (e.g. 'Acme Incorporated'); later on in the same report, the word 'incorporated' is dropped and the company is referred to as 'Acme'. Thus, the appearance of 'incorporated' boosts the probability of a text boundary. Likewise, the word 'see' increases the probability of a boundary because it is more commonly found at the closing of reports, as an invitation for the reader to read a related story. Examples of words which discount the probability of a text boundary are 'he', since it generally assumes an antecedent, and 'Mr', which is commonly used in Wall Street Journal stories (e.g. Mr Smith) after the full name of the person in question has already been provided (e.g. 'John Smith, president of Acme Incorporated').

3.3.4 Performance metrics

In order to evaluate the performance of their procedure, Beeferman et al. (1997) make use of two measures: recall and precision. These measures (which will be used briefly in our discussion of Beeferman et al.'s study) will be frequently referred to in the remainder of the thesis, and therefore they need to be introduced carefully at this point. In what follows a short discussion on these metrics is provided.

Recall and precision are used in information retrieval to represent the performance of computer systems designed to extract documents from a data-

base following a user's query (van Rijsbergen, 1979). Hence, together with 'fallout' and 'error rate', they are referred to as 'performance metrics'. In segmentation analysis, information retrieval metrics have been employed to indicate the proportion of segment boundaries recognized by a particular segmentation procedure (Passonneau and Litman, 1995, p.11). Passonneau and Litman (1995) define the four metrics in the context of segmentation research as in figure 3.1. The four measures are obtained by computing the number of hypothesized segment boundaries and the number of reference segment boundaries. The former are segment boundaries inserted by the segmentation procedure, whereas the latter are the segment boundaries against which the segmentation will be compared; they can be boundaries proposed by readers, boundaries already present in the text, or even boundaries suggested by another segmentation algorithm.

	Reference	
Hypothesized	Boundary	Non-Boundary
Boundary	a	b
Non-boundary	c	d

Recall = a/(a+c)
The ratio of correctly hypothesized boundaries to reference boundaries;

Precision = a/(a+b)
The ratio of correctly hypothesized boundaries to hypothesized boundaries;

Fallout = b/(b+d)
The ratio of incorrectly hypothesized boundaries to reference boundaries;

Error rate = (b+c)/(a+b+c+d)
The ratio of incorrect hypotheses to the table total;

Figure 3.1: Performance Metrics (adapted from (Passonneau and Litman, 1995, p.12)

Recall is obtained by dividing the number of correctly hypothesized boundaries by the number of reference boundaries, that is, by computing

the proportion of reference segments which match hypothesized segments. Precision, in turn, is obtained by dividing the number of correctly hypothesized boundaries by the number of hypothesized boundaries, or in other words, by calculating the proportion of hypothesized segments which match reference segments. Fallout is obtained by dividing the number of incorrectly hypothesized boundaries by the number of reference boundaries, that is, by computing the proportion of reference segments which do not match hypothesized segments. Finally, the error rate is obtained by adding the number of non-matching segments and dividing by the sum of hypothesized and reference segments.

Recall and precision are the two measures which are most often used in segmentation analysis; fallout and error rate are much less common, and therefore they will not be referred to any further. One reason why recall and precision are so common in segmentation research is that they provide complementary perspectives on the performance of a particular segmentation technique. A perfect score on recall indicates that the procedure has identified all of the reference segments in the text or texts. A perfect score on precision shows that the procedure has only inserted segment boundaries that matched reference segments. Thus, 100% recall and precision indicates that the segmentation procedure inserted segments at the places where there were reference segment boundaries only. However, in practice this rarely occurs; segmentation procedures do make mistakes and they insert segment boundaries at places where there are no reference segments, and conversely, they will fail to place boundaries where there are reference segments.

For example, suppose a particular segmentation procedure places 5 segment boundaries in a text in which it was found that there were 10 reference segments. Of the 5 segments, 3 match a reference segment. In this case, recall is 30% ($3 \div 10 = 0.3$), and precision is 60% ($3 \div 5 = 0.6$). On the

other hand, if the text had only 6 reference segments, then recall would be higher, 50% ($3 \div 6 = 0.5$), and precision would still be 60% ($3 \div 5 = 0.6$). However, if the segmentation procedure did not place 5 segment boundaries, but 10, recall would still be 50% ($3 \div 6 = 0.5$), but precision would then be lower, 30% ($3 \div 10 = 0.3$).

A limitation of information retrieval metrics is that a 'segmenting tool that consistently comes close – off by a sentence, say – is preferable to one that places boundaries willy-nilly' , yet both would have the same recall and precision rates (0%) (Beeferman et al., 1997, p.8). In other words, recall and precision do not represent 'near misses' (Passonneau and Litman, 1995, p.11). Thus, it is possible to trade precision for recall by inserting more boundaries in order to raise the chances of matching more reference segments. Ultimately, it is possible to insert boundaries at all possible segmentable places and obtain 100% recall. Nevertheless, if the text has many such segmentable places (e.g. 1000), precision would be drastically reduced; by contrast, if the text has only a few segmentable places (e.g. 2), precision would not suffer. The possibility of tweaking parameters has led Beeferman et al. (1997) to propose a new performance metric which takes into account 'near matches' and is expressed by a single number. A problem with their performance measure is how to define 'near matches'. Promising as this new performance measure is, recall and precision still remain as the most widely used metrics for evaluating segmentation procedures, and the best measure with which to compare different segmentation procedures.

3.3.5 Performance of feature induction model

The full model which incorporates both the short- and long-range models of word co-occurrence and the vocabulary features is referred to by Beeferman et al. (1997) as 'feature induction'. They report on the application of the

feature induction model to the segmentation of two corpora: a subset of the Wall Street Journal corpus (WSJ) comprising 325 KB of data, and 4.3 million words of the 'Topic Detection and Tracking Corpus' (TDT). This latter corpus is a collection of newswire and broadcast news drawn from other corpora. The target segment boundaries were boundaries between texts in each corpus.

There were 757 segments in the WSJ corpus. Of these, the feature induction model placed 792 boundaries with a precision of 56% and recall of 54%. A random segmentation of the same corpus achieved considerably worse results: 17% precision and 16% recall. Two segmentations were carried out in the TDT corpus. Precision ranged from 47% to 60%, with recall between 45% and 57%. The random segmentation of the TDT corpus also did far more poorly, reaching just 12% precision and recall.

3.3.6 Conclusion

The segmentation model proposed by Beeferman et al. (1997) is versatile because it incorporates components which are not specific to written language, namely long- and short-range co-occurrence probabilities, and selected segment boundary features. These components can be adapted to other semiotic systems such as images. Thus, their model can be used for example in video-on-demand applications to locate specific scenes on a video database. In more traditional document processing applications, their model can be applied to information retrieval as well as to document summarization. In information retrieval, their model can be used to locate portions of text which match a user's query by first dividing the text up in topics and presenting to the user only those portions which are relevant to the query. In document summarization, the model can be utilized to provide an initial division of a text into topics which would then be input into another application which would then

summarize each topical segment separately.

3.3.7 Implications

The model proposed by Beeferman et al. (1997) is interesting in that it combines several sources of information in segmenting a collection of texts. It is also a very sophisticated algorithm from a statistical point of view. The authors take great care in providing statistical explanation for their decisions. Another important characteristic of their study is the fact that they compare the performance of their model to random segmentations. The comparison with random segmentation performance helps put the performance of their model in perspective.

The first limitation of their approach is related to its application to finding boundaries between texts rather than within texts. This is despite the author's claims that their algorithm is ready for text internal segmentation applications. Another limitation is related to the fact that the key feature of their segmentation algorithm is word co-occurrence in arbitrary intervals. By concentrating on word co-occurrence the model overlooks the importance of how words relate to each other between clauses and sentences (Hasan, 1984; Hoey, 1991b). It must be conceded that the authors mention in passing that one of the objectives of their long-range model is to show which words are co-occurring within the space equivalent to one or two sentences. However, this does not take into account how sentences connect to other sentences which are not their immediate neighbours. In short, Beeferman et al.'s (1997) algorithm is more suitable for practical applications rather than to provide answer to questions related to how texts are organised in segments and to the role of lexis in segmentation.

3.4 Morris and Hirst

The work of Morris (1988) and Morris and Hirst (1991) has tackled segmentation by using lexical cohesive chains. In this respect, their approach differs from the techniques described so far in this chapter. Another important difference is that while the approaches described so far have been implemented by means of computer programs, Morris and Hirst (1991, Morris, 1988) offer an algorithm which has not been written into a computer program. Such analyses as they present have thus been carried out manually. The reason they have not implemented their proposal is that it depends on a machine-readable version of Roget's thesaurus which was not available at the time.

3.4.1 Lexical chains

Lexical chains are 'sequences of related words (...) spanning a topical unit of the text' (Morris and Hirst, 1991, pp.22-23). The identification of lexical chains is carried out manually by looking up chain candidates in a thesaurus. Each word candidate is assigned a category label number based on the classification of the word in the thesaurus.

The identification of chains begins by excluding closed-set words and by lemmatizing the resulting words, all of which is done by hand. The computation of thesaural similarity is done carefully so as not to exclude word pairs which are not part of the same immediate thesaural category. In addition to identical words, thesaural categorization groups together both those headwords which occur under the same thesaurus heading and those words which do not occur exactly under the same thesaurus heading but which have words which share the same thesaurus heading. Before all possible relationships are computed, a decision is taken as to whether to treat distantly related words as part of the same chain or not. For example, if the word 'cow' is found to

be related to 'sheep', 'sheep' is related to 'wool', 'wool' is related to 'scarf', 'scarf' is related to 'snow', this raises the question of whether it is fair to treat 'snow' and 'cow' as part of the same chain. The authors refer to the distance between members of different groups as transitivity, and decide on a maximum transitivity of one link; in the previous example, this limits the chain to 'cow – sheep' only.

Another important criterion is the maximum size of intervening text between related words. The authors define three sentences as the maximum distance between related words of any single chain. If distance is greater than this, a 'chain return' is computed. The authors argue that chain returns can help identify large scale chains that cut across the whole text.

The texts were segmented by the authors according to their *intentional* structure (Markels, 1983). The distribution of the lexical chains was then compared to the intentional divisions, which indicated a high degree of agreement. This was interpreted as showing that lexical chains can be an indicator of how texts are divided into coherent units.

3.4.2 Conclusions

The authors see their work as a contribution to a structural theory of texts whose main goal is the identification of 'units of text that are about the same thing' (Morris and Hirst, 1991, p.35). The identification of lexical chains can help in Natural Language Processing tasks such as word sense disambiguation. The use of a thesaurus is debated by the authors and they agree that although 90% of the intuitive chains were found in the thesaurus, important relations were not, such as street names and meronyms (e.g. 'light' and 'car').

The study concludes that more comprehensive results could be obtained if an electronic version of the thesaurus had been available, which would

have allowed for the automatic implementation of the lexical chains procedure. Furthermore, the authors warn that even though they found a match between lexical chains and perceived thematic divisions, the match was not exact, which suggests that lexical chains cannot be used on their own to locate thematic divisions. The authors make a clear distinction between 'coherence' ('being about the same thing' as perceived by a language user) and 'cohesion' ('hanging together'). Morris and Hirst argue that while the latter can be implemented objectively, the former remains largely interpretative. Therefore, a more realistic goal of research into lexical chains would be to attempt to find out possible indicators of coherent units not the coherent units themselves.

3.4.3 Implications

The relevance of the work carried out by Morris (1988) and Morris and Hirst (1991) is that they have indicated that lexical cohesion can indicate major divisions in text. One limitation of their approach is that they rely on a thesaurus, which has made it impossible for them to create a computer program to carry out the analysis for lexical chains. This is an indication that although lexical chains may be effective, they are not practical for automatic segmentation. The fact that the texts were divided into segments by the authors only may also be regarded as a limiting factor, since other readers might have provided different segments. Readers have a subjectivity about segmentation, and they often do not agree among themselves as to where to segment texts (Passonneau and Litman, 1993, see section 3.9 on page 120; Hoey, 1996).

3.4.4 Related study: Okumura and Honda

In their study, Okumura and Honda (1994) present an investigation into segmentation using lexical chains. Their methodology follows the technique presented in Morris (1988).

Lexical chains

The methodology used by Okumura and Honda (1994) is based on the algorithm introduced by Morris (1988). One modification is that unlike the original algorithm, the authors incorporate information about the sentence in which each word in the lexical chain is found. The authors believe that the sentence provides 'a preliminary filter' for determination of lexical context which can aid in chain assignment.

Another modification is that the authors included in chains only those words which are part of the same thesaural category, thus being more restrictive than Morris (1988), who also included words of similar categories. The thesaurus used for computing similarity was the Japanese thesaurus 'Bunrui-goishyo', which is similar to Roget's.

Performance

An analysis of five Japanese texts was undertaken. The texts were first segmented manually and the segmentation was compared with where the end-points of lexical chains fell. The comparison yielded an average precision rate of 25% and an average recall rate of 32%.

Conclusions and Implications

The authors conclude that the results are unsatisfactory, yet promising in that they suggest that there is a relationship between lexical cohesion and

text structure. The authors suggest a number of improvements mostly dealing with the introduction of different weighting to the lexical items in the text.

The relevance of Okumura and Honda (1994) to the present study is that they have given further support to the assumption that lexical cohesion seems to be related to segmentation. In this respect, their study corroborates Morris (1988) and Morris and Hirst (1991). Nevertheless, the problems raised during the discussion of Morris (1988) and Morris and Hirst (1991) presented above (see p.105) still apply, namely that thesauri are counter-productive aids in that, while they can help in finding similarity between words, they are by definition limited in their coverage. Further, thesauri work best when they are fine-tuned to the specific text at hand.

3.5 Hearst

Another approach to segmentation to use windows and repetition is TextTiling, a technique introduced and developed by Marti Hearst (Hearst, 1993, 1994b; Hearst and Plaunt, 1993). 'TextTiling' is used in order to divide texts into coherent parts. 'TextTiling' is used to derive broad segmentations of texts rather than to show in great detail what divisions can be made. TextTiling is perhaps the best known of all approaches to segmentation, and so it needs special attention in this chapter. One of the main characteristics of TextTiling is that the task of segmenting texts must not depend on arbitrary units but on existing textual units. Hearst (1994b) chooses the paragraph because this unit is commonly found in different text types; further she believes the paragraph to represent a coherent unit of text[1].

[1] Hoey (1996) takes a different position on this issue; according to him the perception of paragraph internal coherence is only one of the factors which influence the division of a text in paragraphs. Hoey argues that it is surface features such as lexical choices that influence paragraphing the most.

3.5.1 Overview

The main aim of 'TextTiling' is to develop a technique which can be used for information retrieval, that is, for extracting full texts from large databases. A new technique is needed because the retrieval of whole texts would allegedly be more successful if information about the whole text were taken into account, rather than information about isolated words only or groups of words in restricted contexts.

The model used by Hearst follows the work of Skorochod'ko (1972) who has looked at how much word overlap there is between sentences. It is argued that a great degree of overlap would indicate discussion of a specific topic while little overlap would not indicate a clear focus on a topic. In the work of Skorochod'ko (1972), Hearst identifies the text structure known as Piecewise Monolithic Structure as the one which serves as the basis for text segmentation. According to this text structure, discourses are made up of sequences of subtopical discussions which, although interrelated, are discrete.

3.5.2 How TextTiling works

The core algorithm of 'TextTiling' works as follows. First, the text is broken into token sequences, which are pseudosentences of 20 words each. Real sentences are not chosen because they vary in length and this variation would arguably lead to improper comparisons. Second, token sequences are grouped in blocks. A block is generally equal to the average paragraph length of each text, and this is usually 6 token sequences, that is, 6 sequences of 20 words each. Third, token sequences are compared and a similarity ratio is computed based on how many items the token sequences have in common. Finally, the similarity ratios are plotted. Text internal boundaries are located at the places where similarity scores change noticeably, which are shown by valleys

on a line plot (Hearst, 1994a, p.31). Segment boundaries are adjusted to fall between paragraph breaks; as Hearst (1994a, p.30) explains, 'when the lowermost portion of a valley is not located at a paragraph gap, the [segment boundary] judgement is moved to the nearest paragraph gap'[2]. In the end, the results of the segmentation essentially indicate which paragraphs have similar or dissimilar lexis.

In Hearst and Plaunt (1993) the authors use an adaptation of the *tf.idf* measurement to compute similarity. This measure represents the ratio between the frequency of a word in a document and its frequency in an individual text. Those terms which score highly in terms of being more frequent in one document than in the collection as a whole are taken as indicators of the contents of the text. The *tf.idf* measure is adapted by treating each block of text as if it were a complete text, that is, by computing the *tf.idf* for each block in relation to the text as a whole. Once the terms in each block have been weighted according to the *tf.idf* measure, the number of items which adjacent blocks have in common is computed. If two adjacent blocks share many items, this is interpreted as an indication that they must be part of the same discussion or subtopic. The comparison yields a similarity value which is then plotted and examined. Hearst and Plaunt (1993) examined the plots for one text and noticed that peaks and valleys tended to correspond to the topical breaks identified by human readers. They found that dips on the plot curve were indicative of topical divisions, and matched human judgement.

3.5.3 Performance of TextTiling

Several analyses of texts by 'TextTiling' appear in the literature. Hearst (1993) compares the segmentation of three texts by TextTiling to human segmentation. The texts were two popular science articles and one environ-

[2] The consequences of this decision are discussed further below on page 296.

mental impact study, of lengths ranging from 77 to 160 sentences. The results indicate an overall agreement, with TextTiling tending to match readers' inserted divisions by no more than 2 sentences off the correct boundary. In addition, TextTiling proved thorough, inserting nearly always the same number of divisions as the human readers, but never fewer.

In Hearst (1994a) 13 magazine articles were analysed by 'TextTiling'. The texts were between 1800 and 2500 words in length. The results were evaluated against the judgement of seven human readers, who provided information on where they would naturally mark the divisions of the texts. The results indicated that the technique extracted 61% of the total boundaries (recall), while 66% of the boundaries that were extracted were true (precision). However, these results would improve to 78% recall and 83% precision if boundaries that were one paragraph away from the target were counted as matches. TextTiling was also used to segment 10 documents from the Brown Corpus taken at random from the first 300 texts of the corpus. The analysis includes thesaural information, and follows a procedure originally applied by Yarowsky (1992) in sense disambiguation tasks. In the original procedure the aim was to choose between possible meanings of polysemous words by observing the context surrounding the target word and matching the surrounding words to word senses in a thesaurus (Roget's thesaurus 4th edition). Instead of the more comprehensive Roget's 4th edition, the author uses a subset of WordNet, a thesaurus in electronic form (Miller et al., 1990), consisting of 106 categories. The reason for choosing WordNet is that Roget's was not available in full to the public in electronic form. Further, a moderate size sample of thesaural categories was chosen so that the number of categories would be small enough to be manageable by the human judges in performing hand coding.

In addition to human judgement, the computer categorization by

thesaurus was compared to a random categorization. The rationale was that the random categorization should not match the categorization by the human judges. Conversely, the categorization by thesaurus should match as closely as possible the categorization offered by human judges. The results indicated that the categorization by thesaurus matched the human categorization 39% of the time when only the five top categories were included. When seven categories were allowed, the agreement rate was higher: 52%. This is better than the categorization obtained at random, which was 13% correct at best, but it is not better than the results obtained without a thesaurus (61% recall and 66% precision, see previous paragraph). It is also important to note that the agreement between judges was 54%, which shows that there is no single 'correct' categorization of the texts. Mann and Thompson (1987a, p.16) make a similar observation when they discuss the role of subjectivity in analysis (see discussion above on p.60).

3.5.4 Comparative performance of TextTiling

The performance of TextTiling was also tested against segmentation by lexical cohesive chains, as described in Morris and Hirst (1991) and Morris (1988). Morris used a thesaurus to help in the identification of similar chains, but because there were no comprehensive thesauri in electronic form, the indexing was carried out by hand. Hearst (1993) tried to replicate Morris's technique by using an electronic version of the 1911 edition of Roget, even though Morris originally made use of Roget's up-to-date 4th edition in print (Hearst claims the replication remains valid). Hearst reports some difficulties in assigning items to chains. This indicates that chain assignment is dependent on the individual text under consideration. Further, with regard to chain comparison, Hearst (1993) reports that in longer texts, chain overlap was irregular, which made it difficult to place clear boundaries at the end of co-

occurring chains. This indicates that chain segmentation may be sensitive to text size – the longer the text the less straightforward segmentation seems to become.

Morris's (1988) five original texts are also analysed by TextTiling and the results are compared to segmentation by lexical chains. The results appear to suggest an overall discrepancy between the segmentation obtained by the two methods. On the whole, TextTiling is more thorough, accounting for all sentences of all texts. The difference between the two procedures provide further evidence that there is more than one possible segmentation of a given text. Ideally, segmentation by computer should be compared to an independent criterion, such as human judgement.

3.5.5 Implications

TextTiling is relevant because it suggests that text segmentation by computer is feasible. Further, segmentation can be quite accurate, matching to a reasonable extent the divisions which readers place in texts. From an implementational point-of-view, experimentation with 'TextTiling' has suggested that the addition of thesaural information does not necessarily imply an improvement in performance. This is particularly relevant since it suggests that word form repetition can be used as input for segmentation tasks.

3.6 Reynar

The graphical method known as 'dotplot' is adapted for segmentation purposes by Reynar (1994). His paper describes an early implementation of his technique where the emphasis is put not on finding text-internal boundaries but on evaluating the viability of his approach. As such, the goals of the analyses presented in his paper are not on segmenting texts but on developing

a technique for future use.

3.6.1 Dotplot

The segmentation technique used by Reynar (1994) is based on a graphical method called 'dotplotting' introduced by Church (1993). Dotplotting works by representing each occurrence of a word as a series of four points: $(x,x)(x,y)(y,x)(y,y)$. For instance, if word A appears in sentences 10 and 20, then its position can be dotplotted as $(10,10)(10,20)(20,10)(20,20)$. When dotplotted, these points produce a dense concentration of dots around the area which corresponds to sentences 10 and 20. And when all words have been so plotted, the visual effect is that areas of the plot which share repeated words stand out showing that these areas share repeated words in common.

The actual segmentation does not need the plot, though. Rather, an algorithm computes densities of dots for each position and then selects those areas with the lowest outside density as possible boundaries.

3.6.2 Analysis

Segmentation by dotplot was tried on a collection of 600 articles from the Wall Street Journal. Instead of looking for boundaries within each article, the analysis was carried out to find the boundaries *between* articles. The reason is that internal boundaries would have to be placed by human readers, thus adding a subjective dimension to the research design. Prior to analysis, the articles were lemmatized and filtered through stop word lists which eliminated function words.

The results of two analyses are presented. The first analysis consisted of placing boundaries between sentences and checking whether those matched the boundaries between articles. The precision rate for this experiment was 17.5% for exact matches and 30% for close matches (up to three sentences

away from correct location), while recall was 53.1% for exact matches and 91.6% for close matches. For the second analysis, possible boundaries were placed between paragraphs, which reduced the possibilities of making a wrong boundary decision. As a result, precision rates increased: 54.9% for exact matches, and 80.3% for close matches.

3.6.3 Conclusions

The author concludes that his technique seems to yield good segmentation results. Its performance can be improved selectively by increasing precision while reducing recall. The artificial nature of the task of finding boundaries between texts is pointed out, and the need for adapting the technique to find text-internal boundaries is highlighted.

3.6.4 Implications

The notion that repetition can be used to carry out segmentation is reinforced by Reynar's (1994) study. Apart from the mathematical algorithm employed to locate densities and propose boundaries, the fundamentals of Reynar's (1994) approach are quite simple: eliminate function words, determine position of lexical words, plot these positions, and look for dense portions on the plot. The major implication of his study is that a successful segmentation procedure can be developed which is based on finding areas of text which share repetitions. Furthermore, it is possible to suggest that the basics of such a procedure can be simple.

3.7 Humphrey

While the various studies reviewed so far have presented original approaches to segmentation, Humphrey (1996) presents a comparison of two existing seg-

mentation algorithms – *TextTiling* (Hearst, 1993; Hearst and Plaunt, 1993) and *Dotplot* (Reynar, 1994). The Dotplot algorithm (Church, 1993) is a graphic representation of the plotting of repeated words on a chart which relies on a 'dot product equation' to compute the similarity of areas of the text. The comparison of TextTiling and Dotplot indicates that both are in essence the same algorithm, since both of them can be reduced to a common equation. To test this hypothesis, the TextTiling algorithm was first rewritten in terms of the Dotplot equation and was then applied to recovering boundaries between texts. The results indicated that the performance of the original and the rewritten TextTiling algorithms are similar.

3.7.1 Conclusions

The major conclusion is that what perhaps distinguishes the two algorithms is the level of detail they operate on. The TextTiling algorithm would be more suitable for yielding fine-grained segmentations, whereas the dotplot algorithm would be better at providing a more general picture of the internal divisions of texts. The author also concludes that the performance of both algorithms is affected mainly by what is within each window of the text.

3.7.2 Implications

For the present study, the relevance of Humphrey (1996) lies not so much in the mathematical proof of the similarity between two apparently distinct segmentation techniques, but in the fact that the two algorithms are actually similar. This is not so striking if we consider that in simple terms both algorithms rely on counting repetitions of words in portions of text and then applying mathematical formulae to the counts. Although they differ in the specific formulae which they apply, the fact still remains that the starting

point of both techniques is the identification of repeated strings of characters ('words').

3.8 Salton

In this section the approach to segmentation developed by Salton et al. (1994) will be commented upon. They refer to segmentation as 'decomposition'. Their approach is based upon finding similarity between paragraphs in the text.

3.8.1 Similarity maps

The technique presented by Salton et. al. (1994) is based upon the production of *similarity maps* for individual texts. The maps are created by means of tabulating similarity values between pairs of paragraphs. Similarity values are meant to represent whether two paragraphs belong in the same segment or not. Paragraph similarity is computed by means of the comparison of the frequency of selected terms (words or phrases) in the paragraph to their frequency in the text (or text selection) as a whole. These similarity values between paragraphs are plotted in a special chart where the paragraph numbers are laid in a circle across which lines are drawn between those pairs of paragraphs which share a certain level of similarity. The minimum level of similarity is arbitrary.

3.8.2 Segments and themes

The analysis of the maps reveals two major patterns. One, linkage between adjacently located paragraphs. These are called *text segments* and are defined as 'functionally homogeneous text units, a contiguous piece of text that is linked internally, but largely disconnected from the adjacent text' (p.3). Typ-

ical examples of text segments are introductions, and conclusions. Two, linkage between non-adjacent paragraphs. These are termed *text themes*, or 'semantically homogeneous text pieces (...) represented by mutually similar (linked) text pieces' (p.3). For instance, in a text about abortion, the themes following themes were identified: facts of abortion, and implications of abortion.

In addition to divisions into individual segments and themes, it is possible to devise a more sophisticated representation of the segments of texts by computing segment-segment relationships, as well as theme-theme relationships. Segment-segment relationships 'provide information about the overall structure of the document' (p.4). These are identified by computing the linkage between pairs of segments and excluding segment pairs which do not exceed a certain threshold. For example, in an encyclopedia article dealing with the 'American Revolution' the segments which were found were 'causes of the revolution', and 'military engagements in the revolution' (p.4). The problem of finding segment-segment relationships has also been tackled by Phillips (1985) in a different way. Phillips (1985) assumed a segmentation between chapters and then went on to demonstrate that chapters related to one another, and by doing so they formed segments. Salton et. al. (1994), on the other hand, first found the segments by comparing paragraph similarity values and then went on to look for segment-segment relationships.

Theme-theme relationships can also be computed in a similar manner and 'provide information about theme centrality and theme specialization' (p.4). An example is discussed which shows how a text on World War I can be decomposed into a 'central theme' and specialized themes each dealing with 'Naval warfare', 'Turkish activities', and 'Woodrow Wilson' (p.4).

3.9 Passonneau and Litman

The reliability of human segmentation is investigated by Passonneau and Litman (1993). Specifically, they look at how the degree of agreement between readers in segmentation tasks can be computed.

3.9.1 Reliability of human segmentation

The main problem investigated by Passonneau and Litman (1993) is the reliability of human segmentation, that is, the extent to which readers agree on where to place segment divisions. The texts which were used for segmentation were 20 transcripts of conversations. There were 7 readers, and their task consisted of inserting segments according to the speaker's intention, that is, where they felt the speakers had completed one communication task. The readers were further instructed to segment the narratives linearly, that is, hierarchical segmentation was not allowed. It is argued that naive subjects would normally find it too time consuming to divide the texts into nested segments.

They calculated segmentation reliability by computing per cent agreement, which is defined as the 'ratio of observed agreements with majority opinion to possible agreements to majority opinion' (p. 3). Majority opinion is taken to be 4 or more, given that there were 7 readers. Possible agreement equals the number of subjects times the number of boundaries. Finally, observed agreement is defined as the number of times a reader's 'boundary decision agrees with majority opinion' (p. 3). In simple terms, per cent agreement reflects the number of times the majority agreed on where to place segment boundaries *and* not place segment boundaries. The computation of per cent agreement indicated that the majority of readers agreed 89% of the time. They agreed more on where not to insert boundaries (91%) than on

where to insert boundaries (73%). The results also show that the high per cent agreement is significant statistically for agreement both on boundaries and non-boundaries. This suggests that the overall high per cent agreement did not come about as a result of the non-boundaries.

3.9.2 Segments and linguistic variables

The authors investigated the relationship between three types of linguistic variables and the segment boundaries placed by the readers. The three variables are referring expression (new noun phrases and pronouns), discourse markers, and pauses. The results indicate that the reader's segments correspond mostly to the segments suggested by referring expressions. Nevertheless, the rate of correspondence (precision) between segments and linguistic features was always low, namely 25% for referring expressions, 18% for pauses and 15% for cues.

3.9.3 Implications

The key point in Passonneau and Litman's (1993) study is that human segmentation can be a reliable task. Nevertheless, the fact still remains that the majority of readers agreed by not placing any boundaries at all. This is relevant to the present study in that it suggests that a better alternative would be not to use intuitive segments but typographical segments, since this would avoid the problem of reliability.

3.10 Conclusion

In this chapter, a review of key approaches to segmentation by computer has been provided. A few important trends can be abstracted from examining the various approaches described here. These are discussed below with the

aim of providing a framework for the decision process involved in developing the research design for the current investigation.

The first trend which can be observed is in relation to the widespread use of lexical cohesion. A large share of the studies presented in the chapter have computed some measure of lexical cohesion for segmentation purposes (e.g. Hearst and Plaunt, 1993; Kozima and Furugori, 1993; Morris and Hirst, 1991; Morris, 1988; Okumura and Honda, 1994). Lexical cohesion therefore seems to be a linguistic property of texts which renders itself amenable to computer recognition prior to segmentation. As Morris (1988, p.7) notes, 'the determination of lexical chains is a computationally feasible task'.

A second trend can be noticed by examining the ways in which lexical cohesion has been treated: among the various forms in which lexical cohesion can be formalized, a very common approach to lexical cohesion is lexical chains. A number of studies have used lexical chains for segmentation (Morris and Hirst, 1991; Morris, 1988; Okumura and Honda, 1994) and for related tasks (e.g. St-Onge, 1995). A difficulty with formalizing lexical chains is that they necessitate the resolution of anaphora so that chains formed by pronominal reference can be adequately traced through the text. The problem with anaphora resolution is that at the moment there are no approaches which can adequately resolve pronominal reference by computer. Accordingly, Hoey (1991b, p.101) observes that the restoration of pronominal reference is not a prerequisite for the computer-assisted identification of cohesion: 'If an automatic procedure is adopted this step must, at our present state of knowledge, be omitted'. Studies which depend on lexical chains have had to resort to thesauri (e.g. Morris and Hirst, 1991; Morris, 1988; Okumura and Honda, 1994) since without pronominals or thesaural information it is not possible to describe lexical chains. A problem with using thesauri is that they need to be either very extensive or fine-tuned in order

to yield good results. As a result, some studies report problems with the use of publicly-available lexical databases (Stairmand, 1996a,b; Stairmand and Black, 1996), while others have had to delay the automatization of the algorithm because of the lack of a suitable thesaurus (Morris, 1988).

A final trend that could be abstracted from the studies discussed in this chapter is that the majority of them make extensive use of mathematics and statistics. For example, Reynar (1994) makes use of a procedure which is based on complex geometry. In similar vein, Humphrey (1996) shows how two apparently distinct approaches have similar mathematical properties. The central part which mathematics plays in most approaches to computer segmentation serves as a reminder that these approaches have their roots in computational linguistics and information technology, that is, disciplines whose practitioners are fully familiar with mathematics and computer programming. In discourse analysis and applied linguistics in general, though, extensive reliance on mathematics is much less common. When it occurs, it takes a background position in the form of the utilisation of statistical tests but seldom as a centrepiece in the study.

In view of these trends, a general observation which applies to the studies reviewed in this chapter is that in so far as these studies have interfaced with linguistic theory, the interface has been merely utilitarian. In other words, these studies are concerned with the end product of the segmentation, namely the production of computer software. As a result, some decisions which are taken during the process of designing the segmentation algorithm are arbitrary, since they are not informed by previous research into discourse. According to Sparck Jones (1996, p.11) there is a great amount of 'wheel rediscovery' in Natural Language Processing, that is, computational practitioners work on some topic for some time only to find later that 'the linguists have been there before them and have already made some descript-

ive or analytic progress which could with advantage be exploited'. There seems to be some evidence to support the view that previous studies in computational segmentation are not equipped to provide a contribution to the general understanding of how discourse works. This is unfortunate since as was argued before in chapter 2 (pp. 19 ff.) segmentation bears centrally on discourse analysis, and therefore learning more about computer-assisted segmentation should enable us to understand the workings of discourse better. As Sparck Jones (1996, p.14) argues, 'there is much for linguistics to gain from looking both at how computation does things and what it finds'. In fairness, as was pointed out, studies on segmentation by computer do not purport to provide a contribution to discourse analysis and therefore it is unreasonable to criticize them for this. What is being argued here is not that they have not attempted to make a contribution, but that no real contribution has been made.

This criticism apart, an important trend observable in this review seems to be the reliance on lexical cohesion as the measure for computing segments and segment boundaries. Admittedly, there is at least one other segmentation procedure which does not make use of lexis (Hahn and Strube, 1997), but although they focus on segmenting texts, their main aim is the facilitation of anaphora resolution. There are two justifications for the reliance on lexical cohesion. The first is utilitarian: lexical cohesion is used simply because it is computable. In other words, unlike other properties such as topic and coherence which prove difficult to directly identify and quantify, lexical cohesion can be automatically identified and objectively quantified. The second justification is a mixture of theoretical and empirical knowledge, and can be expressed in a syllogism. According to our experience as readers, segments seem to be linked internally by linguistic means. According to linguistic theory, lexical cohesion seems to provide a measure of how texts

or parts of a text are linked. Therefore lexical cohesion should indicate how segments are linked internally.

The major role assigned to lexical cohesion by previous research in computational segmentation seems to be a strong indication that lexical cohesion might be adopted as the basis for segmentation in the present investigation. However, as pointed out above, previous research in computational segmentation has typically relied on lexical chains as a formalization of lexical cohesion. Other approaches to lexical cohesion are available though, for instance lexical repetition, which has already been applied computationally (e.g. Hearst and Plaunt, 1993). Before a decision can be reached, a more detailed look must be taken at the various approaches to lexical cohesion. This necessitates a new chapter.

Chapter 4

Lexical cohesion

In the previous chapter, it was argued that previous research in computational segmentation suggests that lexical cohesion can be profitably utilized for segmenting texts. In what follows, a description of the most influential approaches to lexical cohesion will be provided followed by a critical commentary on the implications of these approaches for the problem of how to segment texts by computer. Initially, it must be stressed that none of the approaches described here have been proposed with the purpose of being used in computer applications, although some of them have been used in this way, as was observed in the previous chapter. The description of each one will then be in terms of their original specifications rather than on how each one can be or has been adapted to computer segmentation.

4.1 Winburne

Thirty-five years ago Winburne (1962) published a short article in which he looked at the role of repetition in the organisation of written texts. His notion of 'sentence attachment' remains particularly important in that it relates in a sense to the work of Hoey (1991b) by showing how repetition spans whole

texts and in so doing integrates text.

4.1.1 Word distribution

The data analysed by Winburne (1962) is Lincoln's 'Gettysburg Address'. He concentrates on identifying the repetition of 'classes of meanings' across sentences. These classes of meanings are called 'sensemes', and each word is termed an 'allosense'. To illustrate his treatment of the data, table 4.1 on the next page reproduces part of the original analysis offered by Winburne; the sensemes are indicated by the letters across the top row, whereas the allosenses appear under each column heading.

The first observation made by Winburne (1962) is that sensemes are not distributed regularly: some are more frequent and regular than others. Those sensemes which are more frequent and regular (for instance those denoted by 'W' and 'X' in table 4.1 on the following page) are taken to be the principal meanings of the text, in that they provide unity and cohesion to the text (p.1097). While certain sensemes appear throughout the text (for example, those in table 4.1 under the headings 'a', 'b', 'c', and 'f'), others appear in the initial sentences only. He attributes to the former the property of helping discourse to advance (p.1097).

4.1.2 Sentence attachment

In addition to noting how words repeat themselves, Winburne (1962) also observed how word repetition has implications as to how sentences repeat each other. He looked at *sentence attachment*, or the sharing of sensemes between pairs of sentences. His counts revealed that all sentences repeat at least one element from other sentences in the discourse. For instance, sentences 2 and 3 are attached by the repetition of 'we' and 'war' (see table 4.1 on the next page). Although the only data reported in his study is the

Gettysburg Address, Winburne (1962) claims that the median number of attachments in 'most English exposition' is 2 per sentence.

4.1.3 Implications

The notion of sentence attachment is clearly relevant to the present study. Winburne (1962) suggests that sentence attachment is not a characteristic of his piece of data only but it can be found in most English discourse. Furthermore, the average number of attachments he claims is true of most English texts bears some similarity to the minimum number of links which make a bond: three links (Hoey 1991), which was chosen to reflect 'higher than average' linkage. It must be said, however, that the sentence attachments and bonds are different concepts mainly because the former reflects repetition of semantic senses while the latter is based on repetition of lexical items.

There are problems with Winburne's (1962) approach. Words are classified in meaning groups without a clear explicitation of the criteria used for grouping them. For example, 'endure' and 'last full measure' share the same meaning group. Moreover, the units which enter into meaning groups vary from single words ('nation') to multi-word items ('87 years ago') without

Sentence number	W	X	a	b	c	f
1	our	dedicated	67 years ago	brought forth conceived	nation	
2	we	dedicated	now	conceived	nation nation	war
3	we					war
4	we	dedicate			nation	
⋮ 8	we	say				did
9	us	dedicated	far			work

Table 4.1: Some sensemes and allosenses in the Gettysburgh Address (adapted from Winburne, 1962, p.1095)

justification or presentation of a rationale.

Despite these problems, Winburne's (1962) study stands out as a predecessor of many contemporary studies of cohesion. One can see in his work the origins, in principle, of the notions of tie (Halliday and Hasan, 1976; see discussion in section 4.2), lexical chains (Hasan, 1989; see section 4.3 on page 136), and bonding (Hoey, 1991b; see discussion in section 4.4 on page 149) (Hoey, personal communication). Winburne's (1962) contribution is all the more important if we consider that at the time of writing the dominant paradigm in linguistic research was syntax. Randolph Quirk's comments attached to the end of his paper criticise Winburne for not paying more attention to 'overt grammatical sequence items' and for getting 'involved in rather slippery judgements of "semantic substitutes' " (Winburne, 1962, p.1099).

4.2 Halliday and Hasan

The single most important reference in the area of cohesion is the work of Halliday and Hasan (1976). Their seminal work has introduced several key concepts which have been taken up by other studies. Important concepts introduced by them are *tie* and *texture*, which will be discussed below.

4.2.1 Definition of lexical cohesion

Before presenting the most relevant points of their work to the present study, it is necessary to define lexical cohesion and for this purpose the original definition provided by Halliday and Hasan (1976) still applies. They define lexical cohesion as 'selecting the same lexical item twice, or selecting two that are closely related' (p.12). Their definition has important implications for the way lexical cohesion can be explored by computational means and will be discussed further below (see section 4.2.4, p.131).

4.2.2 Classification of lexical cohesive ties

An important concept introduced by Halliday and Hasan (1976, p.3) is that of *tie*, or 'a single instance of cohesion'. They describe in detail the various kinds of lexical cohesive ties in English. The two major types of lexical cohesion according to Halliday and Hasan (1976) are reiteration and collocation. Reiteration occurs when there is an occurrence of an identical or related word. The second occurrence can be the same word, a synonym (or near-synonym), a superordinate, or a general word. For instance, given the sentences 'there's a boy climbing that tree', and 'the boy's going to fall if he doesn't take care', a tie exists between the two occurrences of 'boy' which would be classed as reiteration by repetition. If the second sentence were 'The lad's going to fall ...', the tie would have been a result of the reiteration of the synonym 'lad'; if the second sentence had been 'The child's going to fall', the tie would have occurred because of the superordinate 'child'. And if the second sentence were 'the idiot's going to fall ...', the resulting tie would have occurred because of the general word 'idiot'. The other type of cohesion is 'collocation', which is defined as 'the association' of lexical items that regularly co-occur' (p.284). These include items which are members from the same ordered series, for instance 'Tuesday' and 'Thursday'; pairs from unordered lexical sets, like 'basement' and 'roof'; items which are 'parts' of a 'whole', such as 'car' and 'brake'; co-hyponyms, for example 'chair' and 'table'; synonyms and near-synonyms such as 'climb' and 'descent'; complementaries, like 'boys' and 'girls'; antonyms, such as 'like' and 'hate'; and converses, such as 'order—obey'.

4.2.3 Texture and text

Texture is defined by Halliday and Hasan (1976, p.2) as the property of 'being a text'; it is what distinguishes a text from a non-text. A text is defined by Halliday and Hasan (1976, p.293) as a 'semantic unit', as opposed to a grammatical unit. The distinction is a major one in that it leads to the question of how this semantic unit hangs together. Unlike texts, grammatical units such as the clause achieve unity by means of grammatical structure. Since text is non-structural, its unity cannot arise out of grammatical structure, but from cohesion.

An important distinction is made by Halliday and Hasan (1976) with regard to the relationship between texts and sentences: texts do not consist of sentences, rather they are encoded in or realized by sentences. Texts and sentences are different linguistic units – the text is semantic, the sentence is grammatical. As was mentioned above, it is grammatical structure which holds sentences together thus making cohesion within the sentence irrelevant (Halliday and Hasan, 1976, p.9). Nevertheless, this does not in turn imply that cohesion is a relation 'above the sentence'. Cohesive links are perceived across sentences because this is the only source of texture across sentences, given that sentences are structurally independent of each other.

4.2.4 Implications

The definition of lexical cohesion provided by Halliday and Hasan (1976, p.12) applies to the context of the present study; they define lexical cohesion as 'selecting the same lexical item twice, or selecting two that are closely related'. Their definition equates cohesion with repetition and therefore it implies that it is possible to study lexical cohesion by studying repetition. An implication of their assertion is that it opens up the way for the study of

lexical cohesion by computers since computers can be programmed to reliably identify repetition but they cannot be easily made to identify other types of lexical cohesion.

The way in which Halliday and Hasan (1976) approach texture has implications for the possible application of cohesion to segmentation. Although they do not address segmentation as the task of finding the internal boundaries of individual texts, they do make mention of assessing variation in levels of texture as a means whereby one could identify boundaries between texts (p.295). Halliday and Hasan (1976) also observe that cohesion may indicate 'transitions' in the development of texts. They note that 'a transition between different stages in a complex transaction, or between narration and description in a passage of prose fiction might be regarded as 'discontinuities' thus 'signalling the beginning of a new text' (p. 295). Halliday and Hasan (1976) relate this rhythm setting role of cohesion to the paragraph: 'the paragraph is a device introduced into the written language to suggest that kind of periodicity' (p. 296). Although the relationship between textuality and paragraphing is debatable (Hoey, 1985, 1996), the fact that Halliday and Hasan (1976) relate cohesion to internal divisions of written texts is of importance to the present study because it suggests that there may be a mapping of cohesion onto major existing divisions of texts.

4.2.5 Systemic Functional Grammar

In this section, studies contributing to our understanding of lexical cohesion from a systemic functional perspective (Halliday, 1985) will be briefly discussed. Two studies are discussed below, both of which make use of the proposal by Halliday and Hasan (1976) discussed above.

Halliday

Halliday (1985) defines lexical cohesion as the pattern which results from the selection of items 'that are related in some way to those that have gone before' (p.310). Lexical cohesion is regarded as one of the types of cohesive features of the textual component in the functional grammar; the others are reference, conjunction, and ellipsis and substitution. The role of lexical cohesion (and of the other components as well) is to contribute to texture.

Halliday analyses lexical cohesion by identifying referential chains (Martin, 1992, p.140ff), which are sequences of lexically cohesive items joined through lexical relations (e.g. repetition, synonymy, and reference). He argues that referential chains can also be called 'participant chains' if they contain elements participating in the same transitivity processes. For example, the chain *drown + mermaid* $\rightarrow$ *drown + fish* $\rightarrow$ *fish + eat* operates in the conversation shown in figure 4.1. The participants 'mermaid' and 'fish' share in the process 'drown' which helps create texture and a 'dynamic flow of discourse' (Halliday, 1994, p.337). Halliday stresses that it is not the presence of chains in isolation, but their interaction which contributes to coherence.

Nigel: **Drown a mermaid!**
(...)
Father: No, you can't **drown a mermaid**, a mermaid lives in the water. You can't **drown a fish** either, can you?
(...)
Nigel: I liked that fish we saw at the Steinhart, the one that its tail wasn't like a fish. **It was eating** a lettuce.

Figure 4.1: Referential chains in context (Halliday, 1994, p.99)

Eggins

Eggins (1994) analyses lexical cohesion by means of lexical strings. She specifies certain conventions for representing lexical strings. For example, figure 4.2 on the following page shows a sample text analysed for lexical strings according to Eggins's (1994) conventions. Words related taxonomically (i.e. meronymy, hyponymy, class/sub-class, contrast, synonymy, and repetition) are placed vertically, while those in expectancy relations (co-occurrence or process-participant) are depicted diagonally. The particular relations (anaphora, cataphora, etc) in which words enter are depicted in boxes. The lines which connect the items across the diagram also follow a convention as regards the shape of the arrows – they are upward pointing for anaphoric references and downward pointing for cataphoric (exophoric and homophoric references are marked by curved arrows and an overlapping label). Typically, only the main lexical strings, that is, those containing more than three or four items, are shown. She argues that lexical strings can be used as devices for identifying the topic or sub-topic(s) of a text. For instance, she argues that the excerpt in figure 4.2 is concerned with blood and body parts, which is reflected in its lexical strings.

Eggins (1994) also argues that different genres should exhibit different relations in their respective sets of strings. She speculates that technical texts should be characterized by strings showing the 'deep' level of a field, whereas in everyday texts the lexical strings would include items indicative of the 'shallow' end of the field. In this manner, lexical choices 'point upwards to the field dimension of context' (p.105).

Simon: How how you did – have you given blood before?

(...)

Diana: No I do it because I had a daughter who when she was 2 days old needed blood transfusions cause she was getting sort of premature jaundice and things. This was in Geneva. And they rang me up on the Sat - this was Saturday night and said 'You've got to come in and have your blood tests against the donors'.

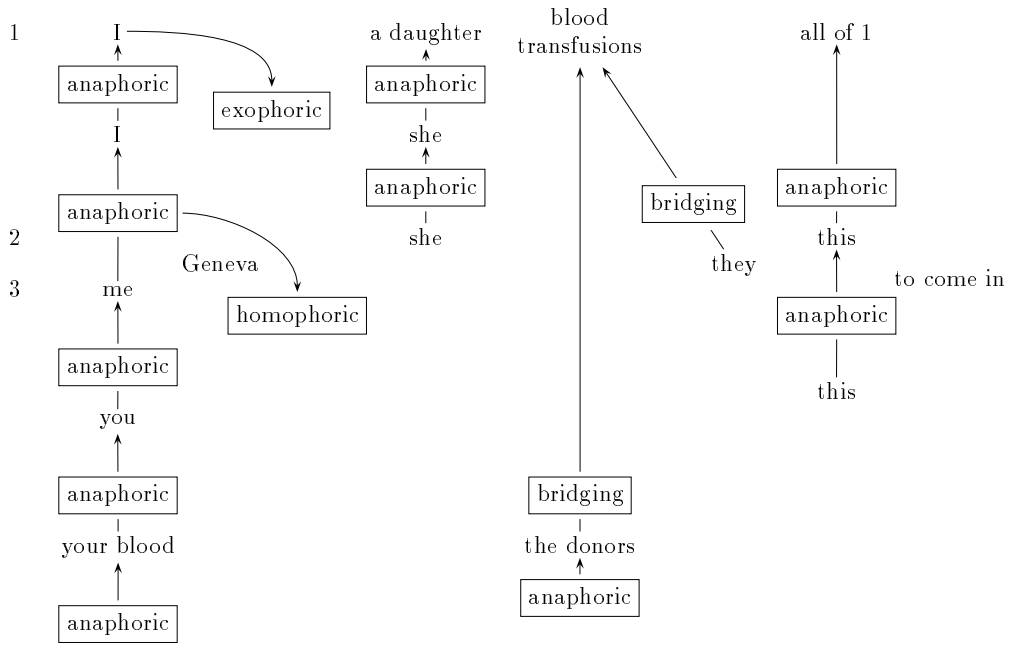


Figure 4.2: Text extract and lexical strings (adapted from Eggins, 1994, p.93)

4.3 Hasan

A major development from the original proposal for the analysis of cohesion in texts proposed by Halliday and Hasan (1976) is the work of Hasan (1989). Her work is instrumental in placing lexical cohesion in the centre of textual research, mainly those aspects which concern the search for the linguistic correlates of coherence. In what follows key notions developed by Hasan (1989) are reviewed.

4.3.1 Semantic relationships

In all, Hasan (1989) distinguishes three types of semantic relationships: co-referentiality, co-classification, and co-extension. In a co-referential semantic relationship, both terms of the tie share the same referents, for instance between 'I had *a little nut tree*' and 'Nothing would *it* bear' (p.73). In a co-classificational relationship, 'the things, processes, or circumstances to which A and B refer belong to an identical class, but each end of the cohesive tie refers to a different member of this class' (p.74), so in the example 'I *play the cello*. My husband *does* too.', each one of the players does his/her own playing, and each playing constitutes a different 'situational event' (p.74). Finally, in a co-extensional relationship, each member of the tie refer to something in 'the general field of meaning', for instance 'golden' and 'silver' in 'A *silver* nutmeg. And a *golden* pear' (p.73).

These three fundamental semantic tie-establishing relationships differ in relation to the typical ways in which they are expressed in texts. Hasan (1989) notes that co-referential and co-classification relations are typically established through 'implicit encoding devices', that is, pronominals, definite articles, demonstratives (in the case of co-reference), substitution, and ellipsis (in the case of co-classification), while co-extensional relations are established

among content-bearing items. She uses the term 'implicit encoding devices' for the former category because 'their interpretation has to be found by reference to some other source' (p.75). She adds that it is exactly the role of these devices in relating a referent to its reference that enables them to function as cohesive devices. Nevertheless, she observes that their role as cohesive devices does not arise simply because they must be interpreted by relating to a previous item within the text, since in some texts the reference is not made explicit. For example, she analyses briefly a short poetic text in which the following lines appear: 'Upended, it crouches on broken limbs (...) It gapes enmity from its hollowed core' (p. 78-79). In the whole passage there is no mention of 'tree' yet the reader relates both occurrences of 'it' to 'tree'. She argues that the interpretation of implicit cohesive devices without their referents is made possible because of the other type of semantic relation, co-extension. More precisely, it is the co-extensional relationships established in the text which help the reader to interpret devices of reference without referents, because items related through co-extension would create a field of meaning which would serve as a guide to the reader. So much so that Hasan (1989) concludes that 'where such [co-extensional] ties do not exist, the relation of co-reference and co-classification are at least problematic if not impossible to establish' (p. 79). In the case of the example cited above, 'it' would have been interpreted as referring to 'tree' because of the co-extensional relations reminiscent of 'tree' which linked 'hollowed core', 'woodflesh', 'splinter', and 'torn root' (p. 79) elsewhere in the lyric.

Co-extensional relations could be termed 'explicit' devices by contrast with the implicit devices which realize co-reference and co-classification. Unlike co-reference and co-classification, the interpretation of which depends on the retrieval of a previous item in the text, co-extension simply requires that speakers 'know the language'(p. 50).

Although the notion of 'general field of meaning' is helpful, it must be delimited so that it becomes possible to explain how co-extensional relations are established. Otherwise, it would be possible to create a sequence of items linked by co-extension such as 'flower, petal, stem, stalk, twig, branch, trunk, tree, wood, log, faggot, tinder, fire, flame' (p. 80) in which 'flower' and 'flame' would feature as being related by belonging to the same 'general field of meaning'. Instead, she argues that different pairs of items in the list are associated by different 'sense relations'.

4.3.2 Sense relations

The sense relations which Hasan (1989) identifies are five: synonymy, antonymy, hyponymy, meronymy, and repetition. When two items are synonymous, their experiential meaning is identical (for instance 'buy' and 'purchase') whereas if they are antonymous, they have opposite experiential meanings (for example 'golden' and 'silver'). When two items are related by hyponymy, one of them represents a general class, while the other represents a sub-class (for instance, 'cat' and 'dog' are co-hyponyms of the superordinate 'animal').

In addition to these three general semantic relations, Hasan includes 'meronymy' and 'repetition'. The former links items which stand in a part-whole relation, as for instance 'limb' and 'root' which are co-meronyms of the superordinate 'tree'. In the case of repetition, 'the same lexical unit creates a relation simply because a largely similar experiential meaning is encoded in each repeated occurrence of the lexical unit' (p. 81). Obvious as it might sound, repetition is arguably the most direct way in which a tie can be created. It is also in many instances the most frequent way, which implies that repetition is a powerful texture forming device. Below in this chapter, the importance of repetition is examined more closely and it is argued that there is empirical evidence to support the view that repetition is a key element

in creating texture. This is turn has important implications for the way the main study presented in this thesis is implemented.

4.3.3 Other relations

In addition to the classification of sense relations in five categories as described above, Hasan (1989) sub-classifies semantic relations in terms of whether they are general or instantial. General relations are 'facts' of a given language, for example, the synonymy relation between 'lady' and 'woman'(p. 81). By contrast, instantial relations are those which are specific to a particular text or message, for instance between 'pleasures' and 'yesterdays' in 'all my pleasures are yesterdays' (p. 81).

Sense relations can be further subclassified between componential and organic relations. The former are those which link components of a message, and these include all those which are linked by the semantic relations discussed so far, namely co-reference, co-classification, co-extension. The latter are formed by ties in which their members are whole messages, such as adjacency pairs (e.g. question-answer), and between clauses ('I'm going to bed because I'm very tired', p. 81).

4.3.4 Cohesive chains

A chain is defined as 'a set of items' which is related to the others by the semantic relation of co-reference, co-classification, and/or co-extension. Based on the kinds of semantic relations which create the chains, it is possible to distinguish between identity chains and similarity chains. The former are those whose members are related by co-reference, so that 'every member of a chain refers to the same thing [or] event' (p. 84). The latter are formed by items which are related by co-classification or co-extension. These chains typically contain elements which 'refer to non-identical members of the same

class of things, events, etc' (p. 84), and therefore the items in similarity chains 'belong to the same general field of meaning' (p. 85). Hasan (1989) further observes that similarity chains can be predicted if we know the field of discourse relevant to a given interaction (p. 84). In other words, if certain semantic groupings are expected given the field of discourse of a text, it is likely that such semantic groupings will materialize in the form of similarity chains.

Hasan's hypothesis is also important in another sense in that it conflicts with an earlier position expressed in Halliday and Hasan (1976), where they argue that chains do not normally reflect the subject matter of a passage. Admittedly, Halliday and Hasan were referring to *subject matter* which strictly speaking is not synonymous with field of discourse. Nevertheless, field and subject matter are related, with subject matter being a second order kind of field (Martin, 1992).

4.3.5 Coherence and chain interaction

The notion of cohesive chains is central to Hasan's analysis of the coherence of texts. Hasan (1989) presents two texts with differing degrees of coherence. Of these texts she asks 'if the two [texts] vary in the degree of coherence, what, if any, patterns of language correlate with this variation?' (p. 88). Her central assumption is that 'cohesion is the foundation on which the edifice of coherence is built' (p. 94), and her initial hypothesis is that the less coherent text is so because it has references which point out of the texts (exophoric), but she refutes this by showing that exophora prevents neither the formation of cohesive ties nor the interpretation of co-reference and co-classification. Her other hypothesis is that the less coherent text is more ambiguous, that is, it contains 'grammatical cohesive devices which could be interpreted in more than one way given the frame of the particular text' (p. 89). She argues,

however, that a text is by default approached as if it were coherent, and therefore readers will tolerate a certain degree of ambiguity. She concludes that ambiguity and coherence are independent (p. 89). Given that neither ambiguity nor exophora, that is, factors that prevent chain formation, can explain the difference in coherence between the two texts, Hasan hypothesizes that it is possible that the two texts vary in relation to the number of tokens each has in chains, and she finds that although the more coherent text has a greater number of tokens subsumed in cohesive chains, so does another completely incoherent text (p. 83). The incoherent text is in reality a series of unrelated sentences with a high degree of repeated items ('a cat is sitting on a fence. A fence is often made of wood...'). She concluded that chain formation is not a good indicator of why the texts differ in coherence. The reason is that when analysing chain formation one is not taking the whole message, but simply separate words, into account. What is needed is a method which will allow for the incorporation of the information about how chains are related to each other as messages. This she terms *chain interaction.*

The justification for the need for approaching coherence via chain interaction is given on the grounds that 'it is only message as message that has textual validity; and it is only at the rank of clause or above that a lexicogrammatical unit is contextually viable: it is only at this rank – or above – that a linguistic unit can encode a complete message' (Halliday and Hasan, 1976, p.91). The way she operationalizes chain interaction is by identifying at least two elements of a chain which 'stand in the same relation to two members of another chain' (p. 91), such relations being those that exist between the constituents of a clause or group (e.g. doer–doing; sayer–saying; doing–done-to, etc).

The diagram in figure 4.3 on page 143 displays the chains that interact in the example text in the same figure. The chains are identified by letters

in brackets (*(a)*, *(b)*, etc), and the relations holding between the chains are identified by roman numbers as described by the key in the figure. Thus, the members of chain *(a)* are in an 'actor action' relation with members of chain *(e)* (e.g. 'girl went'), and therefore the chains interact. Chain *(h)* interacts with chain *(b)*, by means of an 'action acted-upon relation' (e.g. 'took teddybear'). Chains *(c)* and *(a)* interact by means of an 'action and/or actor location' relation ('girl got home'). The relation which causes chains *(l)* and *(n)* to interact is 'saying text' (e.g. 'said words'). And finally, chains *(f)* and *(b)* are in an 'attribute attribuand' relation (e.g. 'lovely teddybear') and therefore they interact.

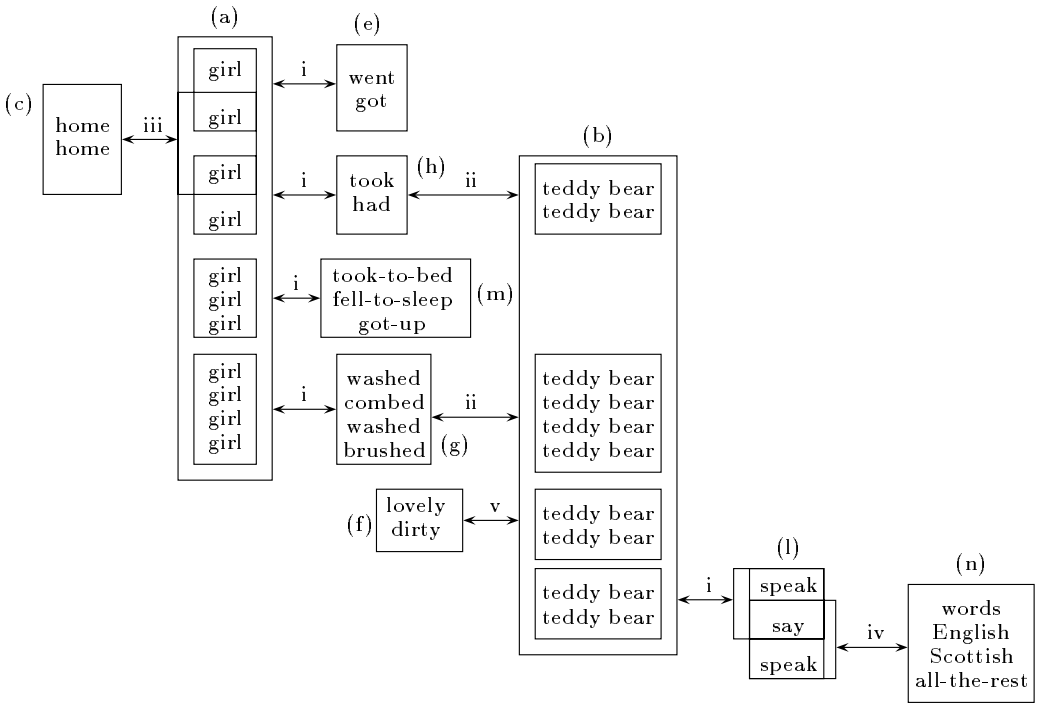
4.3.6 Cohesive harmony

In order to analyse chain interaction, one needs to distinguish first of all between 'relevant' and 'peripheral' tokens, the former being tokens that enter into any kind of chain, the latter being those that do not. Relevant tokens can be further broken down into 'central' (those chain items which actually interact) and 'non-central' (those which do not interact).

The computation of various statistics involving relevant, peripheral, central, and non-central tokens allows one to estimate the 'cohesive harmony' of a text, that is, the 'linguistic correlates of coherence based on chain interaction' (Halliday and Hasan, 1976, p.93). In other words, variation in cohesive harmony is expected to correlate with variation in coherence ('variation in coherence is the function of variation in the cohesive harmony of a text', p. 94). There are two particular ratios that Hasan (1989) identifies as possible indicators of levels of coherence. One is the proportion of peripheral tokens to relevant tokens – the lower the proportion, the more coherent the text should be. This predicts that 'the semantic grouping in the text should be such as

Text

1. once upon a time there was a little girl
2. and she went out for a walk
3. and she saw a lovely little teddybear
4. and she took it home
5. and when she got home she washed it
6. and when she took it to bed with her she cuddled it
7. and she fell straight to sleep
8. and when she got up and combed it with a little wirebrush the teddybear opened his eyes
9. and she started to speak to her
10. and she had the teddybear for many many weeks and years
11. and so when the teddybear got dirty she used it to wash it
12. and every time she brushed it it used to say some new words from a different country
13. and that's how she used to know how to speak English, Scottish, and all the rest.



Key

Letters in brackets:	individual chains
Roman numbers:	relations between chains

Key to roman numbers

i	'actor action'
ii	'action acted-upon'
iii	'action and/or actor location'
iv	'saying text'
v	'attribute attribuand'

Figure 4.3: Chain interaction (Hasan, 1989, p.72, and p.92)

to establish unequivocally certain definite referential domains' (p. 94), the semantic grouping being represented by the relevant tokens, that is, those which enter in chains. The other statistic is the proportion of central tokens to non-central ones – the higher the proportion, the more coherent the text should be. The rationale behind this ratio is that the 'establishment of the definite referential domain is not enough', rather it is necessary that 'speakers stay with the same and similar things long enough to show how similar the states of affairs are in which the same and similar things are implicated' (p. 94).

These two ratios are computed for both the more and the less coherent texts and the results support the hypotheses. The more coherent text has 90.5% relevant tokens and 65% central tokens. The less coherent text, on the other hand, has only a 76% total of relevant tokens, and 36% of central tokens. These figures seem to indicate that indeed cohesive harmony seems to be a linguistic correlate of coherence. Parsons (1990) however showed on the basis of further experimentation with a large set of texts that cohesive harmony is not a reliable measure of coherence.

Focal chains One less quantifiable measure which according to Hasan (1989) could be related to variation in coherence is the presence of 'focal chains' (p. 94). Focal chains are described as long chains which interact with other chains. In the text in figure 4.3, all chains are related to each other via two focal chains, namely chains *(a)* ('girl') and *(b)* ('teddy bear'). These are the chains which in a sense hold the text together by allowing the other separate chains to hook onto one another. Hasan (1989) then concludes that in the case of coherent texts, 'the outcome is that a complete break in chain interaction does not take place – transition from one topic to the next is a merging rather than a clear boundary' (p. 94).

4.3.7 Implications

The work of Hasan (1989) is important for several reasons. First, it systematizes the study of lexical chains introduced by Halliday and Hasan (1976). Second, it makes bold claims about the relationship between cohesion and coherence, some of which needed testing on a large body of data, a task undertaken later on by Parsons (1990, 1996). Finally, her work has implications for a study of segmentation based on lexical facts, even though she does not address segmentation as such. On the whole, her methodology is not directly implementable on the computer, which makes her approach inappropriate for large-scale investigations. The basis of her methodology lies on the notion of chain interaction, which in turn rests upon the analysis of transitivity. Such analysis cannot be implemented successfully by computer because the interpretation of transitive relations depends on linguistic knowledge which is difficult to model on the computer. But the main reason why her methodology is not suitable for the investigation of the relationship between segmentation and lexical cohesion is that in her model grammatical cohesion is a major component. As she puts it:

> to be effective, grammatical cohesion requires the support of lexical cohesion [and] to be effective, lexical cohesion, in turn, requires the support of grammatical cohesion. The reciprocity of these two kinds of cohesion is essential. (Hasan, 1989, p.82)

The notion of 'focal chains' is relevant for the present study of segmentation because it might suggest that coherent texts do not have topical breaks. If this is true, then the major premise which underlies this study would be false, namely that texts contain internal divisions. However, by considering how focal chains interact, it may be possible to relate the notion of focal chains to segmentation. For example, in the sample text in figure 4.3 on page 143, the type of interaction of the two participants in the focal chains

for 'girl' and 'bear' changes at clauses 2, 9, and 13. Clause 1 may be treated as an 'introduction', including the chain for 'girl' only. Clauses 2 to 12 include both chains ('girl' and 'bear'), but there is a major difference in the relationship between the chains within this span: from clause 2 to clause 8, the 'girl' acts on the bear, whereas between clauses 9 and 12 either the 'bear' or the 'girl' acts. Clause 13 presents a 'moral', and like in the introduction, includes the chain for 'girl' only. The major change in the interaction signalled by the interaction of the focal chains is with respect to the 'bear', which switches from 'done to' to 'doer'. Seen in this way, the existence of focal chains would be compatible with a view that focal chains maintained the overall continuity of the text, while chain interactions broke at boundaries of segments (G. Thompson, personal communication, 1997).

4.3.8 Related Study: Parsons

In this section a study applying the methodology introduced by Hasan (1989) is reviewed. Parsons (1990) analyses student compositions for lexical chains; his aim is to investigate to what extent the presence of lexical chains correlates with perceived coherence of the compositions. In particular, Parsons (1990) explores the relation between chain length and coherence by looking at *significant chains*. Although the study is not directly related to the problem of segmenting texts, its findings are of relevance to the general issue of the relationship between lexical cohesion and coherence which is involved in an investigation of segmentation.

The study Parsons (1990) investigates the role of cohesion in student writing by applying the analytical principles developed by Hasan. The texts were 16 compositions written by non-native university students. The analysis involved performing a lexical rendering of each text, which consists of omitting

grammatical words (articles, conjunctions, etc) and restoring ellipsis and pronoun referents. The lexical chains present in the texts were then identified and four ratios were computed: RT/PT (Relevant tokens to peripheral tokens), Ct/nCT (Central tokens to non-central tokens), %CT (percentage of central tokens of the total lexical tokens), and CT/PT (Central tokens to peripheral tokens).

The texts were rated by 12 informants as to their 'communicative effectiveness', which resulted in a classification of the texts into 4 groups. The group which received the best overall rating was composed of native speaker writers only; the second best group was nearly all non-native (except for one); and the other two groups had only compositions by non-native writers in them. The texts were rated again, this time by coherence. The informants' judgements yielded a classification into four distinct groups. The group containing the texts perceived to be more coherent had only compositions by native-speaker writers. The group considered to be the second most coherent was split into non-native and native writers (two apiece). Both sets of informant judgements of the texts were then compared to the analysis for lexical chains as revealed by the percentage of central tokens (%CT). The results indicated a lack of correlation between %CT and 'communicative effectiveness', but a positive correlation was found between %CT and coherence (Pearson-Product r=.427, page 163). The correlation, though weak, is significant at $p<.05$, and it is thus concluded that perception of coherence was associated with percentage of central tokens. This is said to corroborate Hasan's claims that coherence correlates with %CT, but a later comparison with the RT/PT ratio reveals a lack of association, which contradicts Hasan's prediction.

Significant chains Since a greater number of central tokens suggests that longer chains might be contributing to the perception of coherence, the au-

thor then investigates the role of chain length. The term *significant chain* is employed to deal with those chains which present a number of central tokens higher than the average across the texts. It was found that the average chain length was 3.13 tokens, therefore it was accepted that a significant chain would be one which comprised more than 4 tokens. A ranking of the texts in terms of the total of significant chains in them and their perceived coherence suggested a possible association between the two measures (p.173). In order to conduct a more objective assessment of the role of significant chains, the author provides a count of the percentage of significant tokens in each text, or the number of tokens present in significant chains. A correlation coefficient of r=.538 (p<.025) (p.182) was found between percentage significant tokens and perceived coherence, which is higher than that obtained for central tokens.

It is hypothesized that the frequent use of longer chains can account for most of the perceived coherence in the texts. The correlations for percentage of 5 tokens was also significant (r=.586), but for 6 tokens the correlation was low (.334) and not significant at p<.05. The author revises the original concept of cohesive harmony by stating that it is not central tokens but tokens participating in long chains which contribute to coherence: 'It seems that chain interaction alone does not necessarily result in the most coherent texts, but that interaction which organises the tokens into long chains is more likely to lead to coherent texts in which there are more occurrences when one is saying "similar things about similar phenomena"' (p.204). The author squares the value of the significant correlations to estimate the amount of variation in coherence due to cohesion. For central tokens, the squared correlation is .182, and for percentage of 5 tokens .343, which suggests that cohesion accounts for about 34% of the coherence in the texts (p.221). Textual features other than lexical cohesion (e.g. grammatical cohesion) are also

responsible for perception of coherence. The implications are that teaching students to write significant chains might help them improve their writing.

Implications The investigation presented by Parsons (1990) does not approach segmentation as such but its major finding that lexical cohesion does not account for the total degree of coherence perceived in the texts suggests that the presence or absence of lexical cohesive ties must not be interpreted as lack of quality. This must be borne in mind during the analysis for segments because it is possible that some segments, while marked as such by authors (and very probably perceived as such by readers), may not exhibit significant numbers of lexical cohesive links and therefore they may not be identified at all by a method which relies on the existence of lexical cohesion. Significantly, by examining a larger body of data Parsons (1990) obtained results which differ from those presented originally in Hasan (1989). This serves as a reminder that an examination of more quantities of data may present findings which can contradict theoretical claims without disqualifying the original model.

4.4 Hoey

The work of Hoey (1991b, 1988) on patterns of lexical cohesion in text forms the basis of the study presented in this thesis. His approach is based on the notion that lexical cohesion forms clusters among sentences. Methodologically, his work is innovative in that it presents a new method of analysis for dealing with lexical cohesion and investigating lexical cohesion between sentences. Theoretically, his in-depth analysis of the way in which lexical cohesion operates in text stresses the importance of lexical cohesion among the other types of cohesion. His method, it will be argued later, can be adapted to the investigation of segmentation.

4.4.1 Relations to previous work

Hoey's proposal is aimed at harmonizing three insights from previous lexical cohesion studies. First, it is devoted to showing how cohesion clusters; in other words, it builds upon the earlier work of Hasan (1989) and concentrates on how chains interrelate. Second, it draws on the work of Winter (1977), and in particular on the assumption that the fundamental function of lexical cohesion is to repeat. Finally, by following Phillips, it is aimed at identifying long-distance lexical cohesive relationships. In short, the central characteristics of Hoey's approach to lexical cohesion are: it is integrative, repetition-based, and incorporates long-distance ties. The remainder of this section will explain in greater detail how Hoey's proposal works and how it can be applied to segmentation.

4.4.2 Importance of lexical cohesion

Hoey (1991b) stresses the importance of lexical cohesion by noting that even in Halliday and Hasan's example texts, lexical cohesion is the dominant type of cohesion (over 40% of the ties are lexical). We further note below that in the same texts, lexical repetition is the dominant type of lexical cohesion, which has implication for determining how cohesion will be computed automatically in this study. In addition, Hoey considers lexical cohesion to be the only type of cohesion which can establish multiple connections. For him, clusters of lexically cohesive items are arranged in a net-like rather than in a string-like fashion. For instance, in the text displayed in appendix 4 on page 457, 'reader' in sentence 16 links back with 'reader' in sentences 1, 7, 10, 12, and 14, and each of the occurrences of 'reader' in these sentences links to each other. As a result, the links among these occurrences can be represented in a net, as in the diagram on the left in figure 4.4 on page 151. By contrast,

if these links were considered as forming a string, their representation would be like the diagram on the right in figure 4.4. By admitting of multiple links between lexical items, the number of ties proliferates, thus increasing the share of lexical ties. Finally, lexical cohesion lends itself to identification by automatic means (Hoey, 1991b, p.74), which is also crucial for the present study, since this study is aimed at investigating lexical cohesion in a large number of texts.

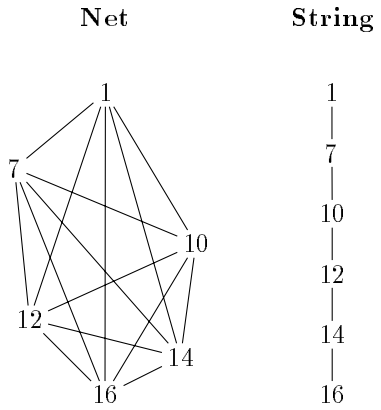


Figure 4.4: Net and string of lexical links (adapted from Hoey, 1991b, p.81)

4.4.3 Lexical cohesion and text organisation

Lexical cohesion, according to Hoey (1991b), relates to text organisation. Other studies have looked at lexical cohesion, but most studies have been dedicated to looking at how lexical cohesion can be classified. Hoey, on the other hand, agrees with Winter (1977) when he states that the common function of the various types of cohesive devices is to repeat (Hoey, 1991b, pp. 16-17). Hoey argues that repetition organises text by creating networks which stretch across the whole text linking separate messages.

The relationship between lexical cohesion and text forms the essence of Hoey's approach. It is therefore necessary to explain in detail how he sets out to investigate this relationship. Hoey believes that text, being a relatively new object in linguistic inquiry, has been approached by means of two metaphors. First, as a sentence; here he includes those linguists who have attempted to formulate a Text-grammar along the lines of sentence grammar, for instance van Dijk (1972). Second, as dialogue: for instance, the work of Winter treats clause relations 'in terms of the questions a reader may ask of a text at any moment' (p.30). Hoey argues that both metaphors have failed to provide means for linguists to tackle the complexities of text because they are based on structural principles. Both are based on the belief that text is structural. A structural description has two characteristics which are not valid for describing text, namely the power to predict certain sequences, and the ability to claim that certain sequences are impossible (p. 193-201). These characteristics are not applicable in the case of the lexical patterning which non-narratives exhibit. As Hoey (1991) puts it, 'it would be a daring person indeed who risked declaring which combinations of elements could not occur together in a text and which had to' (p.204).

A new metaphor is needed if we want to take account of the complex patterns that lexis creates in text. Hoey proposes that a better metaphor would be a 'collection of texts', in that 'such a comparison would build on the premise that texts are made up of interrelated but separate packages of information – sentences – just as a collection of texts might be' (p. 31). One such text that fits this description is the academic paper, since in the bibliography section it makes explicit links to other papers. Taken together, academic papers provide a more promising metaphor for analysing interconnections among sentences, mainly because each paper that cites another is in a sense repeating the other paper. This metaphor enables us to build

in the key notion of repetition, which is central in Hoey's approach (Hoey, 1991b, p.35). Further, it makes it possible to distinguish papers which are more central to the collection to those which are more peripheral, the former being those which include more citations to other papers. Such central papers do in a way incorporate a greater share of the collection. If the terms of the metaphor are translated to their textual equivalents, it then becomes possible to replace academic papers with sentences, and collection of papers with text. This further enables us to distinguish between central and marginal sentences. Central sentences, like central papers, 'make a number of connections with other sentences [and] are germane to the development of the theme(s) of a text' (p. 43), whereas marginal sentences 'contribute less to the development of its themes [and] show fewer signs of connection with the rest of the text' (p. 43).

4.4.4 Lexical cohesion and coherence

Like Hasan (1989), Winter (1979), and Phillips (1985), Hoey (1991b) sees coherence as related to cohesion. This is different from other studies (e.g. de Beaugrande and Dressler, 1981; Widdowson, 1978) which see the two concepts as separate. Hoey observes that coherence is a subjective judgement on text, while cohesion is a property of texts which can be assessed automatically (p. 25). The relationship between cohesion and coherence is not straightforward, though; it is rather a question of 'how the presence of a cohesive tie predisposes a reader to find a text coherent' (p. 12). If cohesion is measured by the amount of cohesive ties, as it was by Parsons (1990, 1996; see section 4.3.8 on page 146), then according to Hoey this will not be a measure of how messages are connected. Hoey argues that cohesive ties are not by themselves criterial of coherence (p. 12), since 'in addition to perceiving ties between words in the sentences we encounter, we also see relationships

between the sentences as whole units' (p. 12). Coherence therefore concerns interpreting messages and judging whether such messages are related or not.

If this assumption is correct, it would be possible to predict that the majority of messages connected by a significant number of cohesive ties will be perceived as coherent. The result of Hoey's analyses indicate that less than 50% of random pairs were coherent (p. 192), while pairs of sentences connected by a significant number of cohesive lexical items are normally coherent (p. 133). In Hoey's words, 'the co-occurrence of the requisite number of repetitions is sufficient to compel a reading of the pairs as intelligible' (p. 126). This prediction was further investigated by Wessels (1993b, 1993a; see section 4.4.10) who, using Hoey's system, found that the presence of significant repetition tended to correlate with coherence as perceived by a number of readers. In summary, Hoey's method of lexical cohesion analysis sees coherence and cohesion as ultimately interrelated.

4.4.5 The sentence

Hoey argues that sentences may be seen as 'miniature packages' of information (p. 33). Their status is part grammatical, part textual: 'in so far as cohesion occurs across clause boundaries, it reveals the sentence to be a textual category; in so far as there are restrictions on the ways one may repeat within a sentence, the sentence is shown to be a grammatical category' (p. 216). If we take the sentence to be a whole unit, the question is raised of how cohesion contributes to creating relationships between sentences. For initial answers, Hoey draws on previous work by Winter (1979) who concentrated on demonstrating how lexical and grammatical devices enable readers to perceive relationships between pairs of sentences. According to Hoey, there are at least three key contributions from Winter which are relevant to understanding how cohesion relates pairs of sentences. Hoey

(1991b) draws from Winter the fact that repetition 'provides a framework for interpreting what is changed', therefore it has information value (p. 20). The role of repetition in showing relations between pairs of sentences is more satisfactorily accounted for if clusters of repetition are considered. Clusters of repetition create relations between sentences that may be at a distance from each other. These insights provide not only the framework for a methodology which looks at how cohesion relates to text organisation, but they also emphasize how important cohesion is for allowing relations between sentences to be perceived. Two fundamental arguments have been put forward above: sentences are textual units, and cohesion links sentences. The next logical step is to establish how cohesion relates to text organisation.

4.4.6 Links and bonds

The system of analysis proposed by Hoey to capture the number of connections between sentences is based on two key notions. The first of these is that of *links*, which occur whenever there is a repetition of an item in two separate sentences. The term 'link' is preferred to the traditional term 'tie' used by Halliday and Hasan (1976) because 'tie' implies directionality (Hoey, 1991b, p.52) while links indicate multidirectionality thus allowing for the creation of webs among lexical items. Furthermore, 'ties' include certain kinds of cohesion devices which do not count towards links (e.g. conjunction, collocation).

The majority of cohesive devices which count towards links are lexical. These include:

simple repetition Two identical items (e.g. bear – bear) or two similar items whose difference is 'explicable solely in terms of different choices from a grammatical paradigm' (Hoey, 1991b, p.58; e.g. bear (N) – bears (N)).

complex repetition Two similar items which share a lexical morpheme, or two identical items of different grammatical classes (e.g. human (N) – human (Adj), dampness – damp).

simple paraphrase Two different items of the same grammatical class which are 'interchangeable in some context' (Hoey, 1991b, p.69), and 'whenever a lexical item may substitute for another without loss or gain in specificity and with no discernible change in meaning' (Hoey, 1991b, p.62) (e.g. sedated – tranquilized).

complex paraphrase Two different items of the same or different grammatical class; this is restricted to three possibilities: (1) Antonyms which do not share a lexical morpheme (e.g. hot – cold); (2) Two items one of which 'is a complex repetition of the other, and also a simple paraphrase (or antonym) of a third' (Hoey, 1991b, p.64) (e.g. a complex paraphrase is recorded for 'record' and 'discotheque' if a simple paraphrase has been recorded for 'record' and 'disc', and a complex repetition has been recorded for 'disc' and 'discotheque'; and (3) When there is the possibility of substituting an item for another (for instance, a complex paraphrase is recorded between 'record' and 'discotheque' if 'record' can be replaced with 'disc').

superordinates and hyponyms Only if they have a common referent and if the hyponym comes first (e.g. 'bear' and 'animals' in 'a drug known to produce violent reactions in humans has been used for sedating grizzly *bears* ... To avoid potentially dangerous clashes between them and humans, scientists are trying to rehabilitate the *animals*').

Non-lexical repetition is also considered to form links. These include: (1) Third person personal pronouns; (2) 'you' and 'we' within quotation marks; (3) Demonstrative pronouns; (4) 'One', as in 'the first one'; (5) 'Do', as in

'do it'; (6) Clausal 'so' and 'not' as in 'they said so', 'they said not'; (7) 'Other', 'another', 'the other', '(the) same'; (8) 'Different' and 'similar'; (9) Ellipsis. These grammatical devices for lexical repetition are grouped into co-reference (type 1 above), substitution (most cases of types 2 through 8), and ellipsis (type 9).

To illustrate the concept of link, which is central to both Hoey's model and the approach to segmentation implemented in later chapters, we shall examine the links established by the first sentence of a short news report:

(1) A drug known to **produce** violent reactions in **humans** has been **used** for **sedating** grizzly **bears** *Ursus arctos* in Montana, USA, according to a report in *The New York Times*.
(2) After one **bear**, known to be a peaceable animal, killed and ate a camper in an unprovoked attack, scientists discovered it had been **tranquilized** 11 times with phencyclidine, or 'angel dust', which **causes** hallucinations and sometimes gives the **user** an irrational feeling of destructive power.
(3) Many wild **bears** have become 'garbage junkies', feeding from dumps around **human** developments.
(Adapted from Hoey, 1991b, p.37)

The links that sentence 1 shares with the other sentences appear in bold. Sentence 1 has four links with sentence 2, namely produce→causes (simple paraphrase), used→user (complex repetition), sedating→tranquilized (simple-paraphrase), and bears→bear (simple repetition), and two links with sentence 3: bears→bears (simple repetition), and humans→human (complex repetition).

The verb 'known' appears in both sentences 1 and 2, but it does not create a link because Hoey considered the contexts in which they occur to be different. Hoey applied in such cases the 'shared context criterion', according to which a link is formed between two items if there is evidence in the immediate context of both items that they refer to the same object or situation. In the specific case of 'known', Hoey argued that the two instances have 'nothing in common with regard to such features as unstated "knower" and topic of knowledge' (Hoey 1991b, p.37; see further exemplification on

p.260).

Links are connections between items, but as was stressed before, Hoey's system is devoted to finding connections between messages, i.e. sentences. The count of links must therefore be made between sentences if a measure of the association between messages is to be achieved.

Hoey (1991b) proposes the concept of *bonding* to account for relations between sentences. A *bond* is established whenever there is an above-average degree of linkage between two sentences. It can be defined as 'a connection between any two sentences by virtue of there being a sufficient number of links between them' (p. 91). Normally, three links constitute a bond. Hoey stresses that the number of links which constitute a bond is relative to the type of text and to the average number of links in the text (p. 91), but the least number of links is three 'because of the greater likelihood of two repetitions occurring in a pair of sentences by chance' (p. 190). For example, the two sentences in figure 4.5 are bonded by three links: writings – writer, political – political, and past – past.

Bonded sentence pairs have certain important characteristics. Bonded sentences normally share common content, and are semantically related or even coherent. In the example in figure 4.5, the first sentence 'specifies what the writer is offering the reader', while the second sentence 'raises the issue of

[1] What is attempted in the following volume is to present to the reader a series of actual excerpts from the **writings** of the greatest **political** theorists of the **past**; selected and arranged so as to show the mutual coherence of various parts of an author's thought and his historical relation to his predecessors or successors; and accompanied by introductory notes and intervening comments designed to assist the understanding of the meaning and importance of the doctrine quoted. [17] What, then, is the advantage which we may hope to derive from a study of the **political writers** of the **past**?

Figure 4.5: Example of bonded sentences (adapted from Hoey, 1991b, p.129)

what the reader might gain from the offer' (Hoey, 1991b, p.129). Importantly, the two sentences in the example are separated by sixteen sentences. When the relatedness is not easily perceived, it is usually because of a restricted number of factors, such as excessive repetition, voice choice, and modal choice (Hoey, 1991b, pp. 134-138).

4.4.7 Repetition matrices

The representation of links in a net as shown above (see figure 4.4 on page 151) reveals how links form multiple connections. However, a net is not an appropriate method for showing in detail the non-linearity of links. For that purpose, Hoey (1991b) uses a *repetition matrix*.

A repetition matrix records the links between a particular sentence and all the other sentences in the text. Hoey distinguishes between repetition matrices which show links itemized, and those which display links counted. In the former, the actual words which form the links between sentences are included, whereas in the latter only the number of links between sentences is given. In the presentation that follows, only the latter kind will be addressed.

The repetition matrix is constructed by drawing a series of rows and columns, one for each sentence in the text. The columns are numbered beginning with the number of first sentence of the text, whereas the rows are numbered starting with the second sentence in the text. The resulting cells are filled in with the number of links between the pairs of sentences represented by the intersection of each row and column. Rows indicate the number of links between a sentence and those which preceded it in the text, while columns represent the links between a sentence and those which followed it. A matrix designed in this manner would be redundant in that the links for each pair of sentence would be recorded twice. To avoid this, the matrix is divided in two along its main diagonal running from the top left-hand corner

to the bottom right-hand corner, and only the bottom half of the matrix is actually utilized.

Figure 4.6 shows a matrix for the five initial sentences of 'Masters of Political Thought', which is reproduced in appendix 4 on page 457. The numbers in brackets ((1), (2), etc) represent sentence numbers; the other numbers in the cells indicate the number of links between pairs of sentences. For instance, the number '6' at the top of the matrix shows that there are six links between sentences (1) and (2); there are also 2 links between sentences (1) and (3), 5 links between sentences (1) and (4), and so on.

	(1)			
(2)	6	(2)		
(3)	2	1	(3)	
(4)	5	1	2	(4)
(5)	1	0	1	0

Figure 4.6: Partial repetition matrix for 'Masters of Political Thought' (adapted from Hoey, 1991b, p.90; see appendix 4 on page 457 for text)

Inspecting a repetition matrix can reveal an important aspect of the lexical cohesion of the text, namely where those sentences sharing a large number of connections occur in the text, which in turn can indicate densities of connection across the text. In the sample matrix in figure 4.6, one can notice a dense area of linkage between sentences 1 and 2 and between sentences 1 and 4, compared to the remaining sentence pairs. In a longer text, such densities are particularly interesting in that they may reveal potential segmentation points (see pilot study 1 in section 5.2 on page 191).

4.4.8 Central, marginal, topic-opening and topic-closing sentences

Bonds can be computed across a number of sentences, not only between individual sentences, as the example in figure 4.5 on page 158 shows. This allows for the identification of sentences which share more bonds in the text, which in turn can lead to the classification of sentences in terms of their degree of bonding. A first classification is between central and marginal sentences. The former are sentences which have a high number of bonds, being by definition 'the most bonded sentences' in the text (Hoey, 1991b, p.265). As with the number of links which constitutes a bond, the number of bonds which constitute a central sentence is also relative, though, and must be decided on the basis of the distribution of bonds in the text under consideration. To illustrate the concept of central sentences, we shall examine the partial matrix for the text 'Masters of Political Thought' presented above in figure 4.6 on the preceding page. According to the matrix, sentences 1 and 2 are the most bonded, sharing six links between them, and are therefore the central sentences in the passage:

> [1] What is attempted in the following volume is to present to the reader a series of actual excerpts from the writings of the greatest political theorists of the past; selected and arranged so as to show the mutual coherence of various parts of an author's thought and his historical relation to his predecessors or successors; and accompanied by introductory notes and intervening comments designed to assist the understanding of the meaning and importance of the doctrine quoted. [2] The book does not purport to be a history of political theory, with quotations interspersed to illustrate the history.
> (Hoey, 1991b, p.78)

The two sentences present the aims of the book, and in so doing they represent the main theme of the passage. As Hoey (1991b, p.43) puts it, these two sentences 'are germane to the development of the theme(s)', and as such they are indeed 'central', which is supportive of the impression gained by examining the matrix.

By contrast, the remaining three sentences have 'low information value'

(Hoey, 1991b, p.45), and their role is to clarify certain aspects of the material included in the volume, providing exemplification of the general nature of the book (a collection of texts), the key authors in it (Aristotle, Augustine), and a note about its limitations (it is not exhaustive):

> [3] It is rather a collection of texts, to which I have endeavoured to supply a commentary. [4] I have tried rather to render the work of Aristotle, Augustine, and the rest accessible to the students, than to write a book about them; and the main object of this work will have been achieved if it serves not as a substitute for a further study of the actual works of these authors, but as an incentive to undertake it. [5] Nor does the commentary make any pretension of being exhaustive.
> (Hoey, 1991b, p.78)

Essentially, what sentences 3, 4 and 5 do is provide support for the two initial sentences, and as such they function as 'marginal sentences' in the passage. Thus, the interpretation of the role of these three sentences supports the prediction made by examining the matrix.

A further classification can be made between topic opening and topic closing sentences. A sentence is topic opening if it bonds with more subsequent than preceding sentences, and it is topic closing if it bonds more times with preceding sentences. The first step in identifying topic opening and topic closing sentences is to calculate the number of bonds each sentence has with its predecessors and its followers. For instance, taking three links as forming a bond, the following listing can be extracted from the matrix in figure 4.6 on page 160:

Sentence	Before	After
1	0	2
2	1	0
3	0	0
4	1	0
5	0	0

According to the table above, the sentence having the most bonds with later sentences is sentence 1, which is bonded to sentences 2 and 4. Sentence 1 is therefore the topic opening sentence in the excerpt. The topic closing

sentences are sentences 2 and 4, since both of them have more bonds with preceding sentences (sentence 1).

Sentence 1 opens a topic which might be described as 'the aims of the book', as the phrase 'what is attempted in the following volume' seems to indicate, whereas sentences 2 and 4 present additional information about the same topic. In this manner, sentences 2 and 4 seem to function to close the topic initiated in sentence 1.

Topic opening and topic closing sentences can also be used as a means for summarising texts. Accordingly, sentences 1, 2 and 4 can be taken as representing a fair abridgment of the passage as a whole: sentences 1 and 2 present the aims of the book in more general terms, while sentence 4 gives supporting detail about the scope of the book. Automatic text summarisation is an important application of the model of analysis proposed by Hoey (1991b), and has been taken up in other studies, some of which are discussed in what follows (see p.164ff. and p.168ff.).

4.4.9 Implications

The work of Hoey (1991b) is ideal for the present investigation by being amenable to computer treatment, and also because it stresses the importance of lexical cohesion among the other types of cohesion. His approach is central not only to segmentation but to a theory of text organisation because it claims a fundamental role for lexis in building text. One implication is that the study of lexical cohesion must be essentially a study of how cohesion organises text rather than how cohesive ties can be classified (p.3). The way in which Hoey views the relationship between coherence and cohesion is also relevant to the previous study. His view that sentences are 'miniature packages' of information agrees with Grimes's (1975, p.108) notion of the sentence as being 'packages of information that are wrapped up and labelled

in a standardized form for the hearer's benefit'. The status of sentences as meaningful units of information in text makes them ideal units for computerized analysis since the computer can be programmed to recognize sentence boundaries.

4.4.10 Related studies

In this section studies which have been based on Hoey's (1991b) methodology are reviewed.

Benbrahim

Benbrahim (1996) and Benbrahim and Ahmad (1994) apply the methodology introduced by Hoey (1991b) to the production of abridgments and term banks. Their major goal is to automatize the analysis so that it can be applied to longer texts.

The study In order to automatize the analysis of bonds, Benbrahim and Ahmad (1994) created a special computer program named 'Tele-Pattan' which carries out an analysis of texts according to links and bonds. Apart from identifying links and bonds, 'Tele-Pattan' has graphic capabilities which allow the user to visualize bond networks in detail. In Benbrahim (1996), 5 academic English texts are examined, and in Benbrahim and Ahmad (1994), both an English and a Welsh text are abridged.

The computation of simple lexical repetition is carried out by simple matching, but for the identification of simple and complex paraphrase the authors employ thesauri, either the Macquarie Thesaurus, which has some 180,000 terms (Benbrahim and Ahmad, 1994) or WordNet, with 164,000 entries (Benbrahim, 1996). Macquarie Thesaurus was replaced with WordNet because the latter has important advantages such as being integrated

into interconnected synonym sets (instead of separate entries identified by ad-hoc labels), and being a computer database. Regardless of the specific thesaurus employed in the computation, the use of a thesaurus allows them to automatically compute complex paraphrase by means of the 'link triangle', i.e. the link which results between two items which are linked to a third by means of complex repetition or simple paraphrase.

Their use of bonds and links is mainly directed towards the production of automatic sentence-based summaries, which can be of four kinds depending on which kinds of sentences they contain: topic opening sentences only; topic opening, topic closing, and central sentences; key central sentences; and finally all bonded sentences ('non-marginal') (Benbrahim and Ahmad, 1994, pp. 30,38). Three of these methods had already been introduced by Hoey (1991b), with the exception of the key central sentence approach. Key central sentences are defined by the authors as those which present a number of bonds calculated as a percentage of the maximum number of bonds presented by any one sentence in the text. For instance, if the most bonded sentence has 10 bonds, a threshold may be set at 70% of 10 bonds which will exclude all those sentences which have fewer than 7 bonds, the remaining central sentences being considered to be key. The authors consider such summaries to be of a 'more precise' kind (p.38), although it is not particularly clear in which way. A further type of summary is introduced that is not based on pulling out individual sentences but whole paragraphs. This method is discussed in Benrahim (1996, pp.115-123). The advantages of this method is that the summaries contain more fluid prose with fewer gaps and, in many cases, the original introductory and closing paragraphs, thus yielding a more readable rendition of the input text.

The authors innovate in offering a comprehensive measure of bonding called 'connectedness density' for each sentence. Connectedness density ra-

tios are calculated for each sentence, and they incorporate information about the size of the text and the direction of the bonding ('before' or 'after' counts). In this way, connectedness density can function as a replacement for total bond counts in deciding on cut-off values for centrality. The formula for the connectedness density ratio is:

$$D_{s_i} = \frac{(B_i^2 + A_i^2)^{\frac{1}{2}}}{N\sqrt{2}}$$

where D_{s_i} is the connectedness density for sentence i, B stands for the number of bonds with previous sentences, A the number of bonds with subsequent sentences, and N represents the number of sentences in the text. So, for a sentence from a 100-sentence text having 10 bonds, 3 of which are with previous sentences, its connectedness ratio would be 0.0539 or $(3^2 + 7^2)^{\frac{1}{2}} \div N\sqrt{2}$. It is not self-evident how useful it is to represent the bonding information for this particular sentence as 0.0539 instead of say 0.1 which is simply the number of bonds divided by the total sentences in the text, since neither of these indexes shows how noteworthy, relevant, or indeed high or low even, a bond count of 10 sentences is.

Among other potential applications of bonding analysis, they cite text-retrieval, and domain-specific and text-specific key word extraction (or 'terminology acquisition' and 'document indexing' respectively). The authors developed a system known as 'Quirk' to accomplish such tasks. The system uses links and bonds to compare a particular text with a corpus so as to determine whether the text is congruent with that corpus. Conversely, the system also uses the same principles in order to extract a relevant text from a corpus.

Benbrahim (1996) offers a detailed count of the number and types of links in a number of texts. He notes that on average, for sentences bonded

Links per bond	Simple repetition	Complex repetition	Simple paraphrase	Simple + complex repetition
2	78%	52%	17%	94%
3	69%	32%	3%	96%
4	63%	15%	1%	92%
5	59%	11%	0%	94%
6	56%	7%	1%	92%
7	60%	7%	0%	97%

Table 4.2: Percentage of types of links in bonded sentences (adapted from Benbrahim, 1996, p.95)

at 2 links, about $\frac{3}{4}$ of the bonds in his texts are formed by simple repetition links; the addition of complex repetition links only increases the coverage by no more than 16% to 94%, while the remaining 6% are completed by the inclusion of simple mutual paraphrase (see table 4.2). He also observes that the contribution of each type of link varies as the number of links required to make a bond increases. So, for 7 links, 60% of the bonded sentences have simple repetitions, but 97% have simple and complex repetitions. Two trends are observable: first, the fastest decreasing share is that of simple mutual paraphrase (dropping from 17% at 2 links to nothing at 7 links); second, the least changing combination is that of simple and complex repetitions, whose participation varies from 92% to 97%. Complex repetitions on their own account for very few links as the bonds increase (7%), while simple repetitions maintain the largest single share (60%) despite a general tendency to drop as the number of links per bond increases.

Implications A major contribution of Benbrahim's (1996) study is the exhaustive counts of the types of links which contribute to bonding at several bond thresholds. His counts suggest that it is generally not necessary to compute all kinds of links in order to obtain a comprehensive retrieval of

all possible bonds in the text. If the analyst has to make a choice of which types of link to record, by computing simple lexical repetitions he/she should account for a great share (much more than half) of the bonds in the text.

In general, Benbrahim and Ahmad (1994) and Benbrahim (1996) suggest that computers can be used for identifying lexical links in texts. The automatic identification of lexical links is a task on which the analysis for segmentation will depend.

Renouf and Collier

Another approach to automatic summarisation is provided by Renouf and Collier (1995), who present an implementation of a summarisation procedure based on Michael Hoey's notion of 'bonding'. They report on their experience in developing a commercial abridgment system based on bonding analysis.

The study Renouf and Collier (1995) use link and bond counts to generate abridgments of expository text. The system works by tabulating the number of links between sentences and then selecting those sentences which bond at a certain level. The number of bonds which count as a bond, as well as the number of bonds which a sentence needs to have for inclusion in the abridgment, can be controlled by the user of the system. The authors point out that at the moment only simple and lexical repetition were handled by the system, even though it would have been desirable to include paraphrases and, importantly, pronominal reference. A sample analysis of a newspaper report is presented in which abridgments are created at different levels of linking and bonding. Even though different abridgements are produced each time, the authors observe that all versions have three sentences in common. These sentences are considered key sentences in that they seem to indicate the main components of the texts. The authors also discuss the fact that one

of the constant sentences was an initial one in the text, which indicates the important role by introductions in newspaper stories. It is argued that all versions of the abridgments are readable, and that since the system works very fast users have a choice of the version which best pleases them without effort. They conclude that their automatic implementation of Hoey's method of analysis seems promising.

Implications The application of bonding analysis to abridging forms part of the original proposal describing bonding analysis in Hoey (1991b). Renouf and Collier's (1995) report does not represent a departure from the original formulation. Similarly to Benbrahim and Ahmad (1994) and Benbrahim (1996) discussed above, Renouf and Collier's (1995) work is relevant in that it presents a computational implementation to finding lexical links. The automatic identification of lexical links will also be carried out in the present study. The fact that they report success in their implementation suggests that using computers to carry out an analysis based on Hoey's method is feasible.

Collier

Links and bonds are not found in running text only; Collier (1994) applies the concept of bonding to the task of sorting concordance lines. While his study is not directly related to the role of lexical cohesion in texts, his application of Hoey's methodology suggests that links and bonds have a role as a general measure of association between any two strings of text.

The study In his application of bonding analysis to concordance line selection, Collier (1994) argues that such a method is necessary because not only have corpora grown in size but the sorting of concordance lines is expected to be accomplished automatically. If corpora have grown, so have the number

of lines the analyst is supposed to sort at one time. The author argues that the application of lexical cohesion to the problem of sorting concordance lines is advantageous because it can lead to the identification of patterns across concordance lines. Therefore the kind of sorting which lexical cohesion permits is different from the usual alphabetical sorting and thus can lead to an improvement in identification of collocational patterns. The underlying assumption is that concordance lines can be cohesive just as sentences can, which draws on the original notion of matching introduced by Winter (1974). Collier distinguishes 'central lines' as being those which form cohesive links with a criterial number of other concordance lines. These should be central in that they might serve as candidates for examples in dictionary entries by being representative of a set of other concordance lines. Similarly, a set of bonded concordance lines should present common linguistic features, that is, similar collocational patterns. In the identification of central lines, the number of links and bonds can be controlled for.

The advantage of the use of lexical cohesion across concordance lines is that it allows for 'gapped' patterns to be picked out which is much more difficult to achieve with simple sorting by fixed position. For example, if the same collocate appears before and after the node word, it will normally not be identified by the usual means of sorting, but the presence of the same lexical item in different lines will count towards a link regardless of the position of the lexical item in relation to the node (this parameter can be adjusted, though). This means that more flexible patterns are capable of being retrieved by using lexical cohesion.

Implications While his study does not relate directly to the role of lexical cohesion in textual organisation, the work of Collier (1994) is noteworthy for the present study in that it suggests that links and bonds are analytical

devices which serve to indicate strength of association between two strings of text, be they sentences or other strings such as concordance lines. This lends more support for the use of links as a device for finding similarity between stretches of text, a task which is at the center of the investigation of segmentation.

Wessels

The work of Hoey (1991b) has also been used for the investigation of the relationship between bonding and quality of student writing (Wessels, 1993b). In her study, Wessels (1993b) looks at whether there is a relation between bonding and perceived coherence in student compositions.

The study In the process of tabulating the frequency of bonding across the student texts, Wessels (1993b) notices that a set of 5 bonded sentences (for instance sentences 1 through 5) can be represented as having 4 bonded sentences (sentence 1 to 2, 2 to 3, 3 to 4, and 4 to 5) or as having 10 bonds (1 to 2, 1 to 3, 1 to 4, 1 to 5, 2 to 3, 2 to 4, 2 to 5, 3 to 4, 3 to 5, and 4 to 5). She calls the former ratio 'degree of bonding', and the latter 'bonding density'. She argues that a differentiation is necessary because 'bonding density' might better account for the level of integration in written text which seems to be expected of student writers, and therefore it might relate to quality of student writing. Initially, the texts were 40 compositions written by students during examination, which were rated for quality based on a four-point scale by two experienced teachers of English. A final score was arrived at for each essay based on the average rating given by the teachers. The final sample was made up by the 13 highest and the 13 lowest scoring essays. The results suggest that the highest scoring and lowest scoring essays did not differ statistically in terms of bonding density, as the less coherent essays had on average .95 bonds

per sentence, while each sentence in the more coherent texts had an average of 1.5 bonds (t=1.48, p=.1542). Similarly, the percentage of bonded sentences seemed not to distinguish between the two groups: the more coherent texts had a slightly higher number of bonded sentences (about 62% of the sentences in each text were bonded) than the less coherent texts (49.5% had bonded sentences), but this was not a statistically significant difference. The author concludes that bonding is a poor discriminator of coherence in student writing and that qualitative measures should be used instead.

Implications The fact that Wessels (1993b) did not find a statistically significant correlation between bonding and perceived coherence suggests that lexical cohesion is not by itself a predictor of quality of writing. This is not surprising since before her study Parsons (1990) had already reached a similar conclusion by observing that the frequency of lexical chains accounted for nearly a third of the perceived coherence in student's compositions. Both studies suggest that despite the fact that bonds and lexical chains are in principle well-suited for explaining writing quality, measures based on these constructs fail to show how texts differ in terms of coherence. This can be taken to mean that although lexical cohesion is an element of texts it does not reflect the quality of texts.

4.5 Pêcheux

A scholar who has used an approach to discourse analysis which is related to lexical cohesion (although he does not refer to it as such) in written text is Michel Pêcheux (Hak and Helsloot, 1995). He has pioneered a system of analysis which he called 'Automatic Discourse Analysis', or 'ADA'. The aims of ADA are to find design domains and hyperdomains which are constituted by connecting stretches from several discourses. It has therefore an intertex-

tual orientation while most approaches discussed so far are predominantly intratextual[1].

4.5.1 Autonomous discursive sequence

ADA works by searching for a particular autonomous discursive sequence in a corpus; an autonomous discursive sequence is typically a sentence, but it may consist of various sentences which display thematic unity. Autonomous discursive sequences are identified manually; they are then broken down into utterances, which are then paired up into 'binary relations'. The list of binary relations is then searched for in the corpus and those relations which present similarities with others across the corpus are called 'quadruplets' (p.169).

To illustrate, figure 4.7 on page 175 presents two autonomous discursive sequences taken from the speech by François Mitterrand at the Socialist Party Congress in 1979. These discursive sequences are found to be related to each other by means of the repetition of 'gouvernement' and 'PC'. Other elements are not repeated but there is parallelism between 'participer' and 'préférer', and between 'union' and 'droite' which brings out the association between the two autonomous discursive sequences.

Formally, these associations, or 'paraphrase-effects' are identified by the coding which assigns each element of the utterances to eight-morphosyntactic categories: (1) F: form of the utterance, that is, voice, modality, tense, etc; (2) DET1: determiner of N1; (3) N1: Noun in subject position; (4) V: verb; (5) ADV: adjectival, verbal, or phrasal verb; (6) P: Preposition governed by a verb; (7) DET2: determiner of N2; and (8) N2: Noun in object position, adjective, or meta-term S reflecting an objective clause or free adjunct (p.194). Thus, two utterances which are related from the autonomous discursive re-

[1] However, Hoey's (1991b) approach has been adapted for investigating intertextuality (Berber Sardinha, 1995d; Hoey, 1995b, cf.).

lations presented above are coded as in figure 4.7.

4.5.2 Contributions of ADA

ADA is seen as a contribution to a sociology of discourse (p.89). This involves seeing how power relations and meanings are expressed in text. The investigation of such meanings is couched in the utterance since Pêcheux recognizes that word frequency alone cannot offer insights into how words are used in metaphorical contexts. For instance, he mentions the fact that the concept of 'freedom' means totally different things to the governor of a prison and to the prisoners themselves.

ADA is offered as a methodology for answering important questions which have been avoided since Saussurean linguistics became established, such as 'what does this text mean' and 'how does the meaning of this text differ from that of another' (p.64). These questions remain relevant, and sadly, largely ignored.

4.5.3 Implications

The work of Michel Pêcheux lends support to the key role of repetition in texts. His account of the similarity between sentences by describing the parallelism revealed by repetition is not dissimilar from how Hoey (1991b) shows the parallelism between bonded sentences. The fact that two different analysts, working in different linguistic traditions, for different purposes, have reached similar conclusions about the role of repetition in assisting in the perception of parallelism and similarity between sentences can be taken as strong indication that the role of repetition in linking sentences cannot be disregarded. This is relevant to the present study in that the methodology used in the analysis for segmentation relies on repetition.

```
F   |DET1|N1          |V         |ADV |P |DET2|N2
----+----+------------+----------+----+--+----+------------
0003|L   |PC          |PARTICIPER|SEUL|A |DS  |GOUVERNEMENT
0000|R   |PC          |PREFERER  |0   |* |L   |GOUVERNEMENT
0000|R   |GOUVERNEMENT|E         |0   |DE|0   |UNION
0000|R   |GOUVERNEMENT|E         |0   |DE|L   |DROITE
```

1. Le Parti communiste n'a participé (avec de Gaulle, Gouin, Bidault et Ramadier) qu'à des gouvernements d'union nationale de concentration républicaine

2. Le point qui nous importe aujord'hui est de savoir s'il est imaginable que le PC change d'attitude, cesse bientôt de considérer les socialistes comme des adversaires principaux, et de préférer le gouvernement de la droite et du grande capital à la victoire des travailleurs. Rien ne le montre. (pp.194-5)

Figure 4.7: Sample ADA analysis

4.6 Conclusion

Despite differences in focus, the majority of approaches reviewed here can be subsumed under two headings: (1) lexical chains or strings, and (2) lexical clusters. The first group is both more numerous (Eggins, 1994; Halliday and Hasan, 1976; Hasan, 1989; Halliday, 1985; Parsons, 1990) and more traditional in that it centres around the original proposal by Halliday and Hasan (1976). This can perhaps explain why the lexical chain approach has been widely used in computational approaches to segmentation (see previous chapter).

The second group consists of one major contribution, namely the approach to lexical cohesion by Hoey (1991b). In addition to being less numerous, the cluster approach is also more recent. It can be seen as a development of previous approaches, notably Halliday and Hasan (1976) and Hasan (1989).

In spite of not having been used by previous studies on computational

segmentation, Hoey's (1991b) approach is the one which most readily lends itself to computerized treatment. As studies on automatic summarization using Hoey's approach have shown, the automatic computation of links and bonds can provide good results in terms of acceptable abridgments of full texts. One of the reasons why Hoey's (1991b) approach has not been used for segmentation might be that in a sense its goal is the opposite of that which can be assumed for segmentation. The principal goal of segmenting texts is to show how a text can be divided into parts; in Hoey's (1991b) approach, on the other hand, the main aim is to show how texts are integrated by lexical cohesion. This apparent lack of fit could perhaps explain why Hoey's approach has not been incorporated in studies on computational segmentation.

A major feature of Hoey's (1991b) approach to lexical cohesion is its inclusion of systematic repetition as a major element in creating cohesion in texts. For this insight he draws on previous work by Winter (1974) who stressed the *meaning sharing* role of repetition. Though in different ways, other researchers have also emphasized the crucial meaning sharing function of repetition. For instance, Pêcheux (1969/1995) has devised a methodology for automatic discourse analysis which draws heavily on the repetition of lexical and grammatical items. Similarly, Winburne (1962) has also been concerned with how sentences attach to one another through repetition. In this manner, important connections can be made between Hoey's (1991b) approach and other studies which have preceded it.

There are also connections between Hoey's cluster approach to lexical cohesion and the other major group of lexical cohesion studies identified above, namely the group concerned with lexical chains and strings. Clearly, repetition also contributes to the formation of chains and strings. In this manner, the boundary between the two major camps seems to have been blurred in

that all approaches reported on in this chapter make use of repetition to a greater or lesser degree in order to assess lexical cohesion in texts.

The review and the critical commentary provided in this chapter present a theoretical basis for choosing lexical repetition as a measure of lexical cohesion. At the same time, this choice would also be fortunate in that it would be readily amenable to computerized treatment. In other words, there is also a practical motivation for selecting repetition. The computation of repetition is totally compatible with an inductive approach to data analysis, and repetition is a measure which can be objectively accounted for. Therefore, the computation of lexical repetition seems an ideal choice for the investigation of text segmentation by computer.

However, several issues need to be resolved before repetition can be adequately used as a criterion for segmenting texts. For instance, which approach best lends itself to computerized treatment? Which approach lends itself to segmentation? The review has suggested that Hoey's approach has been adapted for computer applications, yet it has never been applied to segmentation. On the other hand, lexical chains have been widely used in computational segmentation, yet lexical chains have proved difficult to implement fully. Further, the original aim of Hoey's proposal was to show integration rather than segmentation; therefore on the face of it Hoey's approach does not lend itself directly to segmentation. Nevertheless, Hoey's approach is based on the notion of the clustering of lexical cohesion, which is appealing since commonsensically it is possible to think of segments as clusters of linguistic elements which belong together. We think of a text as a sequence of 'chunks' which each have some kind of internal consistence, and which are fitted together by various kinds of connections. It makes sense to hypothesize that clustering will take place within chunks more than across chunks (though this does not rule out co-clustering of separate chunks at a

distance in the text). All of these are questions which need to be investigated before a definitive computer-aided methodology can be suggested to account for segmentation in written texts. These questions have been addressed in a series of pilot studies which have taken place during the years in which this thesis has been in preparation. Some of the most relevant of these studies are reported in the chapter which follows.

4.7 Summary

Before ending this chapter and moving on to experimentation, it is perhaps useful to provide a summary of the central arguments which have been presented in the thesis so far.

In chapter 2 it was suggested that segmentation is a common task in discourse analysis. However, existing approaches provide a framework for segmentation which is restricted in many aspects. An alternative framework was proposed whose desiderata would have to include: extensive coverage, inductive data treatment, and independent validation. It was argued that the first two requirements could only be adequately met by using a computerized approach to segmentation. The decision was then to review the existing approaches to segmentation.

Chapter 3 provided a description of the major approaches to segmentation by computer. In the chapter, it was observed that a common approach was to identify the lexical cohesion in texts, since the identification of repetition lends itself to computerized treatment. In addition, it was noted that lexical cohesion is empirically viewed as having a natural connection with segments. Nevertheless, there seemed to be a consensus around the use of lexical chains, even though lexical chains are problematic to compute because of the many relations which can enter in the chains, most of which are not self-evident to

the computer. The decision was then taken to survey the major approaches to lexical cohesion in search of possible alternatives to lexical chains.

In the present chapter, the survey of major approaches to lexical cohesion indicated that the only major alternative to lexical chains was that offered by Hoey (1991b) based on the clustering of lexical cohesion among sentences. It was further argued that a key insight in Hoey's approach is that of the central role of repetition in creating cohesion between sentences. By looking at the field of lexical cohesion from the point of view of repetition, it was possible to find a point of contact among the various approaches. Thus, repetition was proposed as a starting point for the investigation of segmentation by computer.

The next chapter begins to address the problem of segmenting texts automatically. The chapter reports on three pilot studies carried out during the four years in which the present thesis was in preparation. The pilot studies were designed to meet the three criteria for an approach to segmentation by computers presented in chapter 2 (see p.83 onwards), namely extensive coverage (the ability to handle large amounts of data), inductive orientation (the ability to refrain from imposing *a priori* categories on the data), and objective evaluation (the ability to assess the performance of segmentation objectively against an independent reference).

Chapter 5

Pilot studies

5.1 Introduction

This chapter begins with an overview of the research pertinent to the investigation of segmentation, and is aimed at showing the relevance of applying Hoey's (1991b) approach to the analysis of lexical patterns to the investigation of segmentation. It will be argued that there was no straightforward obvious way of making use of Hoey's (1991b) approach to lexical patterns as a segmentation method, and therefore a number of attempts had to be made in that direction. These attempts are reported here as pilot studies. Each pilot study revealed important aspects about the way texts segment, and each of them were also controlled by specific criteria.

The chapter begins with an with an overview of the gaps in the research pertinent to the investigation of segmentation, followed by a suggestion for filling these gaps, and ending with a proposal for operationalizing the suggestions put forward.

5.1.1 Overview of previous research

The review of the literature presented in the last three chapters identified several major characteristics of previous research in segmentation using both computational and non-computational means. In the review of existing discourse analytical approaches to segmentation (chapter 2), several key features were mentioned which seem to apply to an appreciable extent to nearly all those approaches.

A first feature is the restricted amount of data typically dealt with in discourse analysis. With few exceptions (Mann et al., 1989; Stoddard, 1991), most studies of discourse structure are restricted to the examination of a few individual texts, which raises the issue of whether the views proposed by such studies are in fact representative of a text type or genre or whether they are only applicable to a restricted number of individual texts (Biber, 1993).

A second feature is the restricted length of the individual texts analysed. Since different text types vary in length, there cannot be a definition of text length which is valid for all texts. By concentrating on texts which can 'fit on the blackboard' (Phillips, 1989, p.8), previous research in discourse analysis has largely been unable 'to see patterns that don't emerge either from modest sets of samples or from introspection and intuition' (de Beaugrande, 1997, p.41). These issues have helped discourse analysis earn a reputation as a field which is 'all program with no analysis, or simple analysis with no program' (Frawley, 1987, p.371).

A third feature includes an interest in labelling segments; in other words, the identification of segments is accompanied by the application of labels ('problem', 'inciting moment', 'establishing a niche', etc) which designate the content or function of the segments in a discourse model. The labelling allows the analyst to incorporate the segments into an organized framework, showing how the individually labelled segments work together as a model.

A fourth feature, which is related to the creation of models through labelling, is the imposition of a top-down orientation towards the data through the application of models. Top-down processing gives rise to the establishment of models which make *a priori* assumptions about the organisation of the data instead of adapting to the reality of the data as they present themselves in the text. As Sinclair (1994, p.13) argues, the analyst should 'trust the text':

> We should strive to be open to the patterns observable in language in quantity as we now have it. The growing evidence that we have suggests that there is to be found a wealth of meaningful patterns that, with current perspectives, we are not led to expect. (...) The first stage should be an attempt to inspect the data with as little attention as possible to theory.

The majority of models are static (Ventola, 1986), and therefore not adaptable to individual variation in text constitution, which can result in a lack of fit between the intended structure as predicted by the model and the actual organisation as realised by the text. As the previous quotation from Sinclair (1994) suggests, it is wiser to aim for a textual description which evolves out of the observation of the patterns in the data than to start with a set of pre-defined categories and impose them on the text.

A final feature shared by approaches to segmentation in discourse analysis refers to the issue of validating the analysis. This aspect has been largely ignored by research in discourse analysis. Most analyses are presented as being 'the truth' (sometimes the only truth) about a text or a genre. Previous research has not tackled the issue of whether two different analysts using the same model would arrive at different segmentations of a text. One exception is Mann and Thompson (1987a, 1987b, 1988), who openly declare that their RST model is interpretive and that separate analyses based on it may diverge; another is Longacre (1983) who recognises the role of intuition in identifying

episodes and attributing episode marking status to certain expressions. It would be a most welcome addition to discourse analysis if proponents of models would be willing to go some way towards showing whether their analyses meet any objective criteria and do therefore lend themselves more directly to replication, or whether the analysis is inherently subjective and is therefore more likely to produce different results in individual circumstances.

5.1.2 Gaps in the literature

From the features discussed above, four gaps in the literature can be identified. The first gap in the literature is with respect to research dealing with large quantities of texts. Research in corpus linguistics does not qualify to fill this gap because corpus linguistics is not concerned with analysis of individual texts but with the analysis of collections of texts regardless of individual text boundaries. The second gap refers to the length of texts normally investigated in research in discourse analysis. As Phillips (1985) argues, the typical amount of data in discourse analysis is that which can fit on the blackboard. Likewise, Biber (1995b, p.344) further notes that discourse analyses are 'typically based on a few thousand words of text'. The third gap is with respect to the need for bottom-up approaches to data analysis. Bottom-up or inductive orientation approaches textual data by trying to induce the segment divisions from the characteristics found in the data, rather than from the opposite direction, by trying to segment texts by imposing elements of a pre-defined model. Examples of inductive data processing in discourse analysis are rare, a notable exception being Phillips (1989, 1985), who looked at how the distribution of lexis produced collocational networks in science textbooks which in turn revealed connections across chapters. The final gap in previous research relates to the lack of concern with validation. One way a model can be validated is by checking it against an independent

criterion, such as an analysis carried out by another analyst, or a valid independent reference. Normally, it is difficult to find other analysts who are willing to analyse the same texts and therefore the first option is less practical. Hence, the second option, namely that of checking the analysis against a valid independent reference, presents itself as a more viable alternative.

The problem arose, though, of choosing a reference for comparative purposes. As argued in section 2.5 (see p.85 ff.), existing divisions in written texts provide a valid reference for segmentation research, mainly because they represent the segmentation decisions supplied by the author(s) of the text. Goutsos (1996b, p.82), for instance, argued that orthographic divisions are the most important means for signalling topic shift in written texts. Three units larger than the sentence present themselves as candidates: paragraphs, sections, and chapters. Recent research into paragraphing (Hoey, 1996) suggests that the insertion of paragraph breaks seems to have less to do with the perception of coherent sub-units of text than with the occurrence of specific paragraph-initial expressions. This finding speaks against the adoption of paragraph breaks as a reference criterion for segmentation.

Compared to paragraphs, a unit which has received considerable attention over the years (Becker, 1965; Berber Sardinha, 1993a; Crothers, 1979; Hoey, 1985; Hwang, 1989; Longacre, 1979; Paduceva, 1974; Rodgers, 1966), research into sections in written texts is much more scarce. While there is no study into decisions for inserting section boundaries comparable to Hoey's (1996) investigation of paragraphing, there is a body of research into the constitution of sections which suggests that they are motivated by certain linguistic characteristics. Biber and Finegan (1994), for example, found that different sections have different linguistic profiles. Swales (1990) identified a range of linguistic features which differentiate sections in research articles. And Berber Sardinha (1995a) noted that introductory sections differ from

other sections in business reports with respect to the distribution of the vocabulary of these texts. All these studies suggest that sections are not simply created by arbitrary decisions taken by writers; rather sections have a linguistic motivation. Despite having received even less attention than sections, chapters also seem to be linguistically motivated (Phillips, 1985). In this manner, sections and chapters seem to be equally good units to serve as a reference criterion. A feature which differentiates between them is their availability; since it would be less restrictive for data collection purposes to choose a unit which is found in a wider range of text types, and since sections seem to be found in more text types than chapters, sections are the best choice.

5.1.3 Filling the gaps

The research presented in this thesis is aimed at filling the gaps indicated above, namely the need for addressing how to deal with large numbers of longer texts, and the need for objectively assessing the analysis. In this chapter I begin to tackle these issues. In order to deal with larger amounts of data, the most logical solution is to make use of computers in the analysis. According to chapter 3, existing approaches to segmentation by computer are inadequate because they generally incorporate arbitrary measures which do not reflect the linguistic realisation of the texts. For example, Hearst (1985) substituted pseudo-sentences for real ones in computing similarity between paragraphs. Youmans (1991) monitored the variation in type-token ratios in even-sized word intervals regardless of clause or sentence boundaries. And Kozima (1993b) measured cohesion within intervals of a fixed length. Invariably, what these studies fail to recognize is the importance of showing how *messages* connect across the text (Eggins, 1994; Halliday, 1994; Hasan, 1984; Hoey, 1991b). Instead, what current computational approaches to

segmentation have shown is how *arbitrary portions* of text behave in texts.

There seems to be an agreement among researchers from various orientations as to the crucial role played by connections among clauses and sentences in creating texts. In other words, according to previous research in discourse, it is not loose words that create texts, rather it is the interrelation among larger units such as collocations (Phillips, 1985; Stubbs, 1996), clauses (Hasan, 1989, 1984), and sentences (Hoey, 1991b) which contribute to the 'Zusammenhang' (Lohmann, 1988) or 'hanging-togetherness' of texts. For instance, Phillips (1985) demonstrated how collocations intercollocate and in so doing create lexical networks which reflect the chapter divisions of textbooks. Hasan (1984) and later Parsons (1990, 1996) showed how clauses enter into cohesive harmony and how this relates to perceptions of coherence. And Hoey (1991b) revealed how sentences connect to one another meaningfully across long distances through the repetition of lexical items. All of these studies share the view that a fundamental pre-condition for analysing text constitution is to focus on the interrelationships among non-arbitrary units. Therefore, a major concern of the research reported here is to adhere to this position and incorporate meaningful rather than arbitrary units in a computer-assisted procedure for segmentation.

In sum, since there is no single computer-based segmentation method which does not introduce arbitrary measures, there is no ready-made method which can be used in the present study. It thus becomes necessary to develop a new segmentation procedure which can fill the above-mentioned gaps while at the same time being informed by research in discourse analysis. Marrying the objectivity of the computer to the rigours of the discourse analyst is not a trivial task. As Sparck Jones (1996, p.14) rightly observes:

> It is something of a caricature to see those engaged with computation as crass technocrats for whom the expression 'non-computational theory' is an oxymoron, and linguists as toffee-nosed snobs

> unwilling to inspect the rude mechanicals' cranks and levers, and huge chasm between the two. But there is a gap that deserves to be bridged because for linguists ...there is everything to be learnt from appreciating the distinctions between assumed, ideal, and real computation.

Ironically, while the adoption of computers makes it possible for more data to be handled, it also poses greater constraints on the selection of a suitable method of analysis. It seems more natural to adapt a discourse model for the computer than to take an existing segmentation algorithm and redesign it to make it acceptable according to discourse analytical criteria, because computer-based algorithms generally incorporate arbitrary decisions on how discourse operates. The logical course of action is to look in the field of discourse analysis for possible models which can be adaptable for the computer. According to chapter 2, discourse analytical models can be roughly divided into content-based and surface-based. Since in the latter the motivation for segmenting is to an appreciable extent provided by surface elements, surface-based approaches seem to be more suitable for the computer. The problem with surface-based approaches is that even those which seem to rely exclusively on discourse markers for segmenting texts, such as Longacre's (1983), seem inadequate to provide an unequivocal identification of the intended segments. As Darnton (1987, p.94) observes, 'it is the existence of the episode which establishes [the] function [of linguistic features] as episode markers, rather than the other way about'. Another problem with using surface features such as 'cue phrases' (Grosz et al., 1989, p.443) to segment texts is that to the extent that they are a closed set of expressions being searched for in the text, they represent a form of top-down processing of the data, and are therefore incompatible with the bottom-up orientation towards the data which is aimed for in this thesis.

Although there is no existing surface-feature discourse model which seems

adequate for computer-assisted analysis, there is a surface feature of language which is readily identifiable by computer: lexical cohesion. As the examples of previous studies reviewed in chapter 4 indicate, the computer is particularly well-suited for identifying lexical cohesion. A range of approaches exist which exploit this capability (e.g. Hearst, 1994b; Kozima and Furugori, 1993; Morris, 1988).

5.1.4 Beginning the investigation

One approach which has both key characteristics mentioned above, namely a focus on how messages are connected and a reliance on surface features, is Hoey's (1991b) model of lexical patterns in text. Major features of his model were described in section 4.4 above (see p.149 ff.). Hoey's (1991b) approach to lexical patterns in text was therefore chosen as the basic framework within which to start the computer-based investigation of segmentation.

It is important at this stage to spell out those aspects of the analytical model proposed by Hoey (1991b) which were implemented in the investigations reported in this thesis. Of the notions discussed by Hoey (1991b), the most central to the present investigation is *links*, or the repetition of a lexical item in two separate sentences. The kinds of link accounted for fully or in part in the analyses presented in this thesis are:

- Simple repetition between identical items (e.g. 'bear' and 'bear'): fully accounted for;
- Simple repetition between similar items (e.g. 'bear' and 'bears'): partly accounted for;
- Complex repetition (e.g. 'used' and 'user'): partly accounted for;
- Simple paraphrase (e.g. 'sedating' and 'tranquilized'): partly accounted for;

- Complex paraphrase (e.g. 'drug' and 'tranquilized'): partly accounted for;
- Superordinates and hyponyms (e.g. 'bears' and 'animals'): partly accounted for.

The identification of non-lexical repetition (substitution, co-reference, and ellipsis) was not implemented; more specifically, the repetition of following elements was not accounted for in the analyses:

- Third person personal pronouns;
- 'you' and 'we' within quotation marks;
- Demonstrative pronouns;
- 'One', as in 'the first one';
- 'Do', as in 'do it';
- Clausal 'so' and 'not' as in 'they said so', 'they said not';
- 'Other', 'another', 'the other', '(the) same';
- 'Different' and 'similar'.

Simple repetition between identical items was the only aspect of the model which was fully represented in the investigations reported in this and subsequent chapters because it is the least troublesome aspect to compute. The identification of the other kinds of repetition discussed in Hoey (1991b) was not implemented in full because of the state of the art in linguistic computing at the time the studies were conducted, and also because of the resources made available to the research project. The difficulty in recognizing certain kinds of links by computer is recognized by Hoey (1991b, p.74) himself: he limited his analyses to the identification of lexical links because these 'offer the possibility of identification by computational means'.

In addition to the notion of links, the only concept in the model proposed by Hoey (1991b) implemented in the analyses reported here is *bonds*, which

was utilised in pilot studies 1 and 2. Other aspects present in Hoey's (1991b) model, such as the classification of sentences as central and marginal, the identification of topic opening and topic closing sentences, and the contextual criterion for avoiding 'chance' lexical repetition were not implemented. Importantly, with respect to the latter aspect, Hoey (1991b, p.57) himself acknowledges that contextual questions 'may be valuable in manual analysis but they are really no use for automatic analysis'.

The details of the programs which implemented the identification of the links in the texts are presented below in section 5.2.2 on p.193ff and section 5.4.6 on p.247ff.

A problem with Hoey's (1991b) analytical framework is that it was designed to show how texts are integrated by lexical cohesion, rather than how they are segmented. There was no straightforward obvious way of implementing Hoey's (1991b) approach to lexical patterns as a segmentation method, and therefore a number of attempts were made on the way to the segmentation procedure adopted for the main study, which is presented in chapter 6. Hence, it was necessary to undertake preliminary research in order to estimate the plausibility of using Hoey's approach for text segmentation. This preliminary stage of the investigation comprised a series of pilot studies, each designed to address a specific issue related to segmentation by computer. Treating this phase of the research project as a set of pilot studies enabled me to develop the tools and the knowledge needed for segmenting texts without the pressure of having to meet the three criteria of extensive coverage, inductive orientation, and objective evaluation all at once.

Methodologically, the goals of the pilot study phase of the research were to develop fully computerized procedures to perform the three major stages in the research:

Computation of lexical cohesion Creation of a suitable computer pro-

gram and immediate application to the data;

Placement of boundaries Development of a methodology for placing boundaries and subsequent implementation on the computer;

Evaluation of performance Development of a methodology for matching boundaries and subsequent implementation on the computer.

5.2 Pilot study 1

The first issue that I needed to tackle was to see whether Hoey's approach to lexical patterns worked for segmenting texts. To this end, I decided to try out Hoey's (1991b) bonding as a pilot study[1]. It seemed best to experiment with one single text, in order to see whether the results would warrant the application of the method to a collection of texts. Thus, at this first stage, the criterion of extensive orientation was not a priority. Nevertheless, the other two criteria were adhered to, namely inductive orientation and objective evaluation.

This section describes the first pilot study conducted as part of the research project which was set up to investigate the automatic identification of segments in written texts. Operationally, there were three distinct stages in the research: first, computation of lexical cohesion, followed by the placement of segment boundaries, and finally, evaluation of performance.

Following the decision to make use of Hoey's (1991b) approach to lexical patterns as a starting point for the investigation of segmentation, the next issue was that of how to explore his system of analysis so that instead of showing how texts are integrated, it indicated how texts are segmented. Integration and segmentation can in fact be considered to be two sides of

[1] This study was originally presented as a Postgraduate Seminar at the University of Liverpool on 15th January 1993 under the title 'Lexical cohesion in business reports'.

the same phenomenon (Goutsos, 1996a; Bestgen and Costermans, 1997, pp. 204-205), and therefore it should be possible to explore the segmentational potential of Hoey's approach.

5.2.1 Data

The text analysed in this study was an 83-sentence business report written in English for a multi-national telephone company that operates in Brazil. The choice of this particular text was motivated by the fact that at the time it was conducted, this piece of research formed part of a larger project[2] whose general objective was the description of business discourse. Business reports were suitable for the task of segmentation since they contain a large number of sections, which conformed to the decision taken to use section divisions as a reference criterion for the objective evaluation of the segmentation. In addition, the fact that all business reports had numerous section divisions suggested that sectioning was part of the generic make-up of this text type. Information which was felt to be of private nature was modified, including for example the name of the company which was changed to 'ACME'.

5.2.2 Automatic computation of lexical cohesion

The links between all pairs of sentences in the text were computed at various bonding thresholds. A bonding threshold is a criterial number of links for considering two sentences as bonded. For example, when the threshold is three links, only those sentences sharing three links or more are included; when the threshold is four links, only those sentences sharing four links or more are included, and so on.

Following Hoey (1991b), the lowest cut-off point was three, so pairs of

[2] DIRECT, or 'Development of International Research in English for Commerce and Technology'.

sentences sharing two links or fewer between them do not feature in the analysis. The links and bonds for the target text were computed by a program[3] developed specifically for the text, called `links` for convenience. The `links` program computed the links shared by all pairs of sentences in the text.

The `links` program was relatively primitive in that it did not allow for the selection of sentence pairs that satisfy a particular bonding level. Therefore, the selection of sentence pairs for the various levels of bonding needed in the analysis was carried out interactively using an ordinary wordprocessor. Despite this limitation, the first aim of the study, namely the computation of the lexical cohesion by computer, was achieved.

Algorithm

This section presents an outline of how `links` works[4]. The basic structure of the program is very simple: read an index listing the words and the sentences in which they occur, process the index by counting the number of repeated words shared by pairs of sentences, and output the count of links into a plain ASCII file. The index which `links` reads in must be prepared beforehand, either manually or using a wordprocessor. In the case of the text analysed here, the index was prepared using the index facility in WordPerfect 5 for DOS. The index has the following format:

```
word_1 sent_1, sent_2*
word_2 sent_1, sent_3*
word_3 sent_4*
```

The star at the end of each entry is used to tell `links` to stop reading that entry and move on the next. This character is needed because some entries

[3] I am grateful to Dr Mike Scott and K Wang for their assistance in developing computer programs at this stage of the research.

[4] Further information on the program can be obtained from the author by writing to: R Paracatu 357 apto 52, 04302-020 São Paulo SP, Brazil

stretched over a number of lines. The user may make all sorts of changes to the index to improve the detection of links. For example, given the following index:

```
cat 1, 2*
cats 1, 3*
mat 4*
```

it would be desirable to merge the first two entries into a single one to reflect the fact that the entries refer to the same lemma ({CAT}):

```
cat 1, 2, 3*
mat 4*
```

Once the user is satisfied with the index, he/she can run it through `links`. The program reads in the index, stores each entry in memory, and calculates the frequency of each word by counting the number of sentences listed in each entry. The frequency count for each word is used in the next step of processing, when `links` deletes those words which occurred only once in the text from the memory. The next step involves the actual computation of the number of links. To compute the links shared by sentences, `links` first builds a record for each pair of sentences in the index, counts the number of links, and lists the words in each pair; for instance, the following would be a list of the records of sentences from the index presented above:

```
1  2  cat
1  3  cat
```

The entry for `mat` does not contribute with a link because it appears in one sentence only (sentence 4).

The results are then output to a plain ASCII file in the same format as the preceding example.

`Links` is a program whose performance in terms of detection of links depends entirely on the information in the index supplied by the user. The program does not have access to the actual text on which the index is based, and so it cannot deal with any aspect of the text that is not reflected in the index. The careful construction and editing of the index is essential to assure that a range of different types of links is detected, and that only lexical words enter into links. The merging of index entries is an essential step in ensuring that a range of different types of links is detected, otherwise only simple repetition (Hoey, 1991b) will be accounted for.

In the analyses presented here, the index was edited manually and entries were merged to provide some sort of lemmatisation of the words in the text.

5.2.3 Analysing the matrix

The decision was taken to start the investigation by examining the *matrix* of repetitions for possible features which would suggest ways of segmenting a text. As explained in section 4.4.7 on page 159, a matrix for Hoey (1991b) is a diagram where the links between pairs of sentences are recorded. In his approach, a matrix is a triangle-shaped 'table' formed by rows and columns corresponding to individual sentences of the text. The main features of a matrix are its leading diagonal, which indicates adjacency, and the reference numbers down the left-hand side and down the diagonal, which indicate the coordinates for each pair of sentences. These features are illustrated in figure 5.1 on page 197. A general principle in reading matrices is that the further away from the diagonal two sentences are, the more distant they are from each other. For example, the pair formed by adjacent sentences 1 and 2 occurs on the diagonal, as shown in figure 5.1; by contrast, the pair formed

by sentences 1 and 6 is not adjacent (there are five sentences between them), and therefore it appears away from the matrix diagonal.

The `links` program output the links between sentences not as a matrix, but as a list. As a result, the lists had to be reformatted as matrices to serve the purposes of this study. This was done in a wordprocessor by means of simple recorded macros which read the list output and rewrote the information as a matrix. Considerable amounts of editing were needed before the appearance of the matrix was considered acceptable.

The matrices generated for the study are reproduced in appendix 1 (p.443 ff.). Each matrix represents a bonding *threshold*, or a cut-off point. The dots in the matrix indicate bonded sentences. Only those sentences which have at least the number of links for a particular threshold are featured in that particular matrix; so, for instance, the matrix for the 3-link threshold includes sentences bonded by 3 or more links, the matrix for the 4-link threshold includes sentences bonded by 4 or more links, and so on.

It was felt that comparing the matrices visually for salient features would be a legitimate place to start the investigation. By comparing the six matrices obtained for each bond threshold, a few trends became apparent. First, as expected, there was a decrease in the number of bonded sentences as the threshold of links increased. Second, the place in the matrix of those sentence pairs which were being eliminated as the threshold increased was not random. Rather, the sentence pairs which tended to disappear were positioned further from the diagonal, while those sentence pairs which remained on the matrix tended to be near the diagonal. In other words, the distribution of bonds seemed to concentrate near the diagonal. This was considered an interesting trend worth exploring for segmentation purposes.

The next issue was how to use the fact that bonded sentences tended to concentrate along the main diagonal of the matrices as a tool for segmenta-

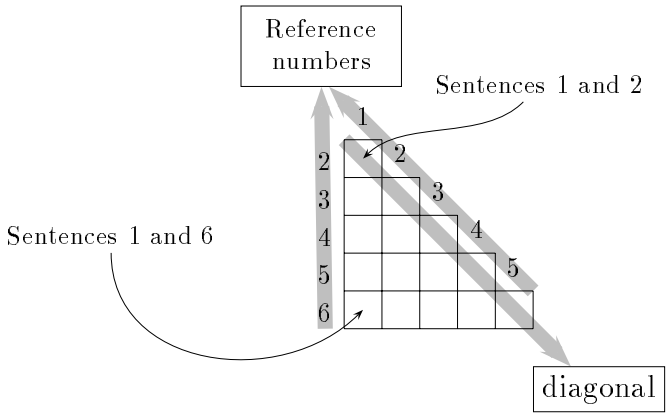


Figure 5.1: Layout of a matrix

tion. By further examining the areas near the diagonal, I noticed that the distribution of bonded sentences in that area was not even. Rather, the bonded sentences along the diagonal seemed to form *clusters*. The clusters were less noticeable at the three-link threshold, because there were more numbers spread across the matrix. This made it more difficult for a pattern to be perceived. Nevertheless, once I had become aware of those patterns, it was possible to identify them on the matrix for the three-link threshold as well. Thus, the reason why the cluster pattern had not been perceived on the three-link threshold matrix was that there was too much information on the matrix.

A simple method which could be applied in order to reduce the amount of information on the matrix is to exclude part of it. A question which presented itself at this stage was which matrix should have its bonding information reduced. What was needed was a matrix which was not in itself a reduced matrix, and therefore the best candidate was the matrix for the three-link threshold. The next question that arose was what part of the matrix should be excluded. Following the observation that the distribution of bonds tended to accumulate near the diagonal of the matrix, the most logical answer was

to exclude those areas away from the diagonal. A further question presented itself at this stage, namely which criteria should be used in order to distinguish between 'near' and 'not near' the diagonal. Clusters were considered to be near the diagonal when they had adjacent elements, that is, adjacent bonded sentences. The criterion that was used was to set a distance off the diagonal which would enable me to capture major clusters of this kind on the matrix. In other words, the distance was just wide enough to include those clusters which presented themselves as major features of the matrix.

Figure 5.2 on the next page shows a sample of clusters that are visible on the matrix. Clusters A and B have adjacent members, and so they qualified as candidates for inclusion in a restricted matrix. Cluster C, on the other hand, did not have any adjacent members, hence its location several sentences away from the diagonal, and was therefore excluded from the restricted matrix. Clusters A and B should remain intact in a reduced matrix. Cluster B was the largest of the two, and therefore its edge could serve as the point away from which the other bonds should be excluded. In view of this, a line could be drawn parallel to the matrix diagonal just wide enough to allow for the inclusion of cluster B. For ease of reference, the line was called 'exclusion line'. Figure 5.3 on page 200 shows the exclusion line applied to the three-link threshold matrix.

Having reduced the information on the matrix while at the same time preserving important information regarding the clusters of bonds, the next step was to find ways in which to use the bonded sentences within the remaining strip of the matrix for segmenting the text. By examining the reduced matrix, it became apparent that even though possible segmenting places could be inserted immediately before and after clusters A and B in figure 5.2 on the next page, most of the matrix could not be segmented in this way since

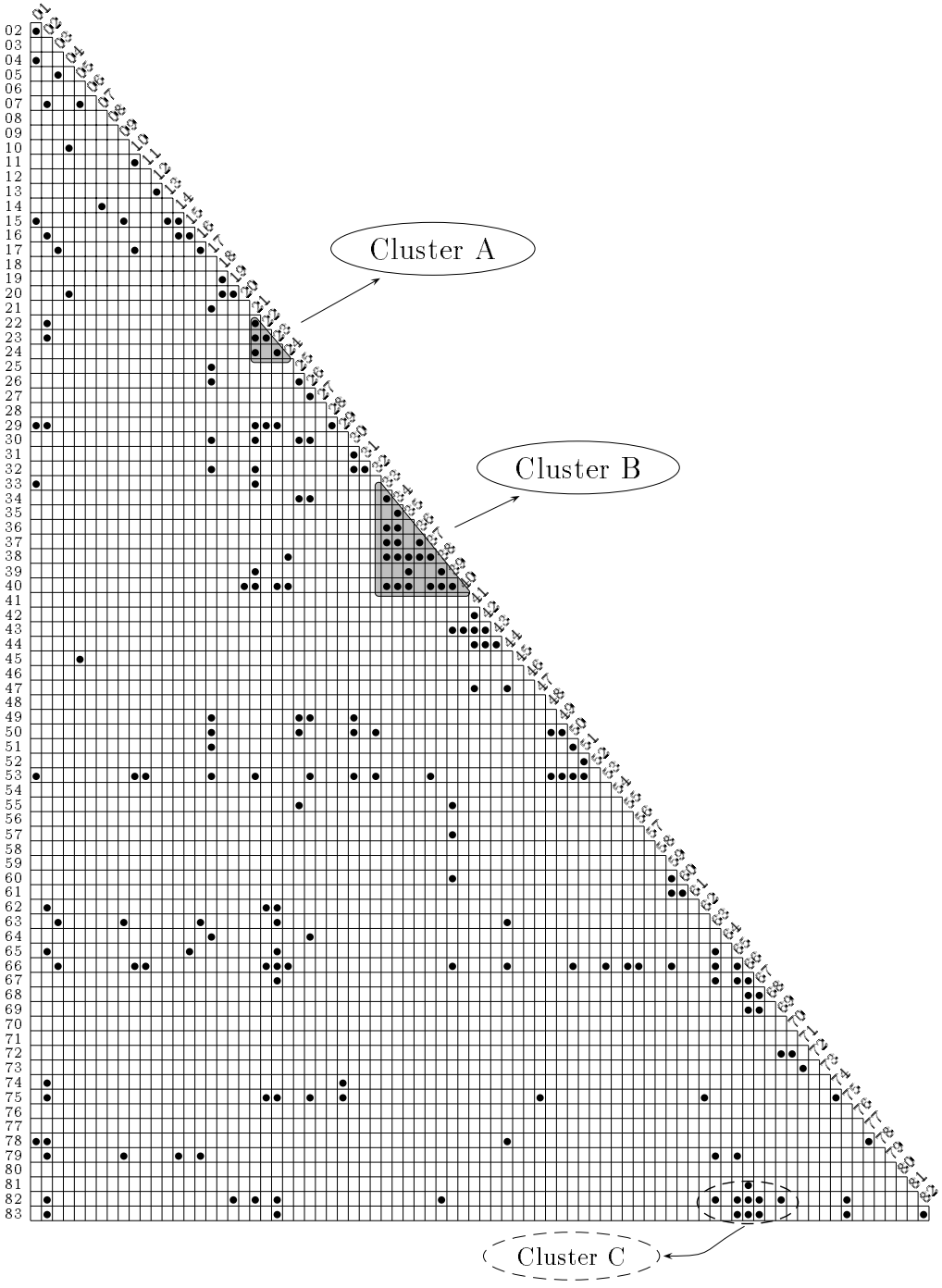


Figure 5.2: Some noticeable clusters in 3-link matrix

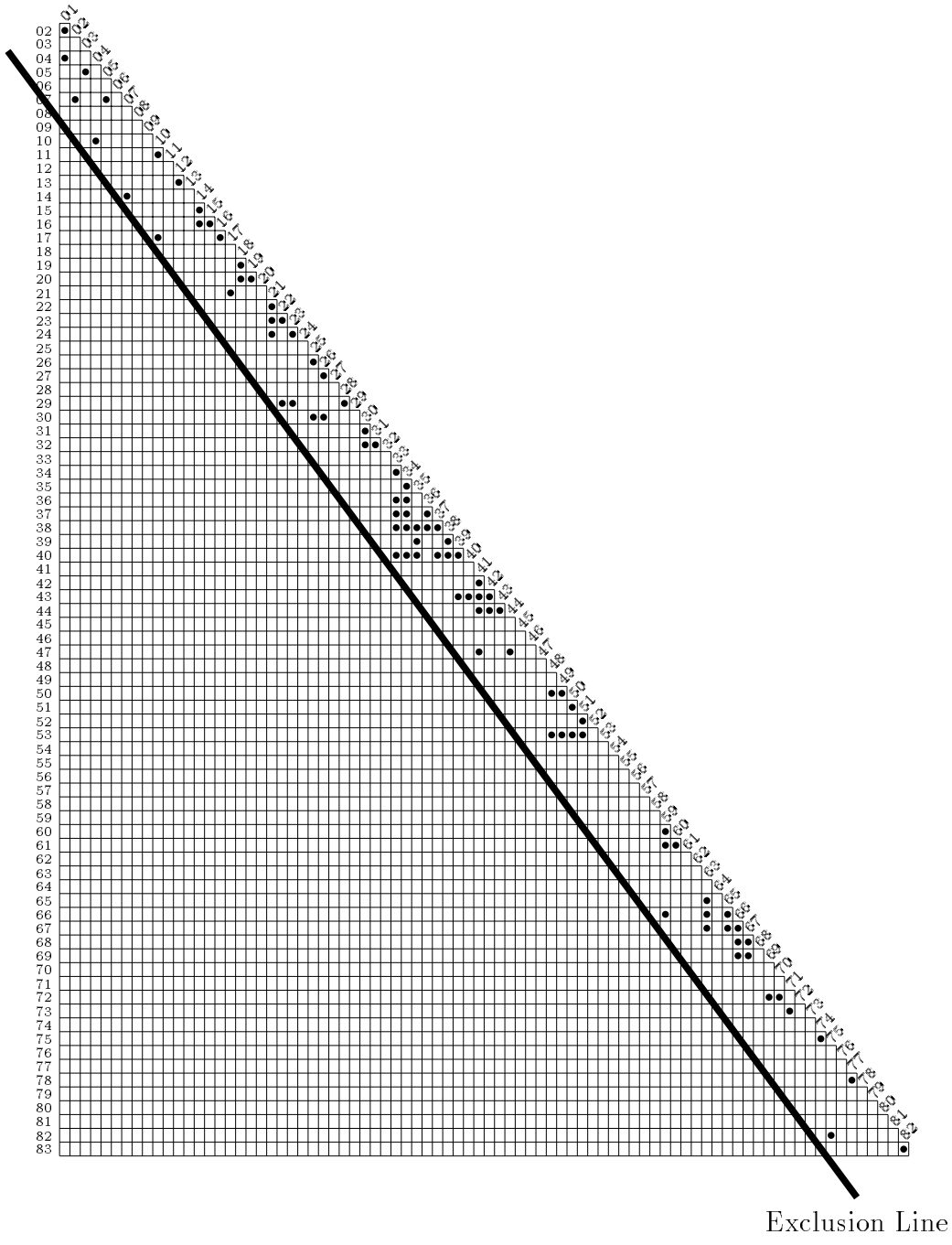


Figure 5.3: Matrix with exclusion line

the remaining bonds did not form clusters as compact as clusters A and B.

In a matrix, pairs of sentences are represented as an intersection between two coordinates, as figure 5.1 on page 197 shows. This makes the matrix ideal for showing the spread of interconnections among sentences. Because of this, the matrix is less suited for showing segmenting places, since the breaks between groups of interconnected sentences are less apparent. This prompted the decision to make use of a different kind of diagram, one which could display possible breaks among the bonded sentences. The diagram developed for this purpose was called *connection chart*, and it consisted of writing down the number of sentences vertically in a single column and then connecting the bonded sentences by a loop. Figure 5.4 shows how a matrix and a connection chart encode the same bonding information. For example, the bond between sentences 1 and 2 is displayed in the matrix (on the left-hand side of the figure) by a dot at the intersection between the coordinates for sentences 1 and 2, whereas on the connection chart (on the right-hand side of the figure), the same bond is represented by an arch connecting the reference numbers for sentences 1 and 2. All of the bonds as shown by the matrix are displayed on the connection chart, and the exact location of a sample of the bonds of the matrix on the chart is indicated by arrows.

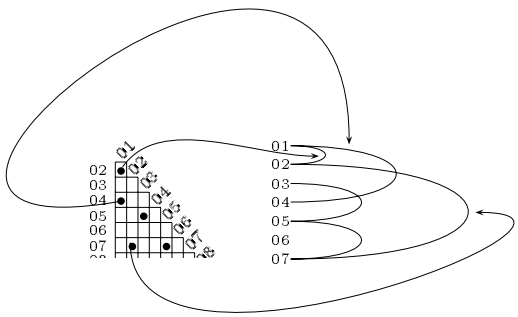


Figure 5.4: Relationship between matrix (on the left) and connection chart (on the right) (a sample of the connections and their corresponding matrix coordinates are indicated by arrows

The information in the matrix reduced by the exclusion line (as shown in figure 5.3) was then transferred to a connection chart. The chart was then segmented by looking for breaks among the connections. As a result, four segments were found in the chart, as illustrated in figure 5.5 on the following page.

The next step was to contrast the segmentation of the text to the original section divisions. Figure 5.6 on page 204 displays the areas corresponding to the segments and the section divisions.

The performance of the procedure can now be estimated (see section 3.3.4 on page 99 for a discussion on 'precision' and 'recall'). There were six matches between segment and section boundaries: sentences 1, 47, 48, 73, 74, and 83. In all, eight segment boundaries were inserted (two for each segment): sentences 1, 47, 48, 53, 59, 73, 74, and 83. This yields a precision rate of 75% (six matches divided by eight segment boundaries). In turn, the recall rate is 22.2%, since there were six matches and twenty-seven section boundaries (one for section 1, and two for each of the remaining thirteen sections).

It is possible to argue that since the first and last sentences of the text are by definition boundaries, they should be excluded from the computation of performance rates. In this case, the precision rate would then be 66.7% (four matches divided by six boundaries), and the recall rate 15.4% (four matches divided by twenty-six section boundaries). Nevertheless, the segmentation procedure described here did not assume boundaries at these sentences by default, and so in another text the first and last sentences might not be picked as segment boundaries, and segment 1 might have begun at sentence 2 or 3. This is possible because the segmentation procedure was not designed to assign every sentence to a segment, as the gap between segments 2 and 3 (sentences 54 to 58 in figure 5.6 on page 204) shows. In other words, it

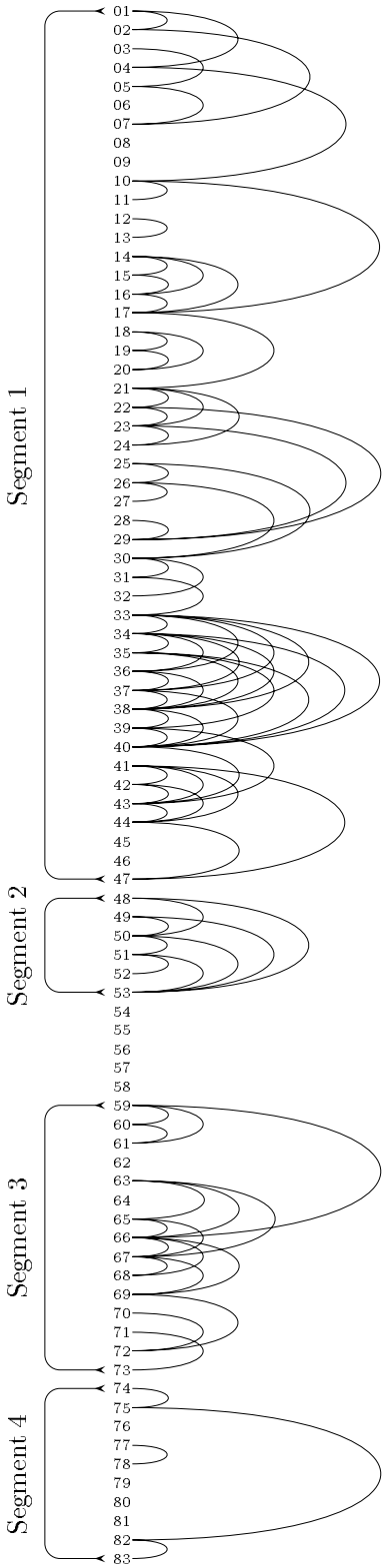


Figure 5.5: Segmentation of the text on connection chart

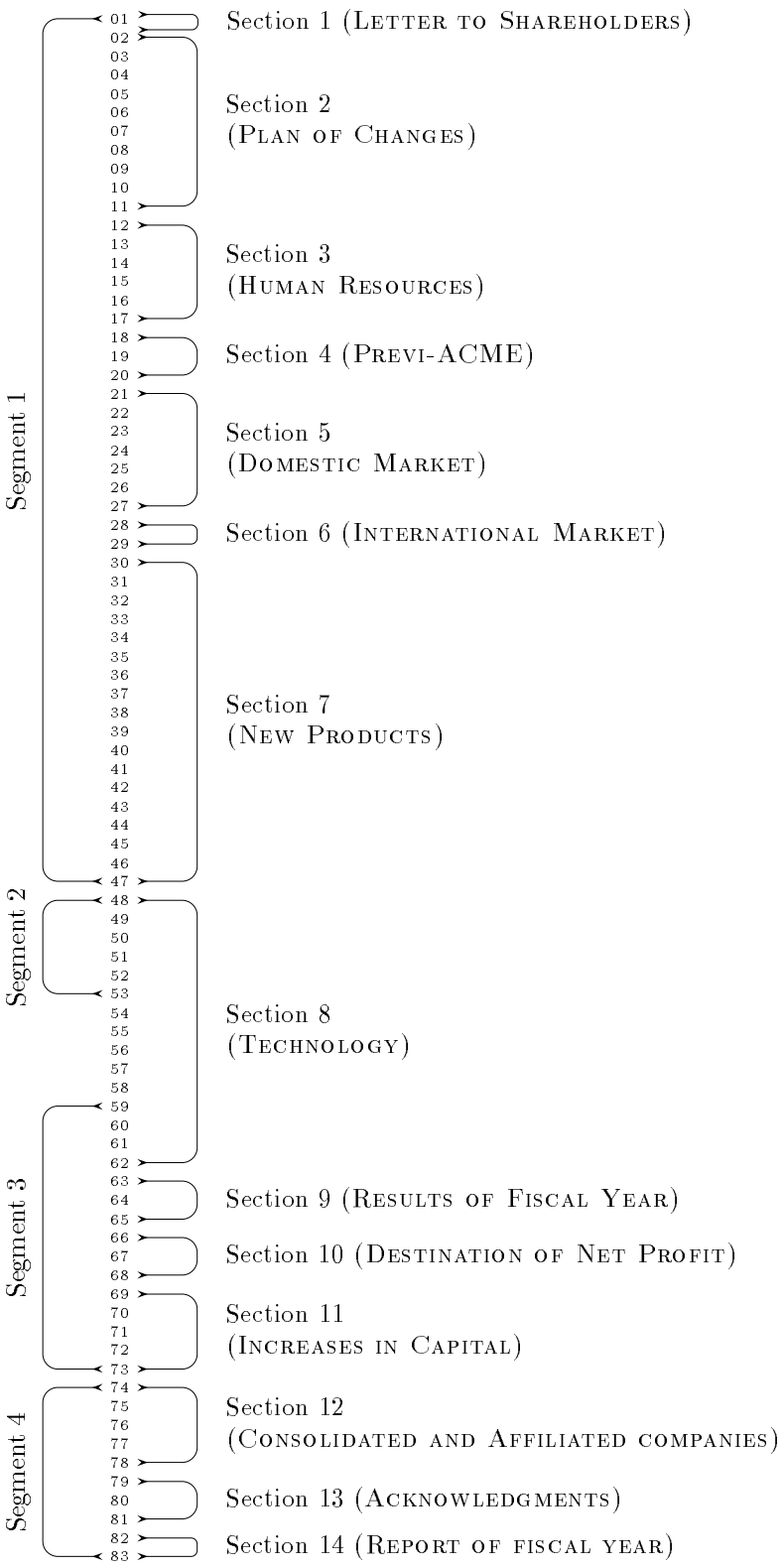


Figure 5.6: Segmentation and section divisions of the text

would not be fair to deduct these two matches since they are not automatic segment boundaries. Thus, the matches at the first and last sentences were not excluded from the computation of performance, and the performance rates for the procedure described here are 75% precision and 22.2% recall. .

5.2.4 Conclusions and future work

The main question which the present pilot study addressed was whether Hoey's (1991b) method of analysis of lexical patterns could in principle be adapted for segmenting texts. The answer is affirmative, since it was possible for the matrix to be rendered into a linear connection chart which was then used as an instrument for observing likely segmenting places in the text.

This pilot study attempted to achieve the three goals which were set for the pilot study phase of the research project. The first goal of the pilot studies referred to the computation of lexical cohesion. A computer program called `links` was developed which identified the cohesive links in the text. The second goal was to develop a methodology for placing boundaries which could be subsequently implemented on the computer. In the present pilot study, a method was developed for placing boundaries through the application of an 'exclusion line' on the matrix and the later rendition of the matrix into a linear 'connection chart'. The exclusion line and the connection chart constituted intermediate steps between the construction of the matrix and the segmentation, and unlike the building of the matrix itself, were not automated. As a result, the procedure was not fully automated, and therefore this goal was not attained. The last goal of the pilot studies was to develop a methodology for matching section and segment boundaries automatically. In the present pilot study, this was accomplished manually, by aligning segment and section boundaries on the connection chart and visually checking for matches. Thus, the procedure needed to be made automatable in order

for the goal to be attained.

Two areas deserved further work, both of which had to do with the introduction of intermediate steps in the utilization of matrices for segmentation. The first related to the use of the exclusion line, which was employed as a means of reducing the number of interconnections between sentences. The second referred to the use of the connection chart, which was introduced as a means of rendering the matrix in a format which was more revealing of breaks. These instruments were introduced for the sake of manually segmenting the matrix. It seemed as though these instruments were superfluous in an automated procedure. Thus, the decision was taken to try to develop a more efficient procedure which did not rely on intermediate instruments such as exclusion lines and connection charts for segmenting a matrix. This was the main motivation for pilot study 2, which is described in the following section.

5.3 Pilot study 2

In the previous pilot study a procedure was presented for dividing a matrix into segments. The procedure was based on the application of Hoey's (1991b) analysis of lexical patterns in text. Pilot study 1 concluded that the procedure seemed to be adaptable to segmenting texts since it was possible to segment a text based on observing the distribution of bonds in a matrix. Nevertheless, there were problems in the actual implementation of the segmentation procedure, since some devices, namely the 'exclusion line' and the 'connection chart', were introduced in order to segment the text manually, and therefore should have no place in a fully automated procedure. Hence, there was a need for a further pilot study in which a new segmentation procedure was developed which did not include manual segmentation devices.

The research undertaken as part of this pilot study is reported in the present section[5].

The data used in this study are the same as for the previous pilot study (see section 5.2.1 on page 192).

5.3.1 Guidelines for alternative segmentation

In pursuing the goal of developing a new procedure for segmenting the matrix it would be advantageous not to restrict the valid area of the matrix as was done previously in pilot 1. Restricting the area of the matrix by applying the exclusion line had an influence on which bonds made their way into the connection chart, and ultimately on the segmentation of the text, since by moving the exclusion line, another set of bonds would have been picked up. In addition, although locating bond clusters in the matrix was crucial for determining where to place the exclusion line on the matrix, there was no formal definition of bond cluster. One could have found several bond clusters in the matrix, and therefore the exclusion line could have been drawn in several places, each of which would have had a different effect on the segmentation. In other words, the notion of bond cluster appeared promising, but it lacked a more precise definition.

In view of these disadvantages, the following desiderata were postulated for a new segmentation procedure:

1. The segmentation procedure should account for the whole matrix
2. The segmentation procedure should not depend on manual segmentation devices

As in the previous pilot, the best strategy for implementing these requirements was to observe the internal shape of the matrix. Although it was

[5] A modified version of this study appeared in Berber Sardinha (1993b).

necessary that the segmentation be based on the entire matrix, it appeared that the best place to begin was to observe the area near the matrix diagonal. As in the previous pilot, clusters of links near the diagonal seemed normally good candidates for segments, especially those which had a triangular shape. Since the matrix is a right triangle (it contains an angle of 90°), it appeared more appropriate to restrict the search to right triangles only. Clusters shaped like right triangles were ideal because they had bonded sentences at strategic points, namely a bond between the first and second sentences, another bond between the next-to-last and the last sentences, and a third bond between the first and last sentences. These points correspond to the three corners of a triangle, and bonds occurring at these places would have the effect of 'tying together' the cluster. Of course, other bonds may occur within the space formed by these three points, but they were not criterial for the delimitation of the triangle-shaped cluster.

A possible tactic was therefore to start with an adjacent bonded pair of sentences on the very edge of the leading diagonal and look down from it to see whether there were any triangular clusters around that area of the matrix. Starting with the very first adjacent bonded pair, namely sentences 1 and 2 (see appendix 1 on p.444), and moving down from there it was not possible to find any triangular clumps near sentences 1 and 2 since the next adjacent bonded pair was sentences 10 and 11. On the other hand, by

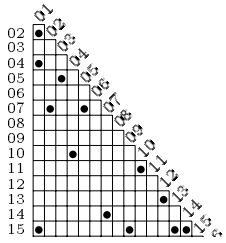


Figure 5.7: Triangle-shaped cluster

drawing an imaginary line that went from the bond formed by sentences 1 and 2, then to the bond between sentences 1 and 15, and finally to the pair between sentences 14 and 15, a triangle-shaped cluster could be outlined. This cluster occupied the top tip of the matrix, and is shown in figure 5.7 on the preceding page. As the figure indicates, a triangular cluster could be seen as a miniature matrix. To reflect their relationship with the matrix, triangular clusters were referred to as *matrix triangles*.

5.3.2 Matrix triangles

The guidelines for identifying matrix triangles are as follows. The location of triangles depends on the identification of three *handles*, one for each tip of the triangle. Handle 1 will be the first adjacent bonded pair of sentences in the text. A provisional handle 2 will be any other bonded pair of sentences located on the diagonal. Handle 3 will be that pair of bonded sentences which is located at the intersection of handles 1 and 2, that is, directly below handle 1 and directly to the left of handle 2. Once a triangle has been located, no other triangle can be superimposed onto it, nor can other triangles be found within it. However, handle 2 of a demarcated triangle can become the starting point (handle 1) for another triangle. Figure 5.8 illustrates these possibilities; the diagram on the left of the figure shows two

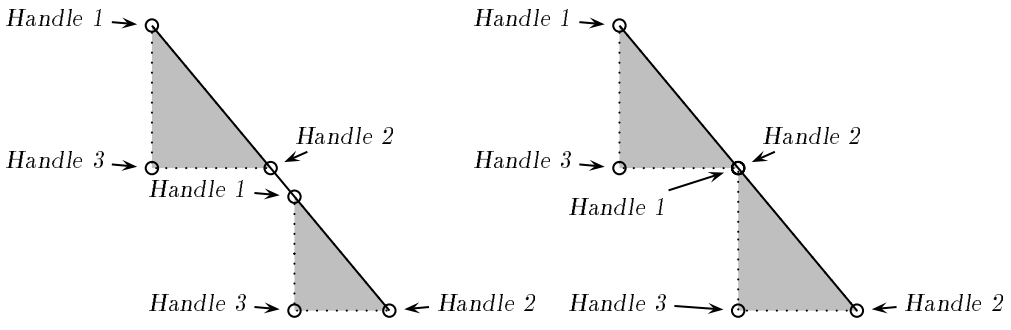


Figure 5.8: Triangle handles

triangle with a gap between them, while the diagram on the right depicts a situation where two triangles share a handle.

The application of these guidelines resulted in the demarcation of segments that were at least two sentences long, which satisfied the working definition of segment introduced in section 1.6 (p.16) of the introductory chapter. Furthermore, these guidelines ensured that the textual areas corresponding to matrix triangles produced segments that were contiguous; this was desirable because the pre-existing sections in the text were also contiguous, and therefore the comparison between segments and section divisions would be more straightforward.

5.3.3 Segmentation

The scheme for identification of triangles as described in the previous section was applied to the whole matrix of the text. This resulted in 8 triangles being identified. These are displayed in figure 5.9 on page 212.

As each triangle represents a segment, the matrix triangles technique yielded 8 segments. The distribution of the segments in the text is shown in figure 5.10 (p. 213). In the figure, the sentences of the text are listed vertically, and the segments corresponding to each matrix triangle are represented by a loop connecting the first and last sentence of the segment.

5.3.4 Performance

The segmentation of the text was compared to the division of the text in sections. Figure 5.11 on page 214 shows where the segments and section divisions occurred in the text. The list of numbers in the centre of the figure indicates the sentences in the text. The loops to the left of the sentence numbers show where each segment begins and ends, and the loops to the right indicate where the section divisions start and finish. As figure 5.11

shows, there were eight matches between segment and section boundaries, namely sentences 1, 18, 20, 21, 29, 30, 66, and 83. This yielded a recall rate of 29.6%, or eight matches out of twenty-seven section boundaries, and a precision rate of 50%, or eight matches out of sixteen segment boundaries.

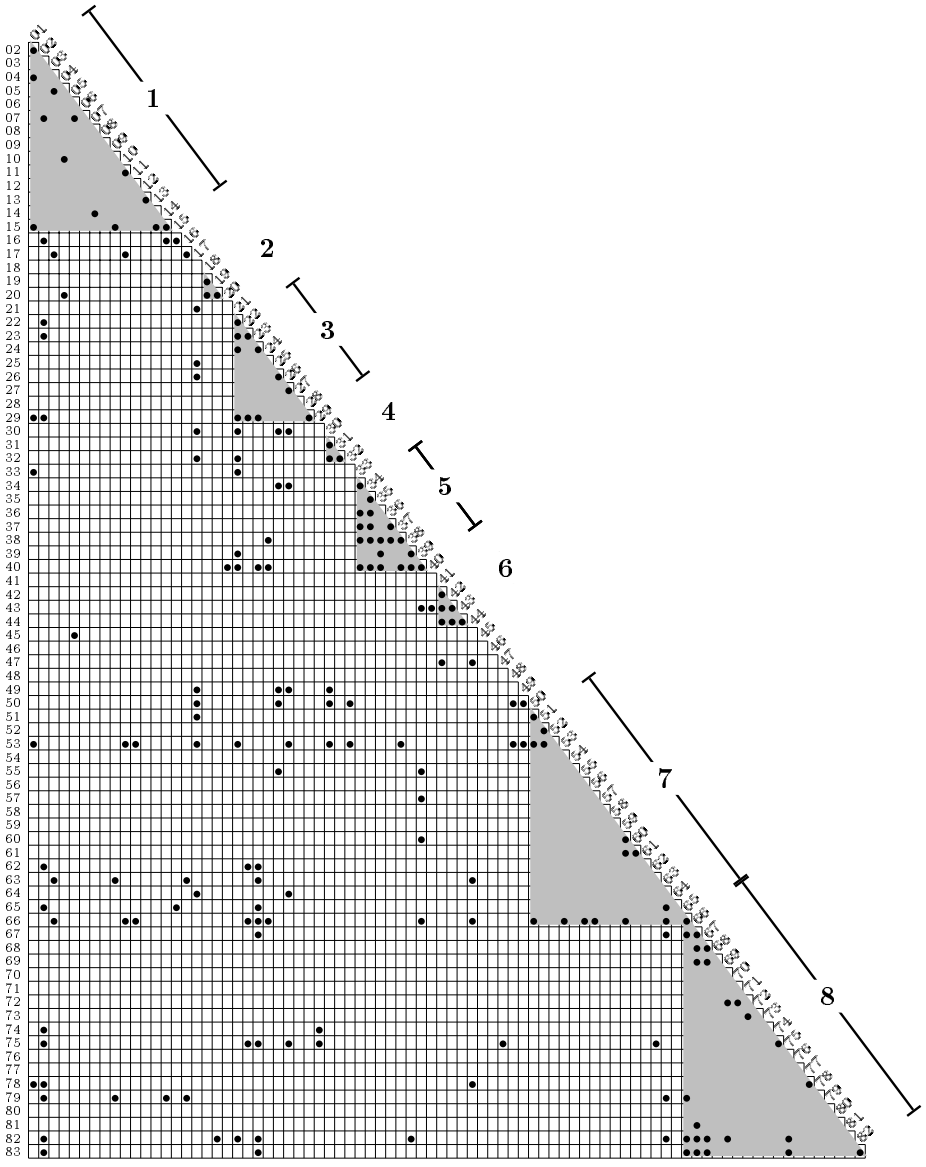


Figure 5.9: Location of matrix triangles

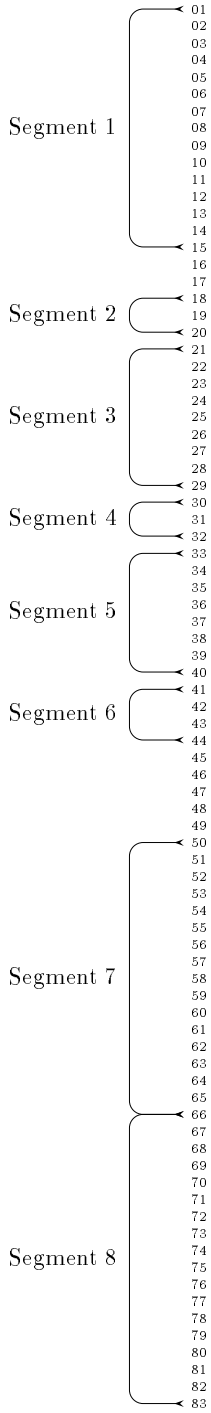


Figure 5.10: Segmentation of the text

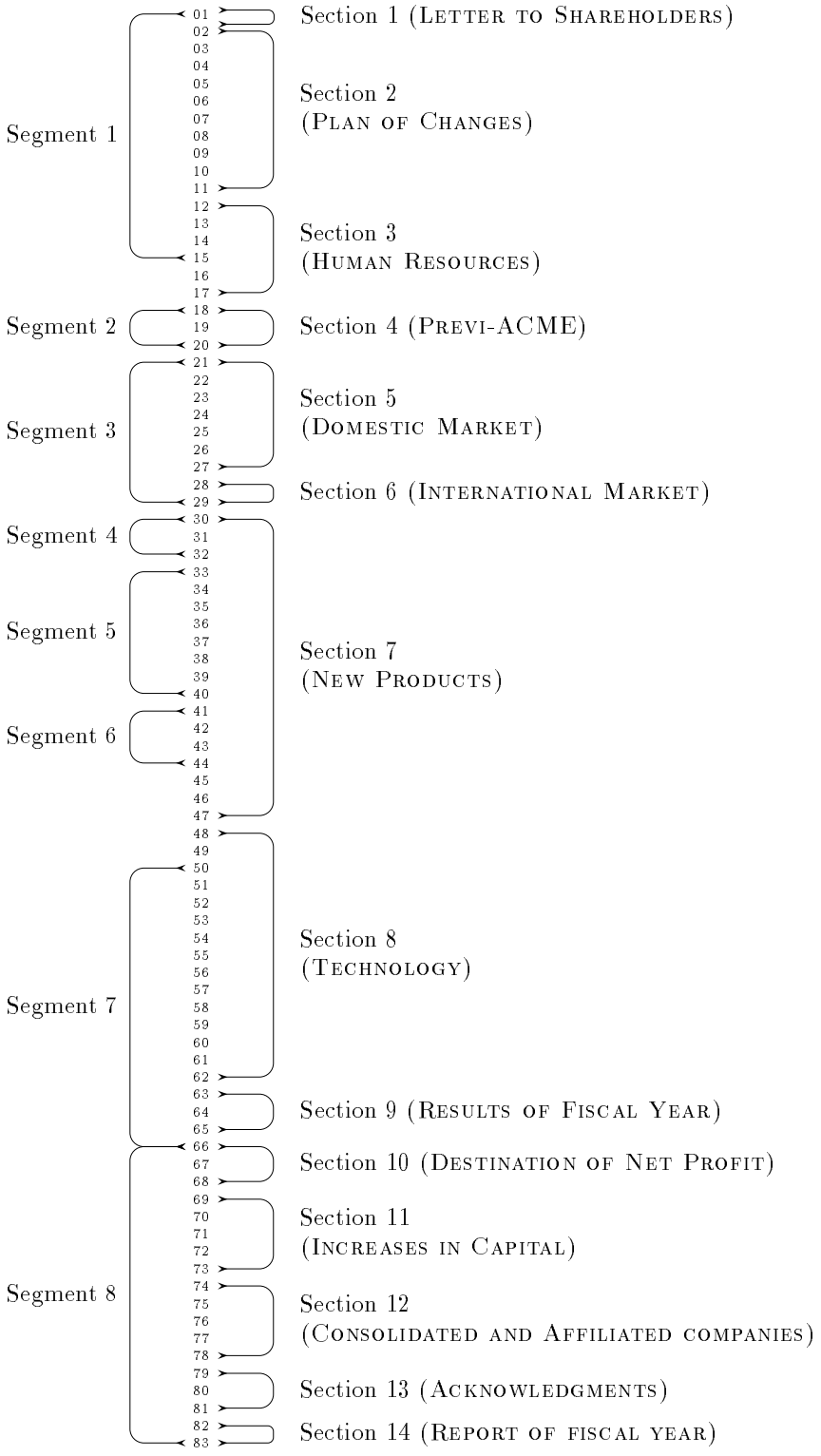


Figure 5.11: Comparison of segments and sections

5.3.5 Comparison with other procedures

The question arose of how the performance figures for pilot study 2 compared against the performance of pilot study 1. Figure 5.12 presents recall and precision rates for both pilot studies. A comparison of the performance of pilot studies 1 and 2 shows that whereas the exclusion line technique (pilot study 1) was much better at proposing true boundaries (precision), the matrix triangle technique (pilot study 2) worked slightly better at recovering more of the existing section boundaries (recall). The advantage of the exclusion line technique in terms of precision has to do with the fewer number of boundaries it placed (four against eight). By the same token, the larger number of boundaries inserted by the matrix triangle technique (sixteen against eight) increased its chances of recovering more of the existing section boundaries, and therefore it achieved a higher recall rate.

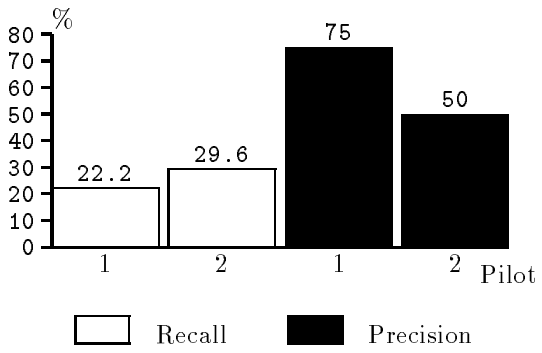


Figure 5.12: Performance of pilot studies 1 and 2

Another question that arose was how the performance of pilot studies 1 and 2 compared against the performance reported by other segmentation techniques reviewed in chapter 3. Figure 5.13 on the following page presents performance levels for the two pilot studies and for three other segmentation procedures, which are explained in what follows. 'Hearst' refers to 'TextTiling' (Hearst, 1993, 1994b,a; Hearst and Plaunt, 1993), a procedure discussed in detail in chapter 3 (see section 3.5, pp.109ff). The figures for

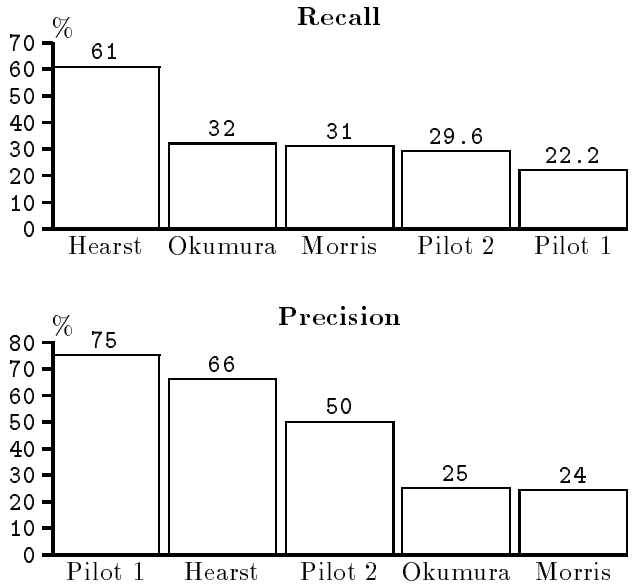


Figure 5.13: Comparison of performance with other procedures

TextTiling are the average values reported in Hearst (1994a, p.32) for the 'blocks' algorithm. 'Okumura' stands for the procedure presented in Okumura and Honda (1994), and the figures are the average values quoted by the authors. 'Morris' refers to the lexical chain procedure presented in Morris (1988) and Morris and Hirst (1991). The values presented for their procedure were calculated especially for the purposes of this comparison since no recall or precision rates as such are given in their studies. The recall rate was obtained by dividing the total of 'exact matches' (on p.99 of Morris, 1988) by the total of 'intention ranges' (quoted separately for each text in several places in (Morris, 1988). Precision rates, in turn, were arrived at by dividing the total matches by the total of chains. The total number of chains includes the possible subdivisions of individual chains; so, for example, text 1 of Morris (1988) has 11 chains: 1, 2.1, 2.2, 2.3, 3, 4, 5, 6, 7, 8, and 9[6].

As figure 5.13 shows, the procedure which presents the best recall rate is Hearst's TextTiling, with 61%. Pilot study 2 is practically tied in second

[6] The individual counts are:

place with Okumura and Morris, all hovering about the 30% mark. Pilot study 1 is the lowest scoring procedure in terms of recall. With respect to precision, it is pilot study 1 which achieved the best rate, with 75%, followed by Hearst, with 66%, and pilot study 2, with 50%. Okumura and Morris achieved considerably lower marks.

This comparison suggests that the performance of the procedures developed for this investigation so far cannot be considered disappointing. As pointed out above, Pilot study 1 achieved the highest precision rate of all, and Pilot study 2 was practically tied at second place in terms of recall. Admittedly, only one text has been segmented so far by any one of the pilot segmentation procedures, so their performance figures must be regarded as preliminary, though promising.

5.3.6 Conclusion

The two goals set for the present pilot study were that the new segmentation technique should account for the whole matrix, and that the segmentation procedure should not depend on manual segmentation devices. The first goal was attained in that all of the bonds in the matrix, regardless of their distance from the diagonal, were included. It is possible to consider the second goal to have been attained as well, since the manual segmentation devices used in pilot study 1 were not employed in pilot study 2. The rules for locating matrix triangles are in principle automatable, and therefore could serve as

Text	(A) Text boundaries: Intentional ranges	(B) Segment boundaries: Chains	(C) Exact Matches	(C/A) Recall	(C/B) Precision
1	13	11	3	23.1%	27.3%
2	19	34	8	42.1%	23.5%
3	9	9	3	33.3%	33.3%
4	13	15	5	40.2%	33.3%
5	10	16	1	10.0%	6.3%
Total	64	85	20	31.3%	23.5%

the basis for a fully computerized segmentation procedure.

5.3.7 Future work

Two improvements can be recommended at this stage. First, matrices should be generated automatically in full. This was suggested before in pilot study 1 but it was an issue which had not been tackled at this point. This prevented more texts from being analysed, which had long-term consequences for the claims being made here in relation to the performance of the segmentation techniques.

Second, the segmentation itself needed to be done automatically. So far, the actual segments had been located by eye. Clearly, this is unacceptable in the long run as it hinders the application of the technique to a collection of texts. Admittedly, progress had been made in this study by eliminating the need for advance delimitation of target segments, which led to the specification of objective criteria for locating the segments on the matrix (via the triangle 'handles'). These objective specifications might arguably find easier implementation on the computer. The goals of subsequent pilot studies had then to include providing fully automatic segmentation.

5.4 Pilot study 3

In general terms, pilot study 2 concluded with the need for a move towards full automation in the analysis. The two specific areas deserving attention in a fully automatic analysis were the representation of the lexical cohesion of the text in matrices and the actual identification of the segments. These two major points were addressed in the pilot study described here[7].

[7] Parts of this study have been presented in (Berber Sardinha, 1996a)

5.4.1 Goals

The general aim of the third pilot study was the development of a segmentation procedure which could place segment boundaries in the text without human intervention. The segmentation procedure had to be designed so as to be able to cope with several texts. This was a guideline which had to be followed throughout the research project, as stated above on p.191. However, it had been ignored so far since the priority had largely been to develop segmentation procedures and not to optimize them. This stage of the research seemed an appropriate time to try to implement this particular guideline. Therefore, the segmentation procedure which was developed in this pilot study focuses on developing a technique which can be automatically applied to several texts.

In pursuing the goal of designing a new procedure for unconstrained segmentation, these guidelines are followed:

Automatic computation of cohesion In the new procedure, lexical cohesion must be computed automatically

Automatic placement of segment boundaries In the new procedure, segment boundaries must be placed without human intervention

Capability to handle several texts The new procedure must be efficient enough to be applied to several texts

5.4.2 Alternative methods

A possible strategy in selecting a new framework from which to choose a new method was to re-evaluate the steps taken so far in the analysis of the example text. A constant in the analysis had been to describe the lexical cohesion in terms of a matrix of links. The problem with matrices is that

there is no simple way of designing a computer program to read them in and inspect their internal shape in the same way that had been done so far in the pilot studies.

One specialized statistical procedure which is designed to deal with data in matrices is Q-Analysis (Davies, 1985), which operates by slicing a matrix into parts which share spatial characteristics. This approach is intuitively appealing for the kind of analysis being developed here. The problem with Q-Analysis, though, is its restricted availability: it is not implemented in any of the major statistical packages (viz. SAS, SPSS, MiniTAB). This made it impossible to use Q-Analysis for the present research.

Other statistical procedures share similar characteristics with Q-Analysis. In fact, Q-Analysis is commonly regarded as being just one of the many types of *cluster analysis* procedures (SAS Institute Inc, 1989a, p.53). The general aim of cluster analysis is the partitioning of a data set into smaller groups of observations. This aim is coherent with what is expected of the segmentation of texts. Unlike Q-Analysis, all major cluster analysis procedures are available as part of statistical packages. Hence, cluster analysis provides a suitable framework for the present investigation.

5.4.3 Cluster Analysis

In this section, a presentation of the statistical techniques commonly referred to as cluster analysis (Alderfelder and Blashfield, 1984; Everitt, 1974) will be provided. The aim of the presentation is to show what motivated the choice of a technique which best suits the task of automatic segmentation of a corpus of texts. Before choosing the actual procedure, a brief introduction to cluster analysis must be given.

5.4.4 Introduction

Cluster analysis is the general name given to a series of procedures which are aimed at partitioning a data set into smaller groups of observations. Cluster analysis is also referred to by other names such as 'partitioning', 'clumping', 'unsupervised pattern recognition', and the more bizarre 'aciniformics' and 'agminatics' (Good, 1977). Regardless of the name, all approaches to cluster analysis share one important characteristic, namely that they do not require the input of *a priori* knowledge about the data (Woods et al., 1986, pp.259-260). Clusters are formed solely on the basis of the similarity (or dissimilarity) among the variables assigned to each observation.

Cluster analysis is not the only statistical procedure devoted to classification of observations in groups. Factor analysis is another of such procedures. Like cluster analysis, factor analysis works by finding similarities or lack of dissimilarity between individual cases based on a measure of relatedness between variables. Those sets of variables which are found to belong together are called 'factors'. An important difference between cluster analysis and factor analysis is that the latter incorporates information about negative correlations between variables thus producing factors to which variables contribute 'negatively' by being absent. Cluster analysis does not take into account negative correlations. Factor analysis has found its way into discourse analysis most notably through the work of Biber (e.g. Biber, 1988, 1995a; Biber and Finegan, 1988). Although factor analysis is said to have a theoretical underpinning, cluster analysis is reputedly an *ad-hoc* procedure. Accordingly, Biber (1988, p.65) warns of the need for a theoretically-motivated research design when using factor analysis. Such restriction does not apply to cluster analysis which is reportedly a much more exploratory set of techniques (Woods et al., 1986, p.259), and was therefore more suitable for an exploratory study into segmentation.

In summary, cluster analysis appeared a more appropriate technique for the current investigation than factor analysis because it is intuitively more directly related to how segmentation had been tackled in pilot studies 1 and 2. In previous pilot studies, segments were identified by searching for 'clusters' of lexically cohesive sentences; hence, a statistical technique such as cluster analysis, which is specifically devoted to identifying clusters, was naturally more appealing. The choice of cluster analysis over factor analysis was also due to the consideration that for the data being analysed there is no reason to suppose segments could be characterised 'negatively' by their lacking certain characteristics. If the data for this study had been coded in such a way that lexical items could have been noted for their absence in certain parts of the texts, then factors instead of clusters would have been more appropriate.

Having decided on cluster analysis, the next step was to choose which kind of cluster analysis to carry out. As said above, cluster analysis is not one single technique, but rather a set of procedures. There are two aspects that needed to be considered in choosing the kind of cluster analysis to be used in the analysis: clustering method, and similarity measure. Each of these aspects will be discussed in detail in what follows. As will be seen, the ways in which different methods and measures produce clusters can vary considerably, and it is in choosing the combination of method and measure that the researcher in part defines what type of cluster solution he/she will obtain or avoid.

5.4.5 How cluster analysis works

All cluster analysis methods have some important characteristics in common. These have been summarized by Rotondo (1984, pp.74-75):

1. Begin with n clusters, each consisting of a single object

2. Find the closest pair of clusters

3. Construct a new cluster by joining the closest pair of clusters

4. If the new cluster contains all n objects stop; otherwise repeat steps 2, 3, and 4

The basic principle underlying all clustering algorithms is that at the outset every observation is a cluster; from then on clusters are joined together until there is only one cluster left. These principles form the basis of what is generally called 'agglomerative hierarchical clustering' which, as the name implies, seeks to arrange the clusters in a hierarchy, that is, smaller clusters are joined into larger clusters which finally merge into a single cluster comprising the entire data set (SAS Institute Inc, 1989a, p.520).

Methods and measures

There are a variety of methods which can be used to perform cluster analysis, such as single linkage, complete linkage, average linkage, k-means and Ward's method. Each one of these works by computing distances between cases and clusters in a different way. In single linkage clustering, only the smallest difference between clusters is used in forming clusters, whereas the complete linkage method uses both the smallest and the largest differences.

Average distance clustering uses information about all cases in the clusters by computing an average distance, which can be of two kinds: either the average difference is amongst the members of two clusters or amongst the members of each cluster. The former method is called *between group average* or UPGMA (unweighted pair group method using arithmethic averages), and it sorts cases into clusters so that the average distance between the resulting clusters is as high as possible. The latter method is termed *within group average* and it assigns cases to those clusters where the resulting average

within the cluster will be as small as possible. Like average linkage, Ward's method also uses information about all cases in the cluster. First an average for each variable across all cases is computed, then the distances between each case and this average are summed up. Those cases are joined that contribute the least to an increase in the sum of distances within the cluster.

Just as there are many methods for cluster analysis, so there are also a number of similarity measures which can be used, such as the Euclidean distance and the City-Block (or Manhattan) measure. An Euclidean distance is obtained by calculating the difference between pairs of cases over all variables, squaring these differences, adding them up, and then taking the square root of the sum. If the square root is omitted, the measure is called 'squared Euclidean measure'. The City-Block or Manhattan measure differs from the Euclidean distance because it does not compute differences between pairs of cases but among all cases. The City-Block measure was used in text research by Phillips (1985).

The first problem in applying cluster analysis was the choice of method. Choosing a particular method would constrain the acceptable choices of measure as well. For example, it is generally recommended that squared Euclidean measures be used with Centroid, Median, and Ward's methods. The choice of method proves to be more challenging than the choice of measure, as some methods tend to produce clusters of certain kinds. For example, it is said that average linkage methods tend to produce clusters of the same variance, whereas the clusters produced by Ward's methods tend to be of similar size (SAS Institute Inc, 1989a, p.56).

In deciding on a method and a measure it is probably best to see which choices have been made in previous studies. These are discussed below.

Cluster analysis and linguistics

Since cluster analysis procedures have been designed for the purpose of classifying data, one might think that linguistics would be a field where cluster analysis would have been widely applied. Particularly in discourse analysis, cluster analysis would be very appropriate given that discourse analysis consists of classifying and labelling discourse features (Schiffrin, 1994). Yet, an examination of the linguistic literature of the past quarter of a century reveals the contrary.

The extent of the use of cluster analysis in the linguistic literature is only marginal. In order to verify this assumption, a search of the Linguistic and Language Behavior Association (LLBA) database on CD-ROM was conducted. The LLBA database spans nearly a quarter of a century of publications (1973 through 1996), and therefore it may be trusted as providing a representative sample of research in language. The expression 'cluster analysis' appeared in only 117 abstracts in the LLBA database, or about once in every one thousand entries. This indicates that cluster analysis is not widely used in linguistic research in general.

In studies dealing with text organisation, there are also very few instances of applications of cluster analysis. One important study which has made use of cluster analysis to investigate text organisation is Phillips (1985), who looked at clusterings of collocations in science textbooks.

Phillips

Cluster analysis was used by Phillips (1985) to identify groups of collocations in eight textbooks ranging in size from 48,000 to 63,000 words. Collocations were extracted by a concordancer (CLOC) for a sampling of about 200 words from each textbook. Each of the 200 words as well as their collocates were arranged in a matrix and analysed for clusters. The method employed for

the computation of clusters was Ward's method because previous studies had suggested it provided a superior clustering ability. The clusters were identified by inspection of dendrograms. Phillips tackled the problem of determining the number of clusters inherent in his data[8] by examining the values of the error sums of squares (ESS) yielded by Ward's method. The ESS indicated the amount of deviation from the cluster means resulting from the fusion of two clusters. If the ESS rose sharply as a result of a particular fusion, then it indicated that the merging of clusters should stop. This was combined with the observation of the contents of the clusters – the exact cut-off point was located at the place where it was felt that the procedure was forming spurious clusters. Before deciding on which clusters were artificial, any 'ragbag' clusters were omitted from further consideration. Those were clusters containing words which never participated in collocation. They are a product of the requirement that the clustering method group all cases in the data, and therefore they do not reflect the structure of the data. In addition to content, the 'ragbag' clusters could be visually identified by being large clusters formed at one single level.

Rotondo

A discussion on the problems arising from the use of statistical clustering in text analysis is provided by Rotondo (1984). Specifically, the author mentions the fact that in text analysis the number of objects to be clustered is normally greater than what is desirable (e.g. Pollard-Gott et al., 1979). The number of comparisons required for a cluster solution is typically equivalent to the square of the total number of cases. So, for an input of 1000 cases, 1 million (or 1000^2) calculations are necessary.

[8] See discussion about methods for determining the number of clusters in section 5.4.5 on page 242.

Importantly, Rotondo (1984) studies segmentation, which he defines as one possible partitioning task (others are sorting and sequence sorting). Segments are said to designate macro-units, or a 'coherent subpart' of text (p.72). The author addresses the problem of finding the right number of clusters in the data by asking a group of subjects to provide a segmentation of the data. The number of segments by each subject is then averaged, and the average serves as the optimum number of clusters, or segments.

The author chooses the single-linkage algorithm because he argues that in the segmentation task, only adjacent segments need to be compared for cluster membership, and therefore there would be unnecessary processing if more information were used in the computation of clusters. His clustering procedure is reportedly capable of handling up to 10,000 cases.

The author first illustrates his method with a 232-word passage from a biology textbook. The passage is part of a text segmented by 63 college students who were asked to mark the boundaries between 'complete thoughts' (p.78). The average number of segments was 6.9 (SD=4.9). A second passage was also segmented by students resulting in 8.54 segments per student. The passage in figure 5.14 on page 229 illustrates two clusters found in the text. The boundary is between sentences 17 and 18. The first cluster, from sentences 12 to 17, is labelled 'classifying supermarket merchandise alphabetically would lead to many practical difficulties', and the second, from sentences 18 to 23, is entitled 'classifying supermarket merchandise according to the nature of the product is more practical and convenient' (pp.79-81).

The results suggest that the students had an implicit understanding that a 'complete thought' typically included more than one sentence. The author also reports that some clusters were as large as or larger than a paragraph, which reinforces the idea that segments are dissimilar from sentences. Further, the fact that most subjects chose to equate the notion of complete

thought with units never smaller than a sentence and usually as large as a paragraph may suggest the psychological validity of segments.

Biber and Finegan

Cluster analysis played a central role in Biber and Finegan's (1988) investigation of 'speech styles' in English. They looked specifically at how adverbials cluster together to signal stance.

An important distinction offered by Biber and Finegan (1988) is between the near-synonymous terms 'genre', 'register', and (to a lesser degree) 'speech style'. Genres are labels assigned according to the 'topic and purpose', register according to the 'relations among participants and other characteristics of the communicative situation', and speech styles according to 'linguistic form' (p.4). They borrow the term 'speech style' from Ervin-Tripp and Hymes, but they extend it to include those aspects described by quantitative methods applied to corpora. The data are taken from the LOB and London-Lund corpora totalling 1.5 million words in 410 different texts. The adverbials are classified into six categories drawn from Quirk et al. (1985), each one containing those adverbials which are close in meaning to the label of the group; the group labels are: 'honestly', 'generally', 'surely', 'actually', 'maybe', and 'amazingly'.

The frequency of each of the eight adverbial types in the texts of the corpora was compared by means of cluster analysis, which identified eight distinct clusters. The number of clusters was determined by the Cubic Clustering Criterion (CCC) statistic. Cluster 1 is labelled 'Secluded from Dispute' and comprises 60% of the spontaneous speeches in the corpora; it is characterized mostly by the use of 'surely adverbials', but also by 'actually' and 'maybe'; cluster 2 is not given a name but it contains face-to-face conversations only (though only 5% of the total) and exhibits a predominance of

(12) Suppose a supermarket manager arranged his merchandise alphabetically. (13) Think of the varied goods to be found under the letter A: (14) abalones, almods, apples, apricots, artichokes, and many more. (15) These would be followed by bacon, baking powder, beans, beef, beets, bread... (16) Imagine the practical difficulties in such a system! (17) Refrigerators for perishable groceries would have to be scattered throughout the store. (18) Actually, in any supermarket we find that the merchandise has generally been grouped according to the nature of the product. (19) In one section we find various kind of canned goods; (20) in another, fresh fruits and vegetables; (21) in a third, meats. (22) Moreover, each of these sections may be further divided. (23) Familiarity with this system of classification enables the shopper to locate groceries easily and quickly.

Figure 5.14: Passage from example text in Rotondo (1984)

'actually adverbials'; cluster 3, 'emphatic shared familiarity', has 60% of telephone conversations and face to face conversations; cluster 4, 'faceless', is the largest cluster with nearly $\frac{2}{3}$ of the official documents, $\frac{1}{2}$ of adventure fiction, and more than $\frac{1}{3}$ of the academic prose, biography/ essays, general fiction and editorials - it has a high frequency of 'actually' adverbials but little use of 'surely' and 'maybe' adverbials; cluster 5, 'emphasis of individual position', has $\frac{1}{2}$ of the interviews, and $\frac{1}{3}$ of the prepared speeches, and no adverbial is remarkably frequent in it, being thus characterized by the absence of adverbials; cluster 6, 'generalized content', has about a quarter of academic prose and official document genres, and 'actually' adverbials are frequent in it; cluster 7, 'cautious', has a fairly large share of adventure fiction (35%) and general fiction (24%), and is characterized by large numbers of 'maybe' adverbials; finally, cluster 8, 'concession to reader/listener', has mostly broadcast (38%) and prepared speeches (33%) – the texts in it make frequent use of 'surely' and 'amazingly' adverbials.

Through the detailed examination of the texts in each cluster, Biber and Finegan (1988) note that the meanings attached to adverbials in context are different from the literal meanings of the adverbials. For instance, 'surely' adverbials are common in cluster 1, but they are used to 'invite affirmation' rather than to 'mark emphatic conviction' (p.19), as in 'you're in this senate

committee of course, aren't you?' (p. 21). They also note that the usage of the same adverbials across different clusters differs slightly. Both clusters 5 and 3 present a high frequency of 'actually' adverbials, but whereas in cluster 3 these adverbials are used in the general sense of emphasis, in cluster 5 they are employed in order to clarify or contrast through emphasis (p.26). Likewise, both clusters 1 and 8 are characterized by the frequent presence of 'surely' adverbials, but the clusters differ in that in cluster 8 'surely' adverbials take on the specialized meaning of 'concession', as opposed to the general meaning of 'assertion' which is associated to the instances of 'surely' adverbials across cluster 1 (p.29).

Other studies

Ledger (1989) investigates the authenticity of works of Plato (Epistles, Hippias Major, and other minor works) by conducting of a cluster analysis of several variables related to Plato's style. Cluster analysis is used to identify those stylometric characteristics which are most certainly related to Plato's established writing. The input to the cluster analysis is the orthographic word, a departure from other stylometric investigations which normally employ grammatical categories for authenticity research.

Karlgren et al. (1995) report on the application of cluster analysis to identify strategies used by Swedish speakers in translating isolated sentences into English. They cluster the possible translations of target sentences to find groups of similar translations.

Hughes and Atwell (1994) present an automatic evaluation of clustering schemes. They argue that an automatic evaluation is advantageous because it does not rely on the intuitive judgements needed in the 'looks good to me' approach, which consists of the analyst evaluating the results of the clustering procedure by inspecting the layout of dendrograms. In their experiments, the

clusters are evaluated by checking to what extent they reflect grammatical categories (nouns, verbs, etc). A score out of 100 is given to each cluster based on how consistently they reflect a single grammatical category. The clusters were formed using two algorithms: by sentence position, and by co-occurrence with a bigram. Each of these algorithms was tested by a combination of 34 metrics (Manhattan, Euclidean, and Spearman Rank Correlation Coefficient) and 8 methods (Single linkage, Complete linkage, Group average, Weighted Group average, Median, Centroid, Centre of Gravity, Ward's). The results indicated that the best scheme was Manhattan metric, and Ward's method, which achieved a score of 76 (i.e. on average 76% of the words in each cluster were assigned to the correct grammatical class). In a separate paper, the authors explain that Ward's method worked better because it avoided producing one-item clusters (Hughes and Atwell, nd, p.2).

Insights from previous literature

There appears to be little consensus in the previous literature on a particular method and measure. One reason for this lack of consensus is the diversity of data types being investigated. Different studies have made choices which are suitable to the specific characteristics of the data they were interested in. This is important in that it suggests that the best approach in the present pilot study would be to run trial analyses on a sample of target data employing a number of possible methods and measures.

An important guideline is provided by Biber and Finegan (1988). They chose a non-hierarchical procedure because they argued that their data were not arranged hierarchically. The same principle can be applied in the context of the present pilot study. Here, the segmentation which it is hoped the cluster analysis will provide is non-hierarchical, that is, it is not the aim here to find subdivisions of segments or groupings of segments. Therefore,

non-hierarchical procedures should be preferred over hierarchical ones. Nevertheless, as argued above, the final choice should be a result of trial testing various clustering algorithms through a sample of the target data.

In the sections below, trials are reported which were carried out on textual data similar to the target data to be segmented subsequently. Instead of trying each of the many clustering methods available, one representative of hierarchical clustering and one of non-hierarchical clustering was experimented with.

Non-hierarchical clustering: k-Means

The procedure known as *k-means* works by first choosing from the cases a k number of observations which are well-separated. These form the initial cluster centres (also sometimes referred to as cluster *seeds*). The analyst has an option of choosing the initial cluster centres herself/himself, or leaving it to the statistical package to do this for her/him. If the computer program is left with the task of choosing the cluster centres, then the analyst needs to specify at least how many clusters he/she wants to split the data into.

A data set is needed for the explanation of the clustering methods, and for this purpose I have chosen a letter published in *the Independent* in May 1995 (retrieved from the electronic version of the newspaper available on CD-ROM), which is reproduced in figure 5.15 on the following page.

The repeated lexical items in the text are listed in table 5.1 on the next page, which are the actual data which will enter in the computation for clusters. The data consist of each lexical item followed by a pair of sentences in which they appear. For the purposes of the explanation of cluster analysis, each individual pair of sentence positions is called the *coordinates* for a lexical item.

1)Sir: Your article on measures to control vehicle air pollution ('Air quality set to remain poor', 5 May) failed to mention one existing legal restraint which is insufficiently used in cities - speed limits. 2)Nearly all forms of vehicle pollution are directly proportional to the amount of fuel burnt, so the faster and more aggressively a car is driven, the worse it pollutes. 3)If the 30 mph speed limit was properly enforced, and drivers could restrain themselves from roaring away from traffic lights, there would be a useful reduction in urban pollution. 4)Drivers in London know they can get away with 40 mph. 5)There are too few of the recently introduced Gatso automatic radar cameras, and they seem to be set at 40 mph - plus. 6)Why not set them at 30 mph? 7)It might mean a mountain of prosecution paperwork in the short term, but in the long term we'd have cleaner air. 8) Yours sincerely, GEORGE BENNETT, Editor, Truck magazine, London

Figure 5.15: Example text for illustrating clustering procedures

Item	Coordinates	
air	1	7
driver	3	4
limit	1	3
london	4	8
mph	3	4
mph	3	5
mph	3	6
mph	4	5
mph	4	6
mph	5	6
pollution	1	2
pollution	1	3
pollution	2	3
restraint	1	3
set	1	5
set	1	6
set	5	6
speed	1	3
vehicle	1	2

Table 5.1: Data for illustrating clustering procedures in alphabetical order

The first step in conducting a cluster analysis by *k-means* is to search the data for two initial cluster centres. Two excellent candidates are 'london' which appears in sentences 4 and 8, and 'pollution', appearing in sentences 1 and 2. These two cases are placed well apart in the text and thus seem good cluster seeds. The squared Euclidean distance between them is 45 ($(1-4)^2 + (2-8)^2 = 3^2 + 6^2 = 9 + 36$), which is higher than for any other pair of cases. Thus, 'london' will be taken as the initial centre for cluster 1, and 'pollution' for cluster 2. The *k-means* procedure works by fitting cases to the closer centre mean, and so each cluster must have a mean. Since each cluster has only one case so far, the means are simply the values for each cluster seed, namely 4 and 8 for cluster 1, and 1 and 2 for cluster 2. Now that the initial cluster centres have been chosen and each cluster has had its mean computed, each case in the list is compared to each cluster mean.

The first case down the list is 'air', which occurs in sentences 1 and 7. Its distance to cluster 1 is equal to 10 (because $(1-4)^2 + (7-8)^2 = 3^2 + (-1)^2 = 9+1$), but to cluster 2 it is larger: 25 (or $(1-1)^2+(7-2)^2 = -0^2+5^2 = 0+25$); therefore 'air' joins cluster 1. The mean for cluster 1 now changes, because of the new member. The values in it are 4 and 8 for the initial seed, and 1 and 7 for 'air', so the mean is $(1+4) \div 2 = 2.5$ for the first sentence coordinate and $(7+8) \div 2 = 7.5$ for the second. The mean for cluster 2 remains unchanged as 1 and 2.

The next case is 'driver' (coordinates 3 and 4). Its distance to cluster 1 is calculated as being equal to 12.5, since $(2.5-3)^2 + (7.5-4)^2 = -.5^2 + 3.5^2 = .25 + 12.25$, while in relation to cluster 2 the distance is 8 (i.e. $(1-3)^2 + (2-4)^2 = -2^2 + (-2)^2 = 4 + 4$). As a result, 'driver' joins cluster 2, which now will have a mean centre equal to 2 and 3, because $(1+3) \div 2 = 2$ and $(2+4) \div 2 = 3$. The cluster centres now are 2.5 and 7.5 for cluster 1, and 2 and 3 for cluster 2.

All cases are processed in the same manner, and are allocated to one of the two clusters. As the cases are ascribed to a cluster, the cluster seed is continuously updated, until in the end, once all cases have been distributed between the clusters, final cluster centres can be computed. The final division of the cases into clusters is shown in table 5.2 on the following page. The final cluster centres for cluster 1 are 3.3333 and 6.1111, and for cluster 2, 1.5 and 3.2. The final cluster distances for each cluster member to the final cluster centre can now be estimated in the same way as they were calculated during the cluster assignment phase. The final distances to the cluster centres are also presented in table 5.2.

As can be seen in table 5.2 on the next page, the initial data in table 5.1 on page 233 were rearranged. Taking just the first three items in table 5.1, namely 'air', 'driver', and 'limit', it is interesting to see that 'air' was assigned to cluster 1, whereas 'driver' and 'limit' ended up in cluster 2. The final assignment of these three cases to two different clusters illustrates how the initial arrangement of the cases did not influence the clustering. It is also important to note that the final assignment makes sense, in that 'driver' and 'limit' appeared much closer to each other in the text than 'air', and therefore 'driver' and 'limit' did in fact belong in the same cluster.

Hierarchical clustering: Between groups average

The computation of the other major type of clustering procedure, *between groups average*, is carried out by first taking each observation as a one-member cluster. In this way, on the first pass through the data, there will be as many clusters as there are observations. The second step consists of matching every case against each other and calculating the average distance between the resulting clusters. For *between groups average*, cluster membership will be decided on the basis of the arrangement which results in the

Cluster	Item	Coordinates		Distance
1	air	1	7	6.234432
1	london	4	8	4.012432
1	mph	3	5	1.345632
1	mph	3	6	0.123432
1	mph	4	5	1.679032
1	mph	4	6	0.456832
1	mph	5	6	2.790232
1	set	1	6	5.456632
1	set	5	6	2.790232
2	driver	3	4	2.89
2	limit	1	3	0.29
2	mph	3	4	2.89
2	pollution	1	2	1.69
2	pollution	1	3	0.29
2	pollution	2	3	0.29
2	restraint	1	3	0.29
2	set	1	5	3.49
2	speed	1	3	0.29
2	vehicle	1	2	1.69

Table 5.2: Final cluster distribution of example data

greater average distance between the clusters.

To illustrate this procedure, consider the first three cases in isolation from the whole of the data set in table 5.1 on page 233, namely 'air', 'driver' and 'limit' to see how these can be grouped in two clusters so that the resulting clusters are as different from each other as possible. There are three possible arrangements into which the three cases can fall: (1) 'air' and 'limit' in cluster 1, and 'driver' in cluster 2; (2) 'air' in cluster 1, and 'driver' and 'limit' in cluster 2; and (3) 'air' and 'driver' in cluster 1, and 'limit' in cluster 2. For each of these situations, the average distance between the clusters must be computed, by working out the Euclidean average (or another measure) across the members of each provisional cluster.

Accordingly, for solution (1), the average is obtained by calculating the distance between 'air' and 'driver' (which is 13, or $(1-3)^2+(7-4)^2 = 4+9$), and between 'limit' and 'driver' (which is 5, or $(1-3)^2 + (3-4)^2 = 4+1$); the average is then simply 9 ($13+5 \div 2 = 18 \div 2$). For solution (2), the same

set of calculations is performed: the distance between 'air' and 'driver' is 13 ($(1-3)^2+(7-4)^2$), and between 'air' and 'limit' is 16 ($(1-1)^2+(7-3)^2$), so the average is 14.5 ($13+16\div 2=29\div 2$). Finally, for solution (3), the distance between 'air' and 'limit' is 16 ($(1-1)^2+(7-3)^2$), and between 'driver' and 'limit' is 5 ($(3-1)^2+(4-3)^2$), thus yielding an average of 11.5 ($16+5\div 2=21\div 2$). By comparing the averages, the largest distance between clusters is 14.5, which corresponds to arrangement 2. Thus, the best clustering solution is that which groups together 'air' in one cluster, and 'driver' and 'limit' in another. In fact, this arrangement, it could be argued, identifies the two items which are truly placed closer together in the text; 'driver' and 'limit' occur in sentences 3 and 4, and 1 and 3 respectively, therefore they both have sentence 3 in common; 'air', on the other hand, occurs in a much later sentence than the common stretch where 'driver' and 'limit' can be found, namely between sentences 1 and 4; admittedly, 'air' also occurs in sentence 1 with 'limit', and there is therefore some ground for arguing that they should have been clustered together but this would not have made the resulting clusters maximally different, which is exactly the purpose of the *between groups average* method.

Dendrogram

The three initial cases in the data have been worked through in detail so that the computations involved in clustering by the *between groups average* method become clear. Normally, when all cases have been dealt with, the results are displayed graphically in what is generally known as a 'dendrogram'. The example data generates a dendrogram such as that shown in figure 5.16 on page 239.

The dendrogram shows by means of lines connecting individual cases or clusters how cases are successively combined into hierarchical clusters. Along

the top a ruler indicates how distant from one another each cluster is, so the further from the left clusters are joined, the more distant they are. For instance, take the first two cases appearing in the dendrogram — the first occurrence of 'pollution' and 'vehicle' are indicated as being very similar, and this is correct since both appear in sentences 1 and 2. As the shape of the interconnecting lines indicate, the six cases which are firstly joined across the top of the dendrogram are in order of appearance 'pollution' (case 11), 'vehicle', 'restraint', 'speed', 'limit', and 'pollution' (case 12).

At a later stage, the last case of 'pollution' (case 13) joins in, thus resulting in a more heterogeneous cluster, which is represented by the length of the line connecting all the members of the cluster together. More cases are incorporated into this cluster, namely 'air' and two occurrences of 'set' (cases 15 and 16). At this point, the dendrogram shows a break, indicated by the lack of early connection between this cluster and the remaining cases. The next listed case, 'mph' (case 10), is first linked instead to 'set' (case 17) (both occur in sentences 5 and 6) and then to 'mph' (case 9, occurring in sentences 4 and 6). Then there is another break and a cluster is formed by 'driver' and 'mph' (case 5), both appearing in sentences 3 and 4, followed by another cluster comprising the three remaining mentions of 'mph' (cases 6, 8 and 7), which have as coordinates sentences 3 and 5, 4 and 5, and 3 and 6. These three individual clusters are joined together at later stages into a single cluster, to which 'london' is finally added at a much more remote distance.

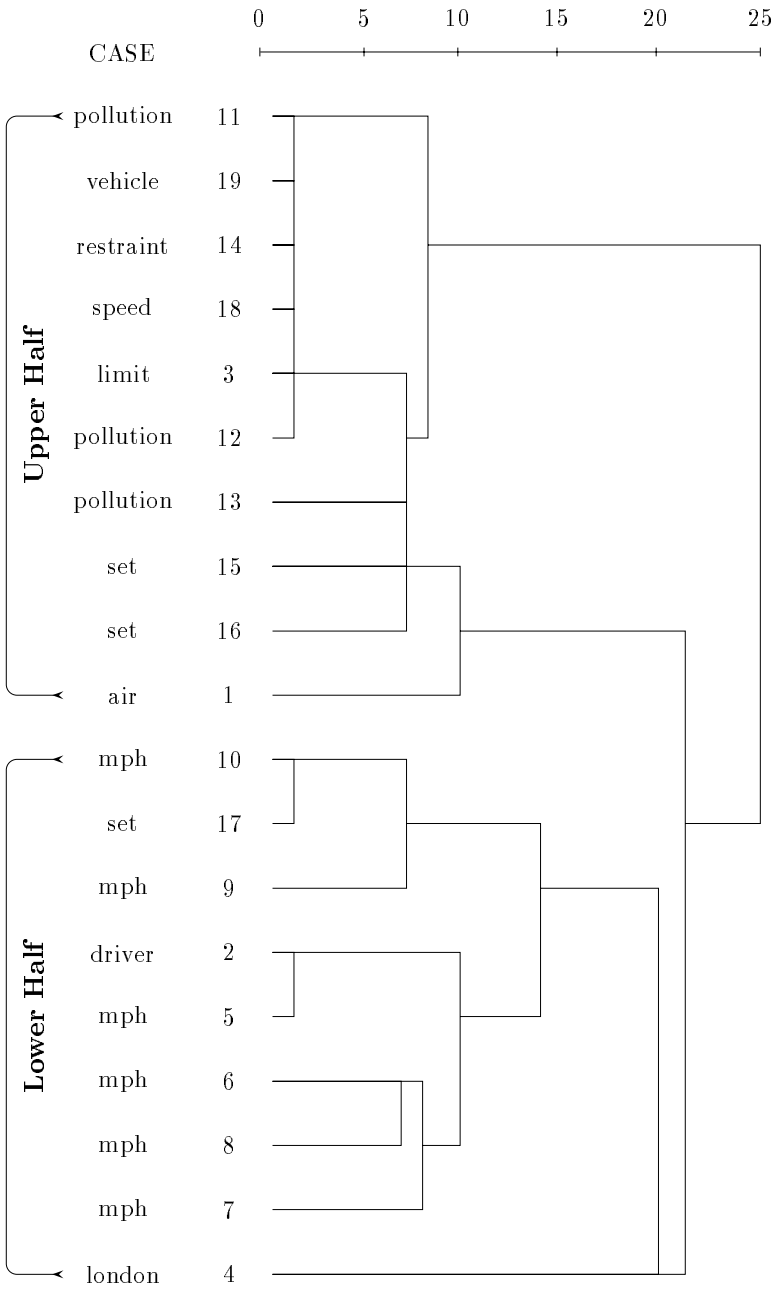


Figure 5.16: Dendrogram for example data using between group averages

The dendrogram is a perfect illustration of the way hierarchical clustering works, since it shows in considerable detail the points where individual clusters are formed and merged and at which level of similarity. For this reason, it also shows the points at which clusters differ the most. In the example data, one major divide was clear in the dendrogram, namely the one which splits the diagram almost in half between 'air' and 'mph'. This break is not only striking because it shows two clusters which exhibit the largest level of difference between them, but also because it indicates a place where one could be confident about dividing the data up into two main clusters. These two portions are identified in the dendrogram in figure 5.16 as an 'upper half' and a 'lower half' cluster.

Comparison of k-means and between groups

Table 5.3 on the next page provides a comparison of which cluster individual lexical items were assigned to by each method. For the purposes of the comparison, cluster 1 from the *k-means* analysis was lined up with the lower-half cluster (i.e. the cluster in the lower half of the dendrogram, as shown in figure 5.16 on the preceding page) from the *between group averages* procedure, and cluster 2 was lined up with the upper-half cluster. The similarity between the two solutions can be assessed by checking whether each lexical item was assigned to the same cluster in both solutions.

The table indicates that of the nineteen cases in the data, sixteen (84%) were assigned to similar clusters: eight cases were assigned to k-means cluster 1 and between groups average lower-half cluster, and eight were classified as k-means cluster 2 and between groups average upper-half cluster. Only three cases were classified in a different way, namely 'driver', 'mph' (case 5), and 'set' (case 17); 'driver' and 'mph' appeared in cluster 2 from k-means and in the lower-half cluster from between groups clustering, whereas 'set' was

Item	Coordinates		k-means	between groups
air	1	7	1	bottom
london	4	8	1	bottom
mph	3	5	1	bottom
mph	3	6	1	bottom
mph	4	5	1	bottom
mph	4	6	1	bottom
mph	5	6	1	bottom
set	5	6	1	bottom
limit	1	3	2	top
pollution	1	2	2	top
pollution	1	3	2	top
pollution	2	3	2	top
restraint	1	3	2	top
set	1	5	2	top
speed	1	3	2	top
vehicle	1	2	2	top
driver	3	4	2	bottom
mph	3	4	2	bottom
set	1	6	1	top

Table 5.3: Cluster membership in k-means and between groups solutions

classified as k-means cluster 1 and between groups upper-half cluster. The comparison indicates that both methods are largely similar, and another criterion therefore must be sought to assist in deciding which method to use.

Choice of a method

An initial guideline mentioned previously could be applied to choosing an appropriate method, namely that the kind of the segmentation intended here is non-hierarchical. This suggests that k-means would be a better choice.

Another guideline could be the demand each method places on computational resources. Given the high number of lexical links which can be found in large texts, it would be more appropriate to choose a method which is not particularly taxing on computational resources. Hierarchical cluster analysis makes heavy demands on computer resources, and consequently it is not generally recommended for large data sets. What constitutes a large data set is relative, but generally data containing between 100 and 100,000 observations

may be considered large. For such data sets, non-hierarchical procedures such as k-means are recommended.

In view of these two guidelines, the option which best presented itself is k-means clustering. Before applying k-means clustering to text segmentation a crucial element of clustering had to be dealt with: the determination of the number of clusters in the data. In the illustration of the methods presented above, the number of clusters was decided beforehand. In real segmentation, the number of clusters must not be decided arbitrarily *a priori*, hence a procedure must be devised which will indicate the number of clusters in the data.

Determining number of clusters

The problem of finding the right number of clusters is perhaps the most crucial in applying cluster analysis to a set of data. A number of statistics exist which can be used for determining the number of clusters in the data. These statistics are sometimes referred to as 'stopping rules'. As with clustering methods, there are no undisputed criteria for deciding on a suitable stopping rule.

A comprehensive comparison of most stopping rules is provided by Milligan and Cooper (1985). They have tested the ability of 30 procedures to recover well-defined clusters generated artificially. Since the data were artificially created, the authors warn that the performance of stopping rules may be different with realistic data (Milligan and Cooper, 1985, p.162). Nevertheless they also note that those stopping rules which fail to recover distinct clusters in the artificial data stand no greater chance of identifying clusters in authentic data, and therefore the results of their experimentation serves as a valid index of the power of most stopping rules.

The actual results indicate that the six best performances were achieved

by Calinksy and Harabasz index, Je(2)/Je(1) ratio, C-Index, Gamma index, Beale ratio, the Cubic Clustering Criterion, and Point-Biserial index. All of these indicated the correct number of clusters for two to five cluster solutions more than 300 times out of the 432 possible; the best procedure identified the right number of clusters 390 times. It would appear that the researcher should use that procedure which has yielded the best performance, but a few considerations should be borne in mind. First, as already mentioned, the performance results are not data independent. It has not been claimed by the authors that the same results would be repeated in another data set, especially if the data are real, that is, data in which the clusters are not so distinct. Second, not all stopping rules are available for the applied researcher to use; in fact most of them depend on extensive mathematical knowledge for their implementation. In certain areas such familiarity with what may be complex statistical formulae must not be expected, which limits the practical applications of many stopping rules.

Since no stopping rule achieved 100% error free clustering, the kinds of errors they make are worth noting. In the case of the CCC, its errors were mainly to do with identifying more clusters than there were in the data (Milligan and Cooper, 1985, p.169). As Milligan and Cooper (1985, p.159) explain, overestimating the number of clusters is the less serious of the two kinds of errors possible in determining the true number of clusters in the data. The more serious error involves assuming there are too few clusters because in this case there is loss of information by merging clusters which should have been left apart. It is granted that finding too many clusters is also undesirable, but at least in these situations the information present in the data may be overrepresented but not missing or disguised.

In the context of the present research, availability as part of a statistical package is a deciding factor in choosing a stopping rule. The Cubic Clustering

Criterion (CCC) is one such commercially available stopping rules, being part of the SAS statistical package. Furthermore, the CCC has also been used in linguistic research. Biber (1995a, p.413) identified the number of multi-dimensional text types in English and Somali by applying the CCC statistic to a cluster analysis. The fact that it has been proved useful in research on language, coupled with the facts that it is ready for use and that it achieved good performance on the comparative tests settled the matter in favour of the Cubic Clustering Criterion.

Evaluation of choices

So far, the choices which had been made were that a segmentation procedure would be used based on the application of k-means clustering informed by Cubic Clustering Criterion. It was now necessary to assess how well this approach fitted into the guidelines for the pilot study.

The first guideline was that the new procedure should be capable of handling several texts. Cluster analysis provides such capability, hence this particular guideline was successfully followed. The second guideline stated that the new procedure should be capable of placing boundaries without human assistance. This is achieved by running k-means clustering in SAS `fastclus`. The number of clusters in turn is based on the Cubic Clustering Criterion statistic which is also provided by SAS. And the actual boundaries are supplied by the position of the disjoint clusters provided by the non-hierarchical clustering algorithm. Finally, according to the third guideline, the computation of lexical cohesion should also be automatic. In the two previous pilot studies this was accomplished by using the `links` program; however, in order to run the data through SAS `fastclus` a different output to that provided by `links` is needed. Therefore a new computer program had to be created to provide data suitably formatted to run in SAS. This is discussed in the

next section.

5.4.6 `Words` program

The program used to compute the lexical cohesion links across the text is called `words`[9]. It is capable of removing non-lexical words from the text and normalising the remaining lexis by stripping out affixes, lemmatising words with inflections and/or derivations, and looking up synonyms in a list. The normalisation is achieved by inputting control files containing the necessary information for each task. The output is a list of lexical words and their respective positions in the sentences of the text (see section 5.4.6 on page 247).

The identification of 'words' by computer is not a trivial matter (Atwell, 1986, p.175). The easiest way to set up a computer to locate word boundaries is to make it identify strings of characters separated by blank spaces and punctuation marks, but this approach immediately leaves out compound items such as 'of course' and 'take up' (Barnbrook, 1996, p.58). For this reason, the way `words` handled the identification of word boundaries involved two steps. The first was the identification of orthographic words, which were strings of characters enclosed by delimiters. The list of word delimiters was input as a control file, and included blank spaces, 'tabs', and punctuation marks. The second step was the identification of multi-word items. These items were specified by the user in a separate input file and could include any items consisting of more than one word, such as 'San Marino'. `Words` simply read in the list of multi-word items and tried to match the strings of isolated words in the text to the entries in the multi-word control file. `Words` did not have the means to check the correctness of the multi-word items, and so it

[9] I am grateful to Kevin O'Donovan and Dr Rob Birch for their kind help in developing `words`.

was entirely up to the researcher to supply as correct and comprehensible a list as possible.

To illustrate the process of locating words, if a text contained the lexical item 'San Marino', `words` would first isolate the strings 'San' and 'Marino' as two separate orthographic words, and would then join 'San' and 'Marino' as one single item, 'San Marino'. With respect to lexical cohesion, if this expression were repeated in another sentence, the consequence would be that only one link would be counted for 'San Marino', instead of two. The possibility of identifying multi-word lexical items was a useful feature, since it meant the program was not restricted to identifying orthographic words only. Nevertheless, during the analysis of large amounts of data (see discussion below on page 336) it was felt that there were far too many items to include in the control file, and therefore there was a risk of inconsistency in the analysis if not all of them were taken into account. As a result, multi-word recognition was abandoned in later stages of the analysis.

The development of `words` took more than six months. During this period the program was both enhanced and debugged. The performance of the program was constantly monitored by running texts through it which were short enough to be analysed by hand. During the development of the program, the analysis provided by `words` was checked against a manual analysis to make sure that the output provided by `words` was always accurate. The control files were changed in the process to ensure that as few links as possible were ignored. However, it became apparent as more texts had to be analysed that it would be unrealistic to aim for the identification of all links in the texts (see discussion on p.336). Thus, the level of normalisation achieved in the analysis was partial.

In other words, it was felt that it was unrealistic to try to finely tune the control files so that all different word forms were lemmatised, all differ-

ent synonyms were correctly matched, and all multi-word groups were adequately tokenized, and therefore it was decided that no further effort would be invested in updating the control files. Thus, the control files used in the analysis contained but a subset of the instructions which would be necessary to normalise the texts in full.

As mentioned above, `words` works by reading in a source text and optional control files. The control files are made active by flagging them on the command line with the proper switch. The current valid switches are displayed in figure 5.17 on the next page, a screen invoked by running `words` without any options. As can be seen in the figure, control files dealing with the following aspects of text handling can be read in by `words`: removal of stop words, stemming (removing affixes), lemmatisation (providing word roots), recognition of word boundaries, multi-word items, and abbreviations. In addition, a table of synonyms can also be supplied, an option which appears listed as 'thesaurus' in figure 5.17. A more detailed account of the way in which `words` performs these operations is presented in section 5.4.6.

Figure 5.18 on the next page shows the `words` output for the example data presented above in pilot study 3 (see figure 5.15 on page 233). By inspecting the original text and the links obtained previously, it is possible to attest that `words` has correctly identified all the links in the text. Notice how the link between 'restraint' and 'restrain' was correctly picked up by adding this entry to the lemmatisation file. The actual layout of the output files was designed to allow their use as data files straight into SPSS and SAS.

Algorithm

The structure of `words` is made up of three main components: modification of input text, identification of repeated strings, and output of results. The

```
Words Version 1.3
Options:

        -r      specify number of links
        -s      specify stop word file
        -m      specify stemming rules file
        -l      specify lemmatisation file
        -p      specify sentence marker file
        -t      specify thesaurus file
        -w      specify word delimiters file
        -g      specify paired word file
        -x      specify abbreviation file
        -o<n>   switch on output file <n>
```

Figure 5.17: `words` options

```
0001 0002 0001x0002vehicle-vehicle
0001 0007 0001x0007air-air
0001 0002 0001x0002pollution-pollution
0001 0003 0001x0003pollution-pollution
0001 0005 0001x0005set-set
0001 0006 0001x0006set-set
0001 0003 0001x0003speed-speed
0001 0003 0001x0003restraint-restrain
0001 0003 0001x0003limit-limit
0002 0003 0002x0003pollution-pollution
0003 0004 0003x0004mph-mph
0003 0005 0003x0005mph-mph
0003 0006 0003x0006mph-mph
0003 0004 0003x0004driver-driver
0004 0008 0004x0008london-london
0004 0005 0004x0005mph-mph
0004 0006 0004x0006mph-mph
0005 0006 0005x0006set-set
0005 0006 0005x0006mph-mph
```

Figure 5.18: `words` output for example data

operation of each one of the components is outlined below[10].

The main point about the structure of the program is that it works by identifying *exact* matching strings. The program was designed to pick up simple repetition (Hoey, 1991b) only, and so it will match `dog` and `dog` but it will not do so for `dog` and `dogs`, or `dog` and `canine`. In order for other kinds of repetition to be picked up, the user has to modify the input text prior to the identification of the repetitions. So, for example, in order for `dog` and `dogs` to be matched, the user would have to tell the program to replace `dogs` with `dog` in the input text (or vice versa), or alternatively, have the program remove the final `-s` in all words of the input text, which would result in `dogs` being changed to `dog`. The facilities to make these modifications are available for convenience within the text modification component of the program, which is described next, but the user would have the option of modifying the input text using other means, such as through the search-and-replace function available in most word processors.

Component 1: Modification of input text The first component of `words` is constituted by 8 modules, each designed to make a specific kind of alteration to the input text(s) prior to the identification of the repetitions. The modules are similar in that they all work on a search-and-replace (or search-and-delete) basis. The execution of each of the modules is guided by control files which contain the target strings, and where appropriate, the strings which must be substituted. The execution of the first component is not obligatory; it may be activated wholly or in part depending on the options flagged by the user on the command line. The instructions for modifying the input text are detailed in control files, each one containing instructions on the search-and-replace or search-and-delete operations applicable in each stage.

[10] Further information on the program can be obtained from the author by writing to: R Paracatu 357 apto 52, 04302-020 São Paulo SP, Brazil

The user invokes a particular control file by specifying it on the command line, so for example:

```
words -s myfile.stp -m myfile.ste mytext.txt
```

would tell `words` (1) to remove the non-lexical words in `mytext.txt` using the instructions in `myfile.stp`, and (2) to stem the words in `mytext.txt` according to the rules in `myfile.ste`. In this case, `words` would tokenise the input file into words and sentences using the default word and sentence delimiters (see items 2 and 4 below), and output a list of repeated words.

The individual modules are explained below:

1. *Remove punctuation marks in acronyms.* This step is necessary to avoid treating the dots that are part of acronyms as end-of-sentence markers when module 4 is executed. The program reads in the control file listing acronyms, searches for them in the input text, and replaces them with the letters in the acronyms without the dots. For instance, given the following control file:

   ```
   U.S.A.
   U.K.
   ```

   the program would look for `U.S.A` in the input text, and if it finds it, it would replace it with `USA`. Then it would do the same with `U.K.`.

   This module may also be used for dealing with another important character: the decimal point, which is the same as the character denoting end of sentence. Without dealing with this character now, the user would have trouble later on during the execution of module 4 where the program must correctly identify the boundaries between sentences.

This could be achieved by a control file whose beginning would look like this:

```
.0
.1
.2
```

and which would go up to `.9`. All of these characters would be replaced with their forms without the decimal point.

2. *Identify word boundaries.* This step deals with finding word tokens. `Words` reads in the control file containing word separator characters and treats each string of characters delimited by such characters as a word. These delimiters are listed in a separate file; the delimiters used for the analyses presented here are:

   ```
   ;''-,:\/"()[]{}<>$%=+#&
   ```

   The blank space, the end of line characters, and the full stop are default word separators, and so these characters do not need to be specified.

3. *Identify certain groups of words as a single lexical item.* The program reads in the list of word groups from a control file and treats each occurrence of those words as a single item. The original boundaries identified in step 2 are updated accordingly.

   For example, a suitable control file might include the following line:

   ```
   San Marino
   ```

   This would cause `words` to treat the string `San Marino` as single word, and upon encountering it in two separate sentences, only one link would be counted, instead of two (`San` and `Marino`).

This option was used in the analysis of the texts in pilot study 3, but not in subsequent studies because the number of word groups in the data was very high, and it was felt that it would not be possible to provide a full account of them.

4. *Tokenise the sentences in the input text*, and number them. Two kinds of numbering were implemented: plain numbers and percentage of the total number of sentences in the text. In a 10-sentence text, the first sentence would be identified as '1' according to the plain-number scheme, and as '10%' in the percentage scheme. The sentence delimiters must be listed in a separate file, with the exception of the full stop (.), which is the default. A suitable sentence tokenisation control file would be[11]:

```
!?
```

5. *Remove non-lexical words.* In this module, `words` reads in the stop words listed in the control file, searches for them in the input text, and deletes them. The stop list is found in appendix 2 on p.448ff.

6. *Stem the words in the text.* The aim of this module is to strip away common prefixes and suffixes, thus reducing some of the words to their base form. The program reads in a file containing common affixes, searches for these strings in the input text, and deletes them. A stemming control file would look like the following:

```
-ed
-s
```

[11] Admittedly, the blind use of these characters would not correctly tokenise certain cases such as sentences separated by ellipsis marks (...) and embedded punctuation as in " 'Hello!', she said."

```
-ing
anti-
un-
```

The alterations performed in this module are 'blind' in that no disambiguation takes place. Hence, by deleting the prefix `anti-`, it is possible to reveal the similarity between `nuclear` and `antinuclear`, but at the same time, a word such as `anticipate` would be reduced to a meaningless string (see the section on limitations on page 256).

7. *Lemmatise lexical words.* This step was introduced in order to cope with those words whose similarity would not be made apparent by simple stemming, such as irregular verbs. `Words` reads in the control file, searches for the strings specified in it, and replaces existing strings accordingly. The basic structure of the lemmatisation control file is:

```
lemma > word_form
```

For example, if the following were part of the lemmatisation control file, `words` would treat each occurrence of 'see' and 'saw' as equivalent:

```
see > saw
```

The lemmatisation control file also handles certain words which were not properly altered during stemming; for instance, if the final `ed` had been removed from `omitted` during stemming, the resulting word form would be `omitt`. The following line in the lemmatisation control file would replace `omitt` with `omit`:

```
omit > omitt
```

The control file can be found in appendix 3 on p.453.

8. *Identify thesaural elements.* This module is aimed at allowing for the identification of repetition between synonyms, antonyms, and superordinates. The format of the of the control file is identical to that used for lemmatisation, as explained above, and so the following would be a suitable control file:

```
vehicle > car, lorry, bus, tractor
```

As in the previous step, `words` reads in the control file, searches for the strings specified in it, and makes the replacements accordingly.

Component 2: Identification of repeated strings The second component is the core of the program, since it is within it that repetitions are computed. The repeated words are identified by locating those strings which are identical. Firstly, a counter keeps track of how many times a particular word has been repeated. Secondly, the program makes sure that all equivalent words share a pointer to the same counter, so incrementing any one of them increments the value for them all. Thirdly, the program builds a tree having sentences for branches and the words in them for leaves, and searches this tree for words that appear in a sentence twice or more, and then marks all but one of these occurrences for ignoring. This is crucial, since according to Hoey (1991b) two occurrences of the same word in the same sentence contribute with one link only. Finally, the program looks for replications up to the end of the file:

```
For each sentence
  For each word in the current sentence
     Get the replaced version of this word
        Start searching for matching strings
          From each subsequent sentence
```

The program then creates a record for each pair of sentences consisting of two numbers identifying the sentences in which they occurred (either plain numbers or percentages), the total number of links, and the words forming the links.

Component 3: Output of links The third component is devoted to the output of the links into a file. Several different formats of output file were programmed into `words` during its development, some of which were later abandoned, having been used to facilitate debugging and/or to assist at preliminary stages of the analysis. There are two basic kinds of output: one listing the links between sentences, and another containing the input file with numbered sentences. The basic format of the former kind of output file is:

```
sentence_number_1 sentence_number_2 total_links repeated_word_1
                                                repeated_word_2 ...
                                                 repeated_word_n
```

For instance, the following is the first line of the output file obtained from the analysis of text 9, as shown in appendix 12 on page 472:

```
0001 0002 4 equatorial guinea mainland gulf
```

This line of output indicates that between sentences `0001` and `0002` there are `4` links, namely: `equatorial`, `guinea`, `mainland`, and `gulf`.

Two variations of this type of output are available: one with sentences represented by plain sequential numbers (1, 2, 3, ...) and the other with individual sentence numbers as a percentage of the total number of sentences. The selection of a particular kind of output is possible through the `-o` switch on the command line followed by an identifier: '1' for plain numbers, and '2' for percentages. The example above was obtained by selecting `-o1`.

The second type of output is simply a copy of the input file whose sentences have been numbered. The text is formatted in such a way that each sentence takes only one line. To illustrate, the following is the first sentence of one of the texts that have been analysed:

```
[[0001]]{{0.56}} sxbrk Deficits in inferior parietal perfusion ...
```

The set of figures in initial position indicate that this is the first (`[[0001]]`) sentence of the text, or sentence 0.56% ({{`0.56`}}) of the total. The code immediately after these figures (`sxbrk`) was placed manually in the texts and shows that this sentence is a section boundary. The segmentation procedure presented in this chapter required the output containing the links only, but the segmentation routine used in the next two chapters (see section 6.11 on page 302)) needed the sentence-numbered text as well.

Limitations `Words` is a simple program which presents several limitations. The main one concerns the component dealing with the identification of repeated strings, more specifically the fact that this component picks up repetition between identical strings only, and as such it implements only a small portion of the model of lexical cohesion as proposed by Hoey (1991b). The component dealing with altering the input text prior to the identification of the repetitions also has some limitations. Since the deletions made by the stemming module are blind, several errors may occur during its execution. For example, the deletion of all word-final `-ing` strings would cause a word such as `sing` to be replaced with `s`, and `singing` would be substituted by `sing`. In this case, if both `sing` and `singing` were present in the same text, the program would not compute the repetition between these two words. The lemmatisation module presents similar problems; a link would be counted, for example, between `saw` (the tool) and `saw` (past of 'see'), if all instances of `saw` were replaced with `see`.

Some of these problems might be avoided if a different approach to the identification of repetition had been used. For instance, in `abridge` (Hoey and Wools, 1995), a program which also implements aspects of the model of analysis proposed by Hoey (1991b), a pattern-matching algorithm is used whereby strings are compared for the number of letters in sequence that they have in common; if the number of letters exceeds a certain threshold (usually five), a match is declared. In this way, `abridge` enables the user to identify the repetition between `president` and `presidency`, for instance, since the two strings have a sequence of eight characters in common. Adopting this approach might have been a more satisfactory alternative to the stemming and lemmatization modules in `words`, but at the same time it still would not have allowed for the identification of the similarity between `see` and `saw`, or `car` and `vehicle`. What prompted the adoption of the exact string approach to the identification of repetition as opposed to the letter matching strategy in `abridge` is that the latter method appeared to be much more taxing on memory resources, and its implementation depended on programming skills which were not available at the time `words` was devised.

The characteristics of `words` described above represent the best compromise, under the circumstances, between efficiency and coverage. It would have been better if a program could have been developed which identified all of the different types of links that Hoey (1991b) describes, but this was not possible given the state of the art in computing and the resources available for the research. Given the state of the art in computing a few years ago when the program was being conceptualized, access to thesauri, lexical databases, and tools for lemmatisation, for example, was restricted because these resources were still experimental and not available to the public.

Computer and manual analysis

It was argued above that `words` allowed for the links in the example data to be detected. A test would be needed to estimate what proportion of all the existing links in a text a program such as `words` is capable of detecting. Ideally, the comparison would be carried out against an *expert* analysis, that is, a very comprehensive manual analysis which incorporated a wide range of lexical cohesive links. One such expert analysis can be found in Hoey (1991b, pp.76-99) where the initial 16 sentences of 'Masters of Political Thought', an academic textbook on political philosophy, are analysed in great detail for lexical links. Hoey's (1991b) analysis will therefore be used as a basis for comparison with the computer-based analysis as provided by `words`.

The analysis provided by Hoey (1991b, pp.86-87) includes the following kinds of links: simple lexical repetition, complex lexical repetition, simple mutual paraphrase, simple partial paraphrase, substitution, co-reference, ellipsis, and deixis. Ideally, a computer program should be able to identify all of these links as well; realistically, the range of links detectable by computer is not as wide. More specifically, substitution cannot be detected on unannotated text without the support of additional software capable of anaphora resolution, and even so the results are not 100% accurate. Deixis and ellipsis present further problems for automatic recognition and they can only be fully detected if the text is manually annotated with codes beforehand. Thus, these three kinds of links should not be considered for comparative purposes since computer-based analysis is by definition as yet ill-suited for recognizing them.

Further, Hoey (1991b, pp.86-87) marks other links as 'arguable', and naturally these should be excluded as well. However, some cases are arguable because they are discourse external, as for example the simple repetition links formed with the word 'reader' as in sentences 1 and 12:

[1] What is attempted in the following volume is to present to the **reader** a series of actual excerpts ... *[12]* In commending the writings which follow to the **reader**'s attention, ...

Although arguable, the repetitions of 'reader' would be identified by the computer with ease, hence these were not excluded from the comparison. The total number of valid links for comparative purposes as found by Hoey (1991b, pp.86-87) in the introductory sentences of 'Masters of Political Thought' is thus 95.

In order to know what share of the total of 95 links would be detected by computer analysis, the sixteen sentences analysed by Hoey (1991b) were entered into the computer and run through `words`. The control files input into `words` were not finely tuned for this particular analysis, that is, the control files instructing the program on how to lemmatise and stem words, remove stop words, and deal with pronoun references and synonyms were the standard files that had been developed so far. The results are presented in table 5.4.

Analysis			Total
	Manual		
Computer	Yes	No	
Yes	48 (50.5%)	25 (100.0%)	73 (60.8%)
No	47 (49.5%)	0 (0.0%)	47 (39.2%)
Total	95 (79.2%)	25 (20.8%)	120

Table 5.4: Computer and manual analysis of 'Masters of Political Thought'

Of the 95 links detected by manual analysis, about half were picked out by the computer (48, or 50.5%). The proportion of links detected by the computer could have been larger if the control files that were fed into `words` had been finely tuned, that is, adapted to the features of this particular

text. The proportion could have been larger as well if the files controlling the handling of the text (stemming, lemmatisation, etc) had been more complete (see p. 336 for a discussion of these issues in subsequent work).

The computer detected several links which had been rejected by the manual analysis. For instance, the computer picked up the link for the repetition of 'selected' in sentences 1 and 6:

[1] What is attempted in the following volume is to present to the reader a series of actual excerpts from the writings of the greatest political theorists of the past; **selected** and arranged so as to show the mutual coherence of various parts of an author's thought and his historical relation to his predecessors or successors; and accompanied by introductory notes and intervening comments designed to assist the understanding of the meaning and importance of the doctrine quoted. *[6]* Very often after a long passage has been quoted a single point has been **selected** for comment; and sometimes this point has been **selected** not because it was the most important, but because it was one which I had something to say.

The repetition of 'selected' was not considered by Hoey (1991b, pp. 56-67) because it failed to fulfill the shared context criterion, according to which two items are considered to form a link if there is something recognizable in the immediate context of the items which showed that the two items were talking about the same object or situation. This is not the case of 'selected', given that the first mention of 'selected' refers to the selection of texts, whereas the second mention refers to the selection of points for comment. In this way, the repetition of 'selected' is not text-forming but chance repetition (Hoey, 1991b, p.56). Context is broadly defined in Hoey (1991b), and he admits that there is 'plentiful scope for dispute over the ways [contextual] questions might be answered' (Hoey, 1991b, p.57). The contextual criterion as proposed by Hoey (1991b) approximates the chain interaction criterion of

Hasan (Hasan, 1989; see discussion in section 4.3.5 on page 140).

In all, the computer found twenty-five extra links. When these are taken into account, the total link count in the text rises to 120, and so does the share of links detected by computer: 60.8% (73 out of 120); by contrast, the share of links picked out by manual analysis is no longer 100%, but 79.2% (95 out of 120). The complete listings of links found by each analysis are presented in appendix 4 on page 457.

It must be stressed that the number of links detected by computer analysis could be improved by using appropriate stemming, inflectional, and derivational rules. These could have been built into `words` but they were not because of a lack of time and resources.

In summary, the analysis of lexical links by computer detected the majority (60.8%) of the lexical links in the text. Yet, the computer did not detect the same range of links which the manual analysis did.

In conclusion, a computer-based analysis is limited in comparison to a manual analysis in that it ignores a number of links. The main advantage of a computer-based analysis, however, is that it makes it possible to analyse long texts reliably. In other words, although the computer misses links and can make mistakes, it will miss the same links and make the same mistakes no matter how many texts it has to analyse; unlike humans, it will not 'get tired', and it can therefore be trusted to be more consistent in tedious jobs. Without the computer it would be very difficult for anyone to locate the links consistently in, for example, a 150-sentence text, whereas for the computer this would be a trivial task. Thus, despite its limitations, the computer is needed in such tasks because it can perform jobs which the analyst cannot. This compensates for the fact that the computer cannot deal adequately with the shared context criterion. Hoey (1991b, p.57) himself admits that '[contextual questions] may be valuable in manual analysis but

they are really no use for automatic analysis'. As a result, Hoey has largely abandoned this aspect of his model, on the grounds that it is not easy to operate (Hoey, personal communication). The manual analysis therefore did not miss any links, and in this sense it was not wrong. Nevertheless, the manual analysis may perhaps have been wrong in its principles that it should not have included this restriction.

With the development of the `words` program all the guidelines for the present pilot study have been followed. Now it is possible to test how the procedure can segment a text. This is reported on in the sections which follow.

5.4.7 Data

The data for the present pilot study were twenty-five encyclopedia articles. The texts were obtained from the 1995 version of Encarta on CD-ROM. The reason why encyclopedia articles were chosen is essentially practical. For one thing they are easily available – and without typing or scanning errors. For another, they all contain several section divisions, which suggests sectioning is an important generic characteristic. The corpus used here is a random sample of texts from the pool obtained by searching for 'countries of the world'. Since the texts are all about countries, the section headings in them are similar ('population', 'economy', 'government', etc.). This adds to the comparability of the texts.

5.4.8 Segmentation of example text

In order to illustrate how cluster analysis is meant to segment these texts, a detailed analysis of a single text will be provided. The example text is about San Marino, and was chosen at random. The San Marino text is reproduced in appendix 5 (p.463). The San Marino text was run through the `words`

program, which computed the 142 individual links in the text (part of the output is reproduced in figure 5.19).

The first step in the actual segmentation of the San Marino text was the determination of the number of clusters in it. This was done by examining the Cubic Clustering Criterion (CCC) values for it. The values of CCC obtained from running the FASTCLUS procedure through the data are shown in table 5.5 on the next page.

The first impression gained from observing the values of the CCC statistic in table 5.5 was that they are all greater than 2 or 3, which indicates good clusterings (Sarle, 1983, p.49). The other important characteristic of the distribution of CCC values is the occurrence of peaks. The highest peaks are for the following number of clusters: three, eight, twelve, thirteen, fourteen, and fifteen. The best choice is not simply the highest peak, but the highest *local* peak, that is, a peak followed by a low valley. In order to locate the local peaks, the CCC values were plotted against the number of clusters (Sarle, 1983, p.49). The plot is displayed in figure 5.20 on the following page. Note that the horizontal numbers (the x-axis) refer to the number of clusters, and not their position or the number of sentences for each solution.

Two peaks are prime candidates for local peak: we have either three clusters or eight; both look about the same height on the chart. The valleys following each of these peaks are indicated in the chart in figure 5.20. The

```
0001 0002  0001x0002republic-republic
0001 0020  0001x0020republic-republic
0001 0022  0001x0022republic-republic
0001 0003  0001x0003Italy-Italy
0001 0021  0001x0021Italy-Italy
0001 0019  0001x0019Rimini-Rimini
```

Figure 5.19: Partial `words` output

Clusters	CCC
2	11.049
3	13.254
4	10.068
5	8.029
6	10.447
7	11.959
8	13.362
9	9.463
10	10.947
11	12.179
12	15.939
13	15.568
14	14.089
15	14.439

Table 5.5: Values of CCC for the San Marino text

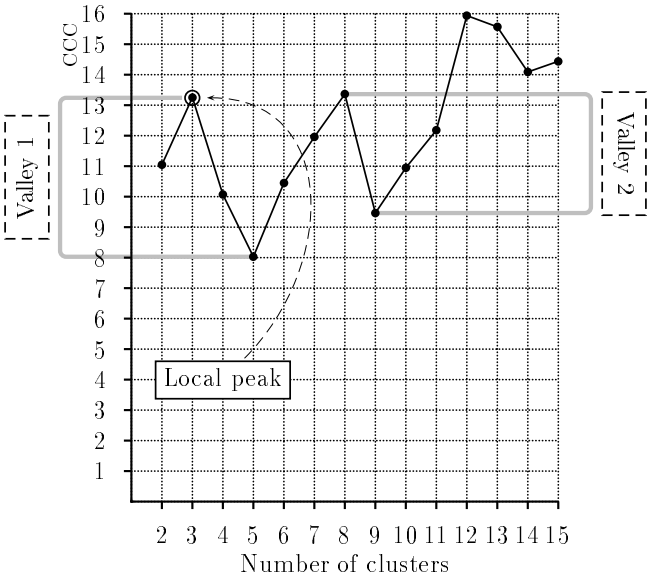


Figure 5.20: Plot of CCC values for the San Marino text

valley after the three cluster solution is labelled 'valley 1', and the valley after the eight cluster solution is shown as 'valley 2'. Lines connecting each of the two peaks to their respective valleys were also drawn in the chart to show the depth of each valley. A quick perusal of these lines indicated that valley 1 was lower than valley 2, and therefore the peak for the three cluster solution was in fact a local peak. Thus, the local peak for the three cluster solution suggested that there were three clusters in the data.

The next step is the location of the three clusters in the text. This is accomplished by plotting the members of each cluster against the sentences in the text, as in figure 5.21. Clusters 1 and 2 are well apart, therefore they will be taken to represent one distinct segment each. Cluster 3 is problematic in that its members appear nearly across the whole length of the text. The decision could be taken to ignore cluster 3 since it is largely overlapping, and overlapping segments are not desired. In addition, the distribution of clusters as shown in figure 5.21 suggests that `fastclus` segmented 1's from 2's in a much more obvious way than it segmented 1's from 3's or 3's from 2's, and therefore cluster 3 could be considered fictitious. In other words, the effect of ignoring cluster 3 would not be great since it would not make much difference to the segmentation, given that the main divisions would still be preserved, namely at the end of cluster 1 and at the beginning of cluster 2.

However, if cluster 3 were ignored, there would be a gap in the distribution of clusters between sentences 11 and 16, which was undesirable. Moreover,

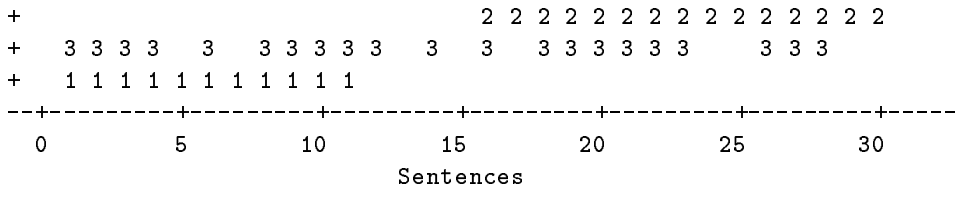


Figure 5.21: Distribution of clusters for the San Marino text

the three cluster solution was arrived at after careful examination of the values of the CCC statistic, and therefore it would be counterproductive to compute a stopping rule statistic and then override its judgement. The whole point of using a stopping rule such as CCC is that it should provide the number of clusters thus allowing for automatic segmentation analysis which is the aim of this pilot study. The best choice was therefore not to ignore the third cluster and settle for a segmentation into three segments. The third segment was allocated to that space between clusters 1 and 2 which was occupied by cluster 3. The numbers assigned to the clusters by `fastclus` were rearranged to reflect the natural order of segments in the text; hence, cluster 1 remained unchanged as segment 1, cluster 3 became segment 2, and cluster 2 became segment 3. As a result, the segment divisions in the San Marino text were placed in the positions indicated in table 5.6.

5.4.9 Performance

There are four section boudaries in the San Marino text, which were as follows: 'Introduction', in sentences 1 and 2; 'Land and Population' from sentence 3 to sentence 9; 'Economy and government', from sentence 10 to 17; and 'History', from sentence 18 to the end. The segments found in the text are shown in table 5.6.

By lining up the text segments and the section divisions, the diagram in

Segment	Sentences
1	1 through 11
2	12 through 14
3	16 through 27

Table 5.6: Segments in the San Marino text

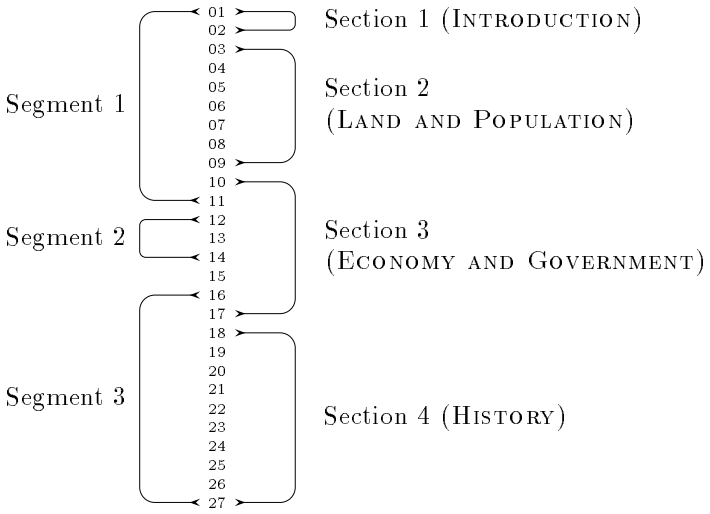


Figure 5.22: Comparison of segments and section divisions in the San Marino text

figure 5.22 is obtained. As in previous pilots, the vertical column of numbers in the centre of the diagram represents sentence numbers, the loops to the left indicate the portions corresponding to segments, and the loops to the right of the sentence numbers show the portions of the text corresponding to the sections. There are no matches apart from sentences 1 and 27, the first and last sentence of the text respectively. Unlike in pilot studies 1 and 2, where segment boundaries coinciding with the

Chapter 6

Development of the Link Set Median Procedure

In this chapter a new investigation into segmentation is presented. The results obtained so far in the pilot phase of the research project are discussed and placed in the context of the investigation as a whole. Then the development of the *Link Set Median* procedure is reported and the procedure is applied to a collection of texts.

6.1 Introduction

In pilot study 3, a procedure for segmenting texts was developed which provided segmentation of a large number of texts without human assistance. The segment boundaries were identified and placed following statistical information. The actual performance of the procedure was low in comparison to the pilot studies. In view of the poor performance, pilot study 3 concluded with the need for the development of a new procedure.

In the development of a new procedure, the aims to be pursued for the research into segmentation as a whole still apply, namely extensive cover-

age, inductive orientation, and objective evaluation. In practical terms, the three aims call for a computer-aided investigation into how lexical-cohesive segments match section divisions. In short, a computer-aided investigation provides the necessary means for the segmentation to be both comprehensive and inductive, while checking the fit between segment boundaries and section boundaries allows for the performance of the segmentation to be evaluated according to an objective criterion.

6.2 Goals

The major goal of the investigation reported in this chapter, as in pilot study 3, is the development of a procedure for unconstrained segmentation. In addition, the segmentation procedure must not depend on human intervention. This is essential because the major aim of the research is to investigate a large number of texts, which would not be feasible if the analysis was manual.

The goals of the present investigation are the same as those for the pilot study 3:

Automatic computation of cohesion In the new procedure, lexical cohesion must be computed automatically

Automatic placement of segment boundaries In the new procedure, segment boundaries must be placed without human intervention

Capability to handle several texts The new procedure must be efficient enough to be applied to several texts

6.3 Alternatives

Two alternatives present themselves at this stage. The first would be to try to improve the procedure developed for pilot study 3 by changing some of

the components of the procedure. Two key components of the procedure are the choice of clustering method and stopping rule. The performance of the procedure crucially depends on these two components, and so it would appear that if a substantial change is to be made in the procedure, it would have to include choosing a different method and a different stopping rule. However, changing these choices would not be an informed decision, since both the method and the stopping rule have been selected on the basis of their previous performance, adequacy to the data, and availability. No other choice of method or stopping rule would be as adequate. Moreover, in practice there is very little room for change, since there are no other stopping rules in the statistical packages available for use in this research project.

The second alternative is to devise a new procedure. In view of the difficulties involved in trying to improve the procedure presented in pilot study 3, this is a more realistic possibility. Importantly, in devising a new procedure insights from pilot study 3 can be utilized. Thus, the choice of devising a new procedure instead of adapting the previous one does not preclude using the knowledge gained by developing the previous procedure and thus presents itself as a better alternative.

A key characteristic of the procedure developed for pilot study 3 was that the segmentation was provided by a statistical procedure, namely cluster analysis. This decision was taken first because cluster analysis offers objective ways of partitioning a data set into smaller groups of observations, which is intuitively analogous to text segmentation. Furthermore, cluster analysis seemed to be able to handle the representation of lexical cohesion in matrices.

However, the bases for the choice of cluster analysis can be rethought. First, cluster analysis did not return a successful segmentation of the data. Therefore while in theory cluster analysis seems adequate for segmentation, in practice the results speak against it. Second, matrices are simply a con-

venient representational device for the lexical cohesion. In fact, the lexical cohesion between sentences of a text can also be represented by other means (nets, lists, etc.) and therefore segmentation does not presuppose matrix handling which in turn does not presuppose cluster analysis. By extension, if cluster analysis is the reason for utilizing statistical procedures, and if cluster analysis is not adequate, then there is no logical reason why segmentation should depend on statistical procedures.

Thus, the segmentation procedure developed here will not be constrained by the range of available statistical techniques. Instead, in developing a procedure the initial strategy will consist of tackling the individual problems which arise from attempting to segment the continuum of lexical cohesive relations between pairs of sentences.

6.4 Sentence similarity

As explained in chapter 1 (p.16), a segment is a sequence of at least two contiguous sentences displaying similarity at the level of lexical cohesion. In this manner, the basic tasks of any segmentation procedure would involve (1) assessing the similarity between contiguous sentences, and (2) assessing the dissimilarity between contiguous sentences. In other words, the decision to place a segment boundary would depend on ensuring that the sentences within a particular segment are more similar to each other than they are to the other sentences in other segments. This course of action is similar to that followed by Longacre and Levinson (1978, p.118), whose strategy for displaying the constituents of a discourse consisted of (i) grouping 'together those sentences that seem to naturally belong together' and (ii) dividing 'the discourse at those points at which it seems to naturally separate'.

A problem with assessing lexical cohesive similarity across contiguous

sentences is that many sentences which readers would normally regard as being similar and thus belonging to the same segment do not share any lexical links. As Thompson (1996, p.147) argues:

> all language users are generally predisposed to construct coherence even from language with few recognisable cohesive signals, if they have reason to believe that it is intended to be coherent.

The same point is made by Brown and Yule (1983) for whom disconnectedness between two adjacent sentences 'must be positively indicated' otherwise the two sentences are interpreted as being related. Goutsos (1996a) elaborates on the problem of continuity of similarity and suggests ways of showing how discontinuity is introduced in the text to show text internal boundaries (see discussion above in section 2.4.4, pp.63ff.). Further, Berber Sardinha (1995e) provides evidence that in some texts (business reports) bonding does not normally occur between adjacent sentences, even though they are clearly intended to be coherent.

In sum, readers would 'have reason to believe' two contiguous sentences are to be interpreted as coherent even where there are no lexical items shared between the sentences. The possibility of coherent pairs of adjacent sentences not sharing lexical items poses a problem to segmenting texts by computing lexical cohesion. The problem lies in the fact that if a segment is considered to be a contiguous sequence of adjacent sequences, then the sharing of lexical items between adjacent sentences cannot be relied upon as a valid criterion for showing that any two adjacent sentences belong in the same segment since it is possible that they will not share lexical items.

6.5 Link set

Instead of looking at the similarity between pairs of adjacent sentences, an alternative would be to look at the similarity between all the sentences with

which each adjacent sentence shares lexical items. This might provide some indication of the degree of similarity between two sentences even in cases where there are no lexical links shared between the adjacent pair. To achieve this, the concept of *link set* must be introduced.

The set of sentences with which each sentence has links can be seen to form a link set. For instance, if sentence 1 has three links with sentence 6 and two links with sentence 4, then its link set[1] would be 4,4,6,6,6, that is, the number 4 is entered twice, one for each link with sentence 4, and the number 6 is entered three times, one for each link with sentence 6. In other words, the figures indicate the sentences with which a particular sentence has links, and the number of times each figure features in the set indicates the number of links shared. The set is ordered in sequential order in the text because this makes it easier for the calculation of a central tendency for the set, as explained below. Other than that, the order of the elements does not matter, that is, the previous set could be represented, for example, as 6,6,4,6,4.

If both sentence 1 and sentence 2 have one link each with sentences 10, 11 and 12, but not with each other, then the fact that they have identical link sets (10,11,12), i.e. they have links with the same sentences across the text, can be used to reveal the extent to which they are similar. Thus, going along with the same example, if sentence 3 had a link set 10,11, then a case could be made that sentences 1 and 2 are more similar to each other than they are to sentence 3. In other words, by comparing link sets it becomes possible not only to assess the similarity of adjacent sentences without depending on the existence of links between the two sentences being compared, but also to obtain some measure of the degree of similarity amongst sentences. If only adjacent sentences were compared for similarity, the only possibilities would be 'there is similarity' or 'there is no similarity'. By comparing link sets,

[1] For convenience, link sets will be displayed in a box henceforth.

the measurement of similarity is not reduced to this dichotomy, rather it is measured on a cline of 'more similar' to 'less similar'.

In addition to the more practical reasons adduced above for the implementation of link sets, there is also an important linguistic motivation for link sets. Given that cohesion is a measure of topic shifts (Hoey, 1991b) and segmentation (e.g. Hearst, 1994a; Kozima, 1993a), the simplest measure of where the cohesion is would be to see every cohesive item as a measure of similarity between two sentences. Lexical cohesion is a measure of similarity (Hoey, 1991b), and therefore similarity can be assessed by looking at the lexis shared among sentences. Since each lexical item is a separate measure of similarity, if there are three lexical items shared there are three points of similarity, hence the similarity can be recorded three times. The notion of link set as a record of similarity is therefore convenient in that it enables the researcher to observe the degree of similarity between two sentences. As a record of lexical similarity, the link set is not entirely different from a repetition matrix (Hoey, 1991b, see discussion above on p. 159), since in a sense the link set can be seen as a 'flat matrix', where the links are not laid out two-dimensionally in rows and columns but one-dimensionally in a single row. The link set preserves the kind of information that is recorded in a matrix but makes the information considerably more convenient to process for the purposes of segmentation.

A problem with comparing link sets is that it is hard to compute the degree of similarity between two sets. Although it is possible to count the number of matches between sets, it is still problematic to decide what will count as a match. For instance, if the link set for sentence 1 is $\boxed{3,5,7}$ and for sentence 2 $\boxed{4,6,8}$, there will be no exact matches between them, yet the two sets are clearly related in that the sentences in them are only one sentence apart from each other. Even if a cut-off point is decided regarding

what would indicate the greatest difference that would still count towards a match, there would still be the problem of how to handle sets of different numbers of elements. For example, if the link set for sentence 1 were again [3,5,7] but for sentence 2 it were simply [6] then how would the two sets be compared? Would 6 be compared to 3, 5 and 7, or just 7? In case 6 is compared to 7 and the difference of 1 is still regarded as a match, what would be made of 3 and 5? Should these be disregarded or should they be used to compute some sort of *dissimilarity* measure?

There are no simple answers to these questions, mainly because matching sets of numbers is in itself a complex task regardless of the application. An alternative would be to compute similarity not between sets but between two key elements, one from each set. By key element is meant a member of the set which can be taken to be a representative of the central tendency of the set as a whole.

Three measures of central tendency exist which could be employed to represent a link set: mode, mean, and median. The mode is the most frequent element of a distribution. In a link set such as [1,1,10,11], for example, the mode would be 1 since it appears more often. The mode has a serious drawback though, which is that since it is unaffected by the remaining elements of the distribution, it can 'camouflage important facts about the data' (Wimmer and Dominick, 1991, p.204). For instance, a link set such as [1,1,99,100,101,102,103] would have a mode equal to 1, yet most of its elements are distributed around 100.

The mean represents the average of a distribution, that is, the sum of the elements divided by the number of elements. In the previous hypothetical link set, the mean would be 72, or $1+1+99+100+101+102+103=507\div7$. A major problem with the mean is that it is affected by extreme scores, or outliers, which have the effect of dragging the mean in their direction. For

example, if the element 1000 were added to the previous link set, the mean would then be 188, or 1507÷8.

The median is the midpoint of a distribution so that 50% of the elements of the distribution lie on either side of the median (Woods et al., 1986, p.19). For instance, for the link set 1,1,99,100,101,102,103 the median would be 100, since $\frac{1}{2}$ of the elements lie above it (i.e. 1,1,99) and $\frac{1}{2}$ of them lie below it (i.e. 101,102,103). For even-numbered distributions, the median is obtained by summing up the two middle elements and dividing them by 2 (Wimmer and Dominick, 1991, p.204). For instance the middle elements of a link set such as 1,1,99,100,101,102,103,1000 are 100 and 101, thus the median is 100.5 (100+101=201÷2).

The choice of a measure of central tendency depends essentially on the type of data being described. In representing the sentences, the point of referring to the first sentence as '1' and to the sentence immediately after it as '2' is to show that the two sentences are ordered, that is, sentence 1 precedes sentence 2 in the text. The actual denomination of each sentence is not important, thus the first sentence could be referred to as 'A' or 'alpha' or any other member of an ordered set. Running numbers are preferable because they form the most traditional set of ordered elements. The fact that the order of the elements is important in representing text sentences indicates that the data in link sets are ordinal. For ordinal data, the median is the most appropriate measure of central tendency. Thus, the median is the most adequate measure of central tendency for link sets.

The usefulness of link set medians is that they provide a way of comparing the similarity of the lexical cohesion between two sentences in running text. The question which arises is how to calculate the similarity or dissimilarity between two link set medians. There is no established means for assessing the similarity between pairs of medians, and therefore a method needed to

be created for that purpose. Since the medians are extracted from the lexical cohesion computed in the text, the criterion for assessing similarity must also be sought in the text.

6.6 Example 1

In order to find a method for assessing link set median similarity, it is necessary to consider a set of (hypothetical) data. For instance, table 6.1 displays six sentences, their link sets and respective medians. For the sake of simplicity, it is assumed that sentences share only one link with any other sentence. The question to be posed here is how similar or dissimilar is each median to their neighbour? For example, how similar is '2' (the median for sentence 2) to '2.5' (the median for sentence 1)? The difference between the two medians is just .5, which intuitively is small since it indicates that their individual link sets differ by less than one sentence with respect to their midpoint. However, there is nothing inherent in a .5 difference which guarantees that the difference is small enough to indicate similarity.

One way to provide an answer is to calculate the average difference across the text and then compare each individual difference to the average difference. First, individual differences must be computed, that is, the difference

Sentence	Link set	Median
1	2,3	2.5
2	1,3	2
3	1,2	1.5
4	5,6	5.5
5	4,6	5
6	4,5	4.5

Table 6.1: Hypothetical data 1: Medians

between each median and its predecessor or follower is calculated. In principle, the difference can be calculated in relation to either the sentence's preceding or following median, since in any case one sentence in the text will be without a difference, either the very first or the very last sentence. For this particular investigation, it was decided to compare each sentence median to its immediate predecessor. It is best to disregard the sign of the difference since there is no reason to distinguish positive and negative differences. Second, the average difference is obtained by summing up the individual differences and dividing them by the number of (non-zero) differences. Finally, the average difference is compared to each individual difference. This yields a categorization of each difference as being either higher or lower than the average. Those differences which are higher than the average can be considered to indicate dissimilar medians.

To illustrate, table 6.2 displays the differences for the data in table 6.1. In this particular case, the average difference is 1.2, or .5+.5+4+.5+.5=6÷5. By contrasting each individual difference to the average difference of 1.2, the only median difference which is greater than the average is 4 (for sentence 4). This suggests that sentence 4 is dissimilar in its link set median to sentence 3. The link set median for sentence 3 is $\boxed{1,2}$, while that for sentence 4 is $\boxed{5,6}$. These two link sets are intuitively different, and the procedure described here

Sentence	Link set	Median	Difference
1	2,3	2.5	–
2	1,3	2	.5
3	1,2	1.5	.5
4	5,6	5.5	4
5	4,6	5	.5
6	4,5	4.5	.5

Table 6.2: Hypothetical data 1: Median differences

has correctly identified these two sentences as being dissimilar. Without the support of the average difference, the decision to differentiate between link sets 1,2 and 5,6 would have been arbitrary.

Since sentence 4 is dissimilar to its predecessor, it would be a suitable point at which to place a segment boundary. By being a boundary sentence, it would mean that sentence 4 would initiate a new segment since it is dissimilar to its predecessor, sentence 3. This would create two segments in this hypothetical text. The first would run from sentences 1 through 3, and the second from sentences 4 through 6.

Graphically, it is possible to represent the data in tables 6.1 and 6.2 in a line chart (see figure 6.1 on the following page). This would assist in appreciating the changes in median difference from sentence to sentence. The median difference for sentence 4 shows up as a *peak* because it is higher than the average median difference, as figure 6.2 on the next page illustrates. 'Peak' seems to be a useful metaphor and therefore it will be applied henceforth to denote those sentences whose median differences are higher than the average median difference.

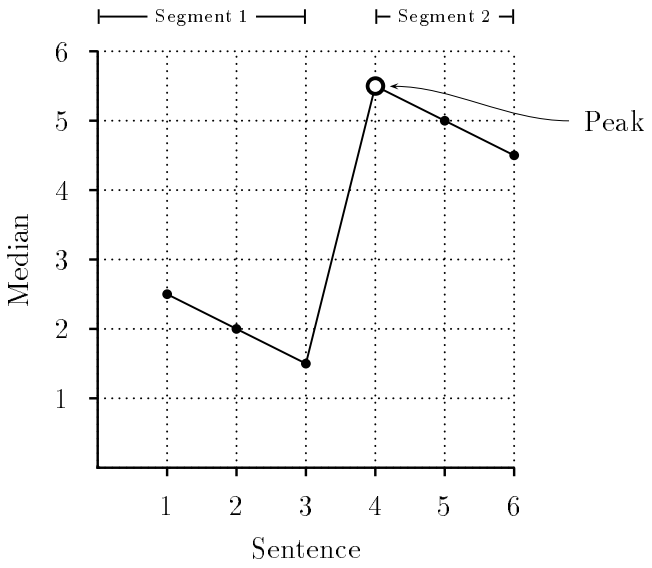


Figure 6.1: Hypothetical data 1: Line chart

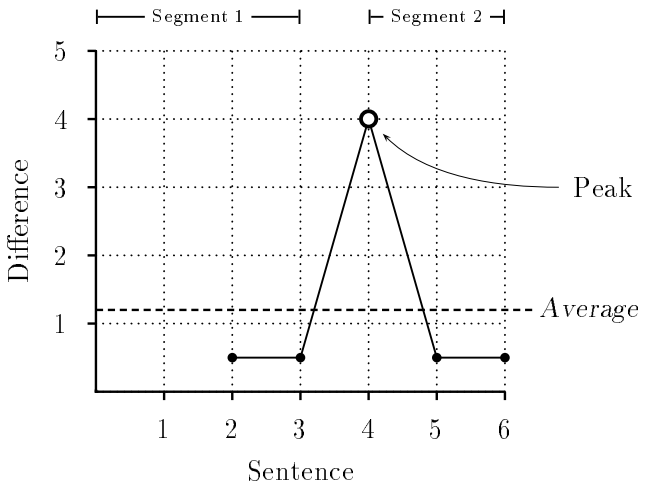


Figure 6.2: Hypothetical data 1: Median difference and peak

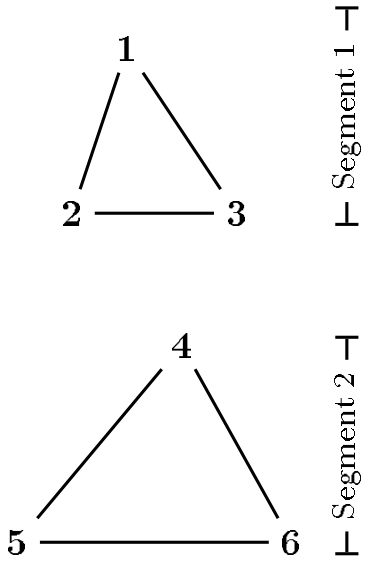


Figure 6.3: Hypothetical data 1: Net

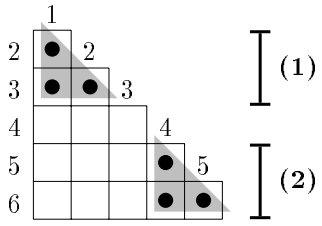


Figure 6.4: Hypothetical data 1: Cluster triangles

It is also possible to represent the hypothetical text in a more familiar schematic form: a net (Hoey, 1991b). By translating the link sets into points in a net, the hypothetical text can be represented as in figure 6.3 on the preceding page. Visually, it is possible to discern two segments, the first comprising sentences 1, 2, and 3, and the second sentences 4, 5 and 6. These are exactly the same segments identified by using median differences. This also suggests that the median difference procedure made the right choice by placing the segment boundary at sentence 4.

Finally, another graphic representation which can be produced for hypothetical data 1 is a matrix (Hoey, 1991b; see also the previous discussion in section 5.2.3 on page 195). In the case of the matrix, what would be particularly interesting would be to see to what extent the segments identified here map onto possible cluster triangles (see pilot study 2 above, section 5.3, pp.206 ff). A matrix representation of hypothetical data 1 is shown in figure 6.4 on the preceding page. The data form two distinct cluster triangles which correspond to the two segments identified by the procedure.

6.7 Example 2

One of the reasons for utilising link sets is that they would assist in showing the relatedness of adjacent sentences even if they did not share any links between them. However, in hypothetical data 1 the same segmentation could have been achieved if one had looked simply for sentences which did not link with their neighbours since this would have identified sentence 4 as a segment boundary. This would have proved right; therefore there would be no need to resort to link sets and median differences to carry out the segmentation of the text.

Sentence	Link set	Median	Difference
1	2,3	2.5	–
2	1,3	2	.5
3	1,2,4	2	0
4	3,5,6	5	3
5	4,6	5	0
6	4,5	4.5	.5

Table 6.3: Hypothetical data 2: Median differences

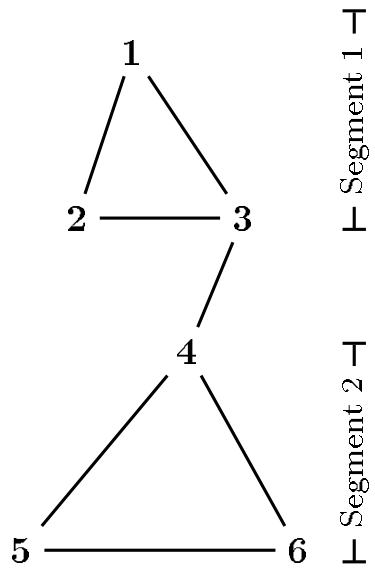


Figure 6.5: Hypothetical data 2: Net

To illustrate the superiority of the proposed method, another set of data must be used in which all sentences are linked with their neighbours. In the data in table 6.3 on the preceding page all sentences are linked to their immediate neighbours. This becomes more apparent in the net in figure 6.5 on the preceding page. In the net, two segments are easily spotted: from sentence 1 to 3, and from sentence 4 to 6, that is, the same ones as in hypothetical data 1. In fact, the only difference between the two data sets is the addition of a link between sentences 3 and 4.

By applying the same procedures explained above, the average median difference for the data set is 1.3 (.5+3+.5=4÷3). The only individual difference which is greater than the average is again for sentence 4, which then becomes the segment boundary. Thus, two segments are identified, from sentences 1 through 3, and from sentence 4 through 6, that is, the same ones as for hypothetical data 1. The difference is that now it would have been impossible to place a segment boundary by checking the linkage between neighbouring sentences only. In this manner, the procedure based on median differences has proved to be robust enough not to be affected by the inclusion of the extra link.

6.8 Example 3

So far, the hypothetical sets of data have contained only one peak, thus there was only one possible segment boundary in each. However, there is nothing which prevents more than one peak from occurring in a text. Crucially, there is nothing in the procedure which deals with adjacent peaks. When at least two adjacent peaks occur, they form a peak cluster. Peak clusters are problematic because they create contiguous segment boundaries thus generating one-sentence segments, which would be incompatible with the

definition of segment as multi-sentence portions of text presented above. A mechanism must therefore be built in the procedure which deals with peak clusters.

The hypothetical data set presented in table 6.4 includes a peak cluster formed by the peaks at sentence 4 and sentence 5. Sentence 4's median difference is 7.5, while sentence 5's median difference is 3, both of which are greater than the average difference of 2.4 for the text (.5 + .5 + 7.5 + 3 + 1.5 + 1.5 + 2 = 16.5 ÷ 7). Figure 6.6 on the following page identifies the peak cluster. In dealing with peak clusters, the best solution is to choose that difference which is greater, in this case, 7.5. Sentence 4 can therefore be referred to as the *major peak* in the peak cluster, and can thus be selected as the segment boundary. As in the previous two examples, the choice of sentence 4 as a segment boundary clearly recovers the two visible segments in the data, as shown in figure 6.7 on the next page.

Sentence	Link set	Median	Difference
1	2,3	2.5	–
2	1,3	2	.5
3	1,2	1.5	.5
4	9	9	7.5
5	6	6	3
6	5,7,8	7	1
7	6	6	1
8	6	6	0
9	4	4	2

Table 6.4: Hypothetical data 3: Median differences

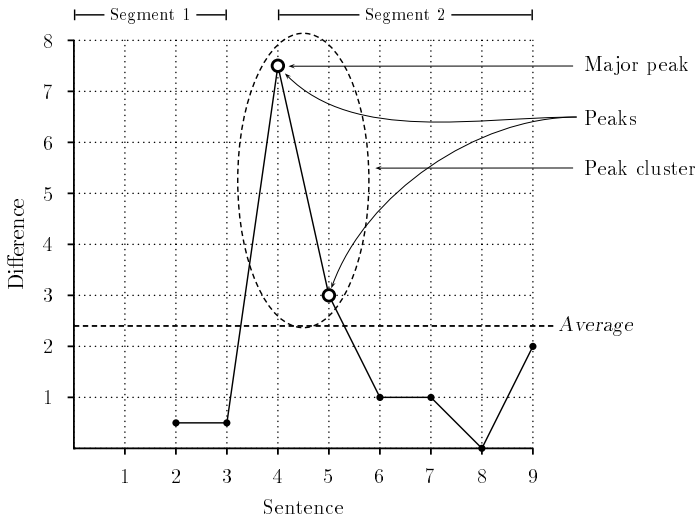


Figure 6.6: Hypothetical data 3: Line chart

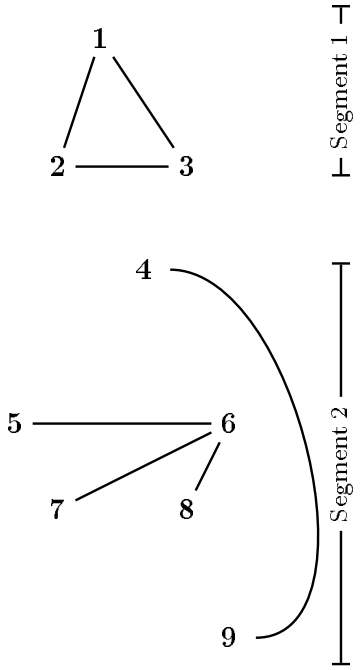


Figure 6.7: Hypothetical data 3: Net

One strength of the link set median (henceforth LSM) procedure developed here is that it is simple, consisting of straightforward arithmetic operations. This allows it to be implemented on the computer using standard statistical packages. SAS, for instance, offers a programming language in which the procedure as described here can be adequately implemented. Another of its strengths is its apparent robustness – the procedure developed here seemed capable of recovering the desired segments in hypothetical data. The next natural step in this investigation is the application of the LSM procedure to real data.

6.9 Data and procedures

The data for this investigation are the same as for pilot study 3, that is, twenty-five encyclopedia articles from the 1995 version of Encarta on CD-ROM (see section 5.4.7 above on p.262). This will enable a comparison of the segmentation results obtained here with those obtained in the last pilot study. The comparison is useful because it will indicate whether the LSM procedure shows any gains in performance.

The LSM procedure was implemented by first running each text through `words` (see description of the program in section 5.4.6 on p.245). `Words` output the links for each text which were subsequently processed in SAS (see section 6.11 on page 302.)[2]. The performance statistics were also output by SAS.

[2] I want to thank Matthew Zack, Jay Weedon, Sue Byrne, Bob Gallop, David Alderton, Tim Borryhill, and colleagues on the SAS discussion list and newsgroup for their help in writing SAS commands.

Random segmentation

One way to evaluate how well LSM performs is to compare it to segmentations of the same texts carried out by other methods. First, it seemed crucial to compare the results of LSM segmentation to *random segmentation*. This would provide an answer to the question 'is LSM segmentation better than chance?' In other words, if one were given a certain number of segment boundaries to insert in the corpus, what are the odds that he/she would find section boundaries? If the segmentation by LSM proves better in comparison to random segmentation, then this will provide some support for the principles upon which LSM segmentation is based. On the other hand, if LSM segmentation does not perform better, then there will be no evidence to suggest that the performance of LSM segmentation is better than chance.

Random segmentation was achieved by a routine written in SAS language which selected a given number of sentences from the data set as segment boundaries. The total of random segment boundaries was the same as the total LSM boundaries. The routine used for assigning random segment boundaries is included as part of the LSM routine, which is explained in section 6.11 on page 302.

Expert segmentation

Second, it is important to compare the results of the segmentation to an *expert segmentation*, or segmentation provided by an expert computer system. This would provide an answer to the question 'how good is the segmentation?' Theoretically, computational segmentation should achieve full precision and full recall; in practice however such marks have never been reported (see comparative charts in figure 5.13 on page 216).

An expert segmentation should therefore provide a realistic ceiling rate to indicate how well one can expect the texts to be segmented by computer.

It is important to bear in mind that the segmentation by LSM as presented here is not supposed to be an information retrieval system, and therefore the aim is not to compare two competing procedures but to evaluate the procedure being developed in the present investigation so that some claims about the relationship between lexical cohesion and text organisation can be made. These two comparisons will be provided below (see section 6.12ff).

Of the segmentation procedures reviewed in chapter 3, the one which reports the best levels of segmentation performance is *TextTiling* (Hearst, 1993, 1994b,a; Hearst and Plaunt, 1993) (see discussion in section 3.5, pp.109ff., and comparison in figure 5.13 on page 216). TextTiling is a fully developed segmentation algorithm used in information retrieval; it was designed to assist in extracting relevant texts from text databases. Importantly, the rationale behind TextTiling is also based on lexical cohesion, hence the comparison to LSM segmentation would be fair.

Where the comparison between LSM and TextTiling would not be fair, though, is with respect to the number of possible segmentation points in a text. As pointed out on p.111, TextTiling operates on the principle that paragraph breaks are the only possible segmentation points in a text (Hearst, 1994a, p.30). This principle restricts the number of possible places where segment boundaries can be placed, and as a result, it makes it easier for the system to place segment boundaries that match section boundaries. An analogy could be drawn with a different kind of software. Suppose there were a computer program that claimed to parse sentences so well that it was able to identify sentence boundaries extremely accurately, but despite that it would only make use of commas and full stops. So, when the program got within three or four words of a sentence break it would shunt itself up to the full stop and place a sentence boundary there, instead of three or four words behind. This would have the immediate effect of increasing the

chance of hitting a true sentence boundary; the match would however not be a result of the accuracy of the parsing, but of the reliance on a pre-existing segmentation. TextTiling utilises the same kind of fudge by insisting that segment breaks must occur at paragraph boundaries.

The criticisms levelled against TextTiling do not mean, though, that TextTiling is not a valid practical working tool, but it means that its value as a research tool is greatly diminished. In terms of the comparison between LSM and TextTiling, if it is found that LSM is capable of achieving a comparable level of performance, it will be a major achievement of the research presented here.

Since it was not possible to 'untweak' TextTiling so that it would consider placing segment boundaries within paragraphs as well, a method for 'deparagraphing' the data was tried. This method consisted of turning each sentence of the texts into a paragraph by simply inserting a blank line after each sentence, the blank line being the paragraph marker that TextTiling was built to recognize. In the texts modified in this way, therefore, there were as many possible segmentation points for TextTiling as there were for LSM. Unfortunately, TextTile behaved erratically when run through the modified texts: it crashed a few times, refusing to process certain text files, and it also returned texts without any segmentation, even though it had segmented them in their original format with intact paragraphs. For these reasons, the idea of using modified texts was abandoned, and the text files run through TextTile for the present investigation therefore contained the texts with their original paragraphing. These problems may have occurred because of the way the particular version of TextTile available for use in this investigation (see next paragraph) was compiled, but since that was the only version available, there was nothing that could be done to prevent TextTiling from restricting the number of possible segmenting places in the texts to the gaps between

paragraphs.

A version of TextTiling was installed on the University of Liverpool's Unix network[3] in order for the expert segmentation to be carried out. This particular implementation, called `tile`, is freely available online and is accompanied by a stop word file which contains several closed set words and general lexical words. This file was replaced with the stop word list used with `words`. This was done to ensure that the two programs were similar in the filters they applied to the input texts.

6.10 Boundary placement and matching

As in the pilot studies, two kinds of boundaries are considered in the analysis: section boundaries and segment boundaries. Section boundaries are those sentences which have a section heading and are therefore the onset of a section in the text. In this respect, the first sentence of the text is problematic because it marks the beginning of a section even if it has no section heading. Even if no sections as such are demarcated in the text, the first sentence can still be considered the beginning of a section simply because it is the beginning of a text. TextTiling makes use of this strategy and by default places the onset of a tile at sentence 1. Being a natural section boundary, it would not be fair to include the first sentence of the text as a valid section boundary. Thus the first sentence of the text was not included in the computation of the performance of any of the types of segmentation considered below.

Segment boundaries are those sentences chosen by each segmentation procedure to initiate a segment. In the case of LSM, it is important to distinguish between *provisional boundaries* and *final boundaries*. The former are those

[3] I want to thank Peter Kulawec for his help in compiling the source code.

sentences considered by the LSM algorithm as a possible segment boundary, whereas the latter are the definitive boundaries inserted by LSM. Each peak in the distribution of median differences is equivalent to a provisional boundary. By contrast, a final boundary is the best provisional boundary of a cluster. For random segmentation, the provisional boundaries are sentences chosen at random by the SAS routine; their total is the same as the total of LSM peaks. This secures a level playing field, since by making sure that random segmentation considers as many provisional segment boundaries as LSM, it has as many chances to place definitive segment boundaries and ultimately of finding section boundaries as LSM segmentation.

As argued on p.292, one reason why not all peaks can be final boundaries, is that adjacent peaks would create one-sentence segments, which would be inconsistent with the working definition of segment as a multi-sentence portion of text (see section 1.6 on page 16). Random boundaries can also be placed in adjacent sentences thus giving rise to the same kind of problem as peak clusters. The arbitrary solution would be to choose one boundary at random from a two-sentence peak cluster or one random boundary from a two-sentence random boundary cluster. The best solution is to treat a boundary cluster as a *boundary zone*. A boundary zone is therefore a sentence or contiguous group of sentences where peaks occur. This amounts to treating each boundary zone as contributing to only one final boundary. For segment demarcation purposes, the final boundary is positioned at the onset of the boundary zone.

Within a boundary zone a number of section boundaries can occur. But if a boundary zone marks the onset of only one segment, then the question remains of how many matches are allowed per boundary zone. The only logical option is to allow at most one match per boundary zone, thus avoiding the situation where one boundary could count as more than one match, which

would yield nonsensical precision rates of over 100%.

If there is more than one section boundary in a boundary zone, the question arises of which of these should be considered a match. For LSM segmentation, if the height of the peak where the section boundaries occur can be used as a disambiguation criterion, in these cases a match is counted for that section boundary which occurs at the highest peak in the boundary zone. For random segmentation, the earliest match can be computed. A detailed account of the matching decisions is presented in figure 6.8 on the next page.

LSM segmentation

1. Allow only one segment boundary per boundary zone;
2. Place segment boundary at boundary zone onset;
3. Compute at most one match per boundary
   (a) If there is only one match, compute that match;
   (b) If there is more than one match, compute only that match which occurs **at the highest peak**.

Random segmentation

1. Allow only one segment boundary per boundary zone;
2. Place segment boundary at boundary zone onset;
3. Compute at most one match per boundary
   (a) If there is only one match, compute that match;
   (b) If there is more than one match, compute only that match which occurs **at the earliest sentence**.

Figure 6.8: Matching algorithms

6.11 Segmentation algorithm

This section describes the structure of the segmentation routine written in SAS which formed the basis for both the LSM and the Random segmentation of the texts analysed in the main study[4]. There are two main components in the segmentation routine: one dealing with the actual segmentation of the texts, and the other providing the output of the results to a file. The segmentation component is the most important one, and will be described in more detail. This component is subdivided into four main parts: data input, the LSM segmenter, the Random segmenter, and a final module which computes the performance of the segmentation. The major steps in the routine are explained below in the order in which they were executed.

Segmentation component

Data input

1.*Read in each text with numbered sentences*, as provided by `words` (see p.256). The texts were formatted by `words` in such a way that the each sentence was printed on a separate line of text. The texts were annotated by hand to show whether a sentence was a section boundary or not, in which case those sentences which were section boundaries were preceded by the tag `sxbrk`.

This step is implemented by the following SAS command:

```
data temp; infile 'myfile.1a' delimiter='[]{} ';
input sent psent sec $ ;
drop psent;
if sec eq 'sxbrk' then sec=sent;
else sec=.;
run;
```

[4] A copy of the source code can be obtained from the author by writing to: R Paracatu 357 apto 52, 04302-020 São Paulo SP, Brazil.

A variable called `section` is created to store the information about whether each sentence is a section boundary (value=1) or not (value=0).
2.*Read in the list of links for each text* supplied by `words`. The format of the links listing is:

```
sentence_number_1 sentence_number_2 total_links repeated_word_1
                                                repeated_word_2 ...
                                                 repeated_word_n
```

The information for each sentence is read into the variables `sent1` and `sent2`. The following SAS code implements part of this step:

```
data tk001a; infile 'myfile.1b';  input sent1 sent2; run;
```

Because the listing does not include links between a sentence and those which preceded it, the data need to be read twice. This time, the order of the variables is inverted (`... input sent2 sent1; run;`). This will ensure that the link sets to be formed next will be complete.

LSM segmenter

1.*Create the link sets* by crosstabulating `sent1` and `sent2`:

```
proc freq data=temp noprint;
 tables sent1*sent2/list nopercent norow nocol out=temp2;
run;
```

2.*Compute the median for each link set.* This is accomplished using the `univariate` procedure in SAS:

```
proc univariate noprint; var sent2;
output out=temp3 median=median;
by  sent1; run;
```

The medians for each text are then saved into a separate file, in addition to the following information: text number, sentence number, and whether the sentence is a section boundary (1) or not (0). Procedure `report` is used to print the data to the file:

```
proc report nowindows noheader nocenter ps=1000;
 column text sent median sec; run;
```

3.*Disregard those section boundaries which occur in sentence 1.* This is accomplished by:

```
data temp; if sent1=1 then section=.; run;
```

4.*Calculate the median difference* by (1) subtracting the value of the current median from the immediately preceding one, and (2) taking the absolute value of that difference:

```
data temp (drop=nextsent); set temp (firstobs=1);
nextsent=median; set temp; change=nextsent-lag1(median);
if lag1(text) ne text then change=0;
abchange = abs(change);
run;
```

5.*Compute the average median difference* by calculating the mean of the medians for each text. This is accomplished thorugh procedure `univariate`:

```
proc univariate data=temp noprint;
 var change; output out=temp4
 mean=meandiff; by text;
run;
```

6.*Identify the provisional boundaries and boundary zones.* The identification of the former is accomplished by by locating those individual median differences which exceed the average median difference; the boundary zones are

then marked as boundaries occuring in adjacent positions. The implementation of this latter step takes many lines of code, but its core is the following:

```
data temp1;
 set temp; by text;
 retain group;
 if first.text then group=0;
 lastrise=lag1(risefall);
 if risefall>. and (lastrise=. or first.text)
  then group=group+1;
  else if risefall=. then delete;
 drop lastrise;
run;
```

7.*Insert segment boundaries* at the onset of each boundary zone. This step also takes many lines of code, but its main part is the following:

```
 data new; merge   temp1   temp2;   by text group;
   retain groupmx; drop min groupmx;
   if first.group then groupmx=0;
   if cut=min and not groupmx then do;
      minrval=min; groupmx=1;
 end; run;
```

8.*Compute matches* by looking for section boundaries within each boundary zone. More precisely, compute a match whenever `section` equals 1 within boundary zones.

Random segmenter

1.*Count the total number of LSM segment boundaries placed in the corpus* and store that value to be used as the total number of boundaries to placed randomly in the corpus:

```
 proc summary data=temp print; var group;run;
```

2.*Insert random segment boundaries.* The core of the random segmentation algorithm is the following routine, which chooses which sentences will be random segment boundaries:

```
data exact(drop=k n);
retain k &sample n;
if _n_=1 then n=total;
set country nobs=total;
if ranuni(747088789)<=k/n then
 do;
  output;
  k=k-1;
 end;
n=n-1;
if k=0 then stop;
run;
```

Computation of performance

1.*Estimate the total number of sections and boundaries in each corpus.* This is implemented by using procedure `summary`:

```
proc summary data=temp;
 class text;
 var group section cuts match
 rcut rmatch;
 output out=temp n=;
run;
```

2.*Calculate recall and precision rates.* The two performance rates are obtained for LSM and random segmentation separately. For each type of segmentation, the recall rates are computed for each text by dividing the total number of matches by the total number of sections, and the precision rates are obtained for each text by dividing the total number of matches by the total number of segment boundaries. This is achieved by:

```
data temp;
 crecall=match/section; cprec=match/cuts;
 rrecall=rmatch/section; rprec=rmatch/rcut;
run;
```

Output component

The performance rates are saved to a separate file using the `print` and `printto` commands:

```
proc printto new print='myresults.txt'; run;
proc print noobs width=min;
 var text sents section group cuts match crecall cprec
  random rcut rmatch rrecall rprec;
run;
```

6.12 Results

The results of the segmentation of the corpus by LSM are presented in table 6.5. The individual results by text are shown in appendix 8 (p.467).

In all, 841 provisional boundaries were inserted in the corpus, each corresponding to a peak in the distribution. The 841 boundaries were distributed into 430 boundary zones, each contributing one final boundary. Considering only the final boundaries, the ratio between boundaries and sections was 1.08 final boundaries to a section, which shows that the number of boundaries and sections was roughly equivalent. Had the number of boundaries been dramatically higher, the chances of obtaining matches would have been higher as well. Of course what matters is where the boundaries were placed, and this is indicated by the statistics involving the number of matches.

For the calculation of recall and precision rates, the total of final boundaries is used. The recall rate of 31.75% indicates that about $\frac{1}{3}$ of the segment boundaries matched section boundaries. Similarly, the precision rate of 29.53% indicates that about $\frac{1}{3}$ of the segment boundaries were true text

Texts	25
Sections	400
Provisional boundaries	841
Final boundaries	430
Matching boundaries	127
Recall	31.75%
Precision	29.53%

Table 6.5: LSM segmentation performance

boundaries.

The performance figures indicate a much better segmentation performance than that obtained in pilot study 3.

6.13 Random segmentation

The performance of the random segmentation is presented in table 6.6. The performance breakdown by text is shown in appendix 9 (p.468).

As for LSM segmentation, performance rates use the total of final boundaries, not the total of provisional boundaries. Both recall and precision figures indicate that random segmentation was outperformed by LSM segmentation. The recall rate was 23.25%, while for LSM it was higher at 31.75%. This means that random segmentation was able to locate about one section in every four final boundaries placed, while LSM segmentation retrieved about one section in every three final boundaries. Similarly, the precision rate for random segmentation was lower than for LSM: 17.32% versus 29.53%. This indicates that less than one random boundary in five was truly a section boundary, while about one LSM boundary in three was true. In conclusion, LSM segmentation appears to be better than what is expected by chance.

Texts	25
Sections	400
Provisional boundaries	841
Final boundaries	537
Matching boundaries	93
Recall	23.25%
Precision	17.32%

Table 6.6: Random segmentation performance

6.14 Expert segmentation

The performance figures for the expert segmentation are presented in table 6.7 (the breakdown by text is shown in appendix 10 (p.469). TextTiling placed 150 boundaries across the corpus, all of which matched section boundaries, hence the 100% precision rate. These 150 boundaries inserted by TextTiling recalled 37.5% of the total of 400 sections in the corpus.

Expert segmentation has a better performance than LSM segmentation. The striking difference is with respect to precision. While precision for LSM segmentation was 30%, for the expert segmentation it was 100%. The difference was much smaller with respect to recall: 37.5% by TextTiling, and 32% by LSM.

The expert segmentation serves to offer performance level targets against which the performance of LSM can be evaluated. In other words, the results of the expert segmentation can be interpreted as what is reasonable to expect given the texts in the corpus. As for precision, the expert segmentation indicates that a perfect score is attainable (100%), and so LSM achieved about $\frac{1}{3}$ of the maximum score possible. As for recall, the results show just a 5% difference between what LSM achieved and what is reasonable. Given that LSM segmentation approximates performance level targets, it seems fair to conclude that it provides a good level of segmentation performance.

Texts	25
Sections	400
Provisional boundaries	150
Final boundaries	150
Matching boundaries	150
Recall	37.5%
Precision	100%

Table 6.7: Expert segmentation performance

One possible explanation for the superior precision performance by TextTiling could be that in the genre of encyclopedia reports there is a great degree of match between paragraphing and sectioning, and so TextTiling was capable of exploiting this feature by relying on paragraph breaks. However, TextTiling is not aware of generic features, and works by inserting boundaries only between paragraphs. By doing so, it reduces the number of potential boundary places and the margin of error (see introduction to TextTiling on p.110 and previous discussion on p.296). Hence, a more realistic explanation is that the performance of TextTiling is attributable to this fudge. As its previous results indicate less than 100% precision, though, this strategy was not as productive with other data. The fact that using layout information is permitted in TextTiling stresses the instrumental character of the procedure. In other words, what matters for TextTiling is how many sections are located so that a query can return more appropriate texts. By contrast, in the procedures being developed as part of the investigation reported in this thesis, performance matters so long as it is informative of the nature of the relationship between lexical cohesion and discourse organisation.

6.15 Assessment of LSM

It is important to place the performance rates obtained so far in the context of previous research. Figure 6.9 on the following page is an expanded version of figure 5.13 presented above (see p.216) and shows the performance of three main segmentation procedures developed elsewhere. The performance of TextTiling refers to the results of the 'expert' segmentation reported in the present investigation. The reported performance of the procedure presented in Okumura and Honda[5]. is labelled as 'Okumura'. And the performance

[5] Average values quoted in Okumura and Honda (1994).

of the procedure developed in Morris[6] (1988) and Morris and Hirst (1991) is identified as 'Morris' in the chart. Figure 6.9 also includes the performance rates obtained in the pilot studies presented in the previous chapter. The figures in the chart have been ranked in descending order.

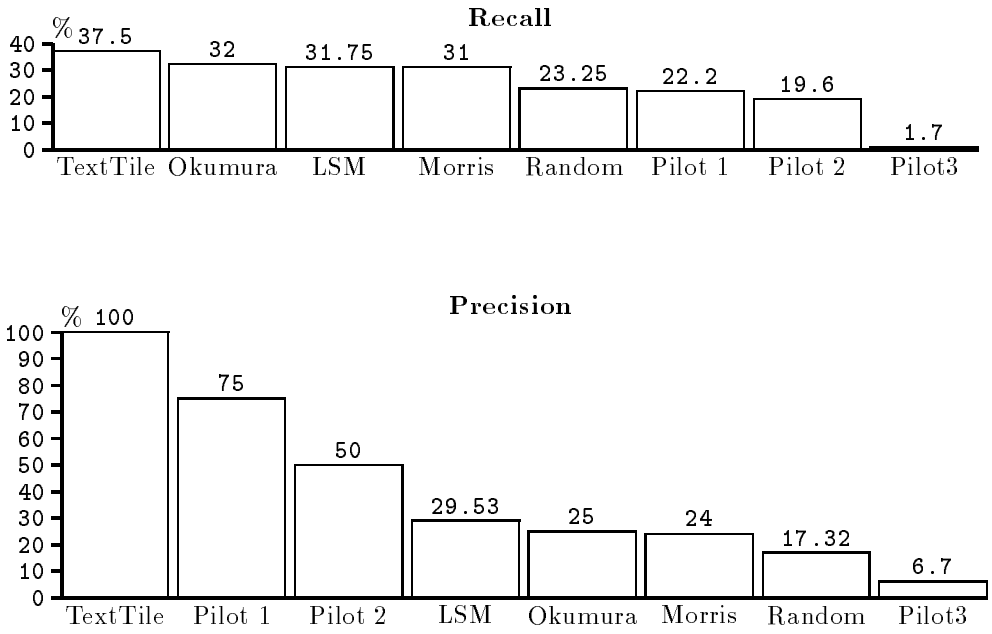


Figure 6.9: Comparison of performance with other procedures

The recall figures indicate that TextTiling achieved the highest rate. Okumura, LSM, and Morris can be considered tied at second place, with figures centering around 31% to 32%. LSM can be considered to have achieved a good ranking, as it performed as well as other segmentation techniques, with the exception of TextTiling; nevertheless, considering the fudge inherent in TextTiling, the recall results obtained by LSM are not disappointing. There is another aspect of the performance of LSM, namely that it scored higher than the pilot studies. This is important in that it shows that progress was made in the research reported in this thesis in developing a better segmentation procedure. In third place is random segmentation, with about 23%. The

[6] For an explanation of how these performance figures were arrived at see fn.6 (p.216).

three pilot studies performed below random, especially the cluster analysis procedure developed in pilot study 3, which scored well below random.

The precision figures also present a good picture as far as the LSM procedure is concerned. It was outperformed by both TextTiling and pilot studies 1 and 2. Again, the fact that TextTiling achieved 100% recall is suggestive of the tweak contained in it which adjusts it to take only paragraph boundaries. LSM is ahead of Okumura and Morris, and importantly, it is ahead of random segmentation. Pilot studies 1 and 2 did well, but the fact that their figures refer to only one text renders their performance less convincing.

The only procedure to have scored below random was again pilot study 3. This suggests that the clustering procedure developed in pilot study 3 might in fact have picked up the opposite phenomenon to segmentation. This negative result therefore deserves careful examination and interpretation, and perhaps warrants a separate study. At the moment, it is possible to speculate on a few possibilities. There are three possible sources of error in the procedure.

The first source of error is the actual clustering method; perhaps the chosen method, k-means, was not suitable for the task of showing separate clusters, even though it was in theory adequate given that segments are non-hierarchical and the k-means method returns non-hierarchical clusters.

The second source of error could lie in the way the data were coded for cluster analysis. Each repetition was coded as a separate case, and each case had two variables, one for each sentence in which the repetition took place. Thus, if the word 'house' appeared in three separate sentences (e.g. 1, 4, and 10), it would be recorded as three cases (case 1: 1, 4; case 2: 1, 10; case 3: 4, 10). The immediate effect of this coding is that the number of cases in each data set could be very large, which in turn increased the chances that the clustering algorithm might make a mistake and throw up

spurious clusters. For instance, if case 1 is coded as '1, 4' (meaning an item is repeated in sentences 1 and 3), case 2 as '1, 10', and case 3 is coded as '2, 3', the clustering algorithm might cluster together cases 1 and 2, because they have sentence 1 in common, and place case 2 in a separate cluster. The result one would expect would be different, though, since it would be more in keeping with the idea of segmentation to keep cases 1 and 3 in the same cluster, since they occur closer to each other in the text, even though they do not have any numbers in common.

The third source of error could be the Cubic Clustering Criterion (CCC) statistic. The local peaks of the CCC statistic were taken as indicative of the best number of clusters in the data, but in the event the number of clusters reported across the data by CCC were always very similar, normally two. This is suspicious, given that the texts had varying numbers of segments, and therefore one would expect this diversity to be reflected in the number of clusters reported by CCC. In short, the CCC statistic may not have worked properly with the data, and this may have been a result of the way the data were coded, or it may have been an indication that the CCC was not a good stopping rule for the data.

The performance figures for individual texts are favourable to LSM. Random segmentation never performed better than LSM; the highest perfomance rates for random segmentation are 40% recall (texts 5 and 8) and 44.44% precision (texts 15 and 16, see appendix 9 on page 468), while for LSM the best figures are 80% recall and 50% precision (texts 9 and 20, respectively, see appendix 8 on page 467). The highest recall rate was lower for expert segmentation than for LSM: 75% (text 17, see appendix 10 on page 469).

In conclusion, the comparative analysis of the performance of several segmentation procedures suggests that the procedures developed in the pilot studies perform generally well, with the exception of pilot study 3. The

results of the random segmentation present evidence that the good performance obtained by LSM segmentation does not seem to have been achieved by chance. This suggests that LSM segmentation is not arbitrary. The results of the expert segmentation suggest that at least with respect to recall LSM segmentation is very close to the maximum practical level. On the whole, the comparative analysis indicates that LSM segmentation is among the best options for segmenting texts available.

6.16 TextTile and LSM

The purpose of comparing segmentation by LSM to segmentation by TextTile was not to judge which procedure is best in competitive terms, but to help put in perspective the levels of performance achieved by LSM. The comparison is also important because it can assist in finding out how the two procedures complement each other. In other words, it may be possible to know whether LSM, which does not perform as highly, is simply retrieving fewer of the same section breaks as TextTile or whether the section boundaries located by each procedure are different. If the former were correct, and one ignored the fact that, unlike TextTiling, LSM is not dependent on paragraph breaks, then it would not be unwarranted to conclude that LSM was simply a poorer version of TextTile given that its recall rate is slightly lower (37.5% against 31.75%); if the latter proved to be correct, though, it would be possible to assume that the two procedures complement each other and that used together they might achieve higher performance.

Table 6.8 on the following page displays the total section boundaries recalled by each procedure alone and jointly by both. It must be explained that the totals for each procedure refer to those section boundaries recalled by one procedure and not by the other. For example, TextTile identified

LSM	80 (34.8%)
TextTile	103 (44.8%)
Both	47 (20.4%)
Total	230 (100%)

Table 6.8: Section boundaries recalled by LSM and TextTile

the boundaries of 150 sections (see table 6.7 on page 309); of those, 103 were identified by TextTile and not by LSM; likewise, according to table 6.5 (p.307), LSM located the boundaries of 127 sections, 80 of which were located by LSM and not by TextTile. Jointly, the two procedures identified only 47 (20.4%) of the different section boundaries recalled in total. Therefore, the majority of the individual section boundaries (about 80%) were identified by either of the two procedures. Appendix 11 on p.470 presents a breakdown of these figures by text.

It appears that the two procedures identify different section boundaries, therefore LSM cannot be considered a 'poorer' version of TextTile. Separately, the two procedures never achieved over 40% recall (37.5% for TextTile and 31.75% for LSM). Their combined recall rate, however, is 57.5% (i.e. 230 matches out of 400 sections). This is a considerable improvement, since it means that more than one out of every two sections is located by the two procedures.

6.17 Summary and terminology

To summarize the stage we have now reached, LSM works as follows. A link set is formed for each sentence of the text containing those sentences with which each particular sentence shares links. The links are represented individually by the sentences in which they occur, so that if there are two links between sentence 1 and 2, and three links between sentence 1 and 3, the link

set for sentence 1 is represented as 22333. A median is calculated for each link set so that the comparison of link sets becomes feasible. The median is simply the midpoint of the ranked distribution of the elements within the link set. In this manner, there will be 50% of the elements of the link set on either side of the median. Once the medians are obtained, the difference between each link set median and its predecessor is calculated. The idea is that large differences, or peaks, will indicate segment boundaries since they will signal those adjacent sentences which have distinct lexical cohesive patterns. In order to know which neighbouring medians differ, an average difference is computed for the whole text. Each individual difference is then compared to the difference average, and those differences which exceed the average (regardless of the sign of the difference) are considered segment boundaries. Since adjacent sentences can have higher than average differences, they can become segment boundaries. In these situations, a peak cluster is said to form. Peak clusters are undesirable because these contiguous boundaries will create one-sentence segments, which are incompatible with the definition of segment as a portion of text consisting of at least two sentences. The terms 'peak' and 'peak cluster' are useful for describing the ups and downs in the plot of median differences. However, they are not meaningful descriptors of what is going on in the segmentation. Since peaks are being observed in order for segment boundaries to be inserted, a better term for a 'peak' or 'peak cluster' would therefore be 'boundary zone', since peak clusters indicate a zone in the text where a boundary can occur.

The terms used during the exposition of the principles behind LSM are glossed below:

Link set: Ordered list of individual links between a particular sentence and the other sentences in the text.

Median: Midpoint point of a link set.

(Median) Difference: Difference between a median and its predecessor.

Average median difference: Mean of greater-than-zero median differences for a single text.

Peak: Those sentences whose medians differences are higher than the average median difference..

Peak cluster: Set of at least two adjacent peaks.

Boundary zone: Sentence or sentences where a peak or peak cluster occurs.

Major peak: The highest of the peaks in a peak cluster.

6.18 Full example

A final example will be presented in this section before moving on to the conclusion of the chapter. First, the numerical elements involved in segmentation by LSM, randomly, and by TextTile are presented. Then the actual segmentation is illustrated by a chart showing the segment boundaries introduced by LSM. Finally, the actual text segments are presented and commented on.

The text used to illustrate segmentation is text 9, which was chosen because it achieved the best recall and precision rates in LSM segmentation (see appendix 8, p.467). Text 9 itself appears in appendix 12. The sentences are numbered so that they can be referred to in link sets. Those sentences that are section boundaries have their numbers marked in bold. The text is about 'Equatorial Guinea', a country in western Africa. The links between all pairs of sentences in the text are listed in appendix 13 (p.474 ff.). These individual links were then used to create the link sets, which appear in ap-

pendix 14 (p.479 ff.), together with the medians for each sentence. These link sets refer to sentences having one link or more with other sentences.

Table 6.9 brings the actual segmentation of the text. The sentences are listed one per row in the 'Sn' column. The location of the five *section boundaries* is shown in the column headed by 'SB' by tick marks ($\surd$). The *medians*, as identified in the previous appendix (p.479 ff.) are reproduced in the table down the 'Md' column next to their respective sentences. The *difference* between each median and its predecessor was computed, and this is presented in the column marked 'Dff'. The *average median difference* for the text was estimated at 8.1, which appears at the top of the table on p.319. Each median difference was then compared to this value, and those differences higher than 8.1 were considered *peaks* and identified as such by a check mark down the 'Pk' column. Peak clusters are not identified formally, but they correspond to those sequences of contiguous check marks found down the 'Pk' column. *Boundary zones* are identified in the column marked 'BZ'. Boundary zones are identified for both LSM and Random segmentation (down the 'LSM' and 'Random' columns respectively). The *segment boundaries* for LSM and Random segmentations appear in the columns headed by 'B'. For TextTiles, the boundaries are tiles, are shown down the 'T' column. Finally, the *matches* for each segmentation are identified in the columns marked 'Mt'; for LSM and Random, matches occur when a section boundary falls within a boundary zone; for TextTile, matches occur when a tile and a section boundary coincide.

The LSM segmentation produced 14 peaks and 8 boundary zones. For each boundary zone one segment boundary was created. Four of the five sections in the text fell within these boundary zones and were counted as matches. This yields a recall rate of 80% ($4 \div 5$) and a precision rate of 50% ($4 \div 8$).

Key
Sn: Sentence; **SB:** Section boundary; **Md:** Median; **Dff:** Difference; **Pk:** Peak; **BZ:** Boundary zone; **B:** Boundary; **Mt:** Match; **T:** Tile.

Average difference: 8.1

					LSM			Random			TextTile	
Sn	SB	Md	Dff	Pk	BZ	B	Mt	BZ	B	Mt	T	Mt
1		13.0						√	√			
2	√	16.5	3.5								√	√
3		9.5	7.0									
4		9.0	0.5									
5		11.0	2.0					√	√			
6		8.0	3.0									
7		16.0	8.0									
8		6.0	10.0	√	√	√						
9		5.0	1.0									
10		4.0	1.0									
11	√	17.0	13.0	√	√	√	√	√	√	√	√	√
12		5.0	12.0	√	√			√				
13		13.0	8.0					√				
14		12.5	0.5									
15		28.5	16.0	√	√	√		√	√			
16		5.0	23.5	√	√							
17	√	17.5	12.5	√	√		√					
18		13.0	4.5					√	√			
19		13.0						√				
20												
21		18.5										
22		36.0	17.5	√	√	√						
23		31.5	4.5									
24		11.0	20.5	√	√	√						
25		30.0	19.0	√	√							
26	√	8.5	21.5	√	√		√					
27		7.0	1.5					√	√			
28		16.0	9.0	√	√	√		√				
29		11.5	4.5									
30		13.0	1.5									
31		20.0	7.0									
32		32.5	12.5	√	√	√						
33		35.0	2.5					√	√			
34		33.5	1.5									
35		32.0	1.5									
36		29.5	2.5									
37		30.5	1.0									
38		31.5	1.0					√	√			
39	√	14.0	17.5	√	√	√	√	√				
40		41.0	27.0	√	√							
41		40.0	1.0									

Table 6.9: LSM segmentation of text 9

The random segmentation resulted in 13 peaks and 8 boundary zones. Only one of the five section boundaries fell within a boundary zone, thus only one match occurred. The recall rate is therefore 20% (1 ÷ 5), while the precision rate is 12.5% (1 ÷ 8).

The TextTile segmentation inserted two tiles, both of which matched sections. As a result, the precision rate is 100% (2 ÷ 2), while the recall rate is 40% (2 ÷ 5).

The LSM segmentation is graphically demonstrated in the chart in figure 6.10. The basic layout of the chart is as follows: the scale running along the bottom of the chart indicates sentence numbers; the vertical scale marked down the left-hand side of the chart gives the median and median differences for each sentence. The dotted plot line shows the values of the medians for each sentence, and the solid plot line represents the median difference. The section boundaries are identified by a vertical dashed line.

The elements pertaining to the actual segmentation were identified as follows. The thick dashed line running across the chart near the bottom represents the average median difference for the text (8.1). Whenever the median difference plot line rose above that line a peak was counted. The actual position of the peaks appear at the very top of the chart marked by

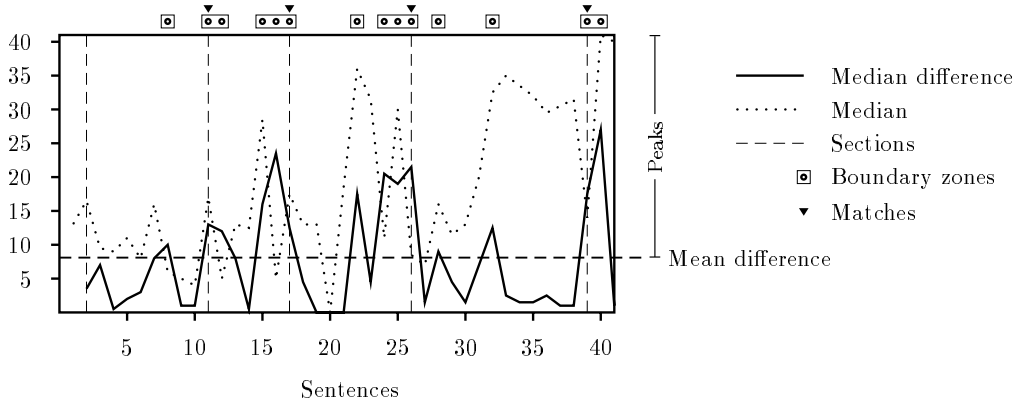


Figure 6.10: Segmentation of text 9 by LSM

a small circle (○). Boundary zones are marked at the top as well by a box (□) surrounding the peak positions. When a section line (vertical dotted line) coincided with a boundary zone, a match was computed, and these are signalled by small triangles (▽) above the boundary zone markings.

In what follows, a commentary will be provided on the segmentation of text 9. The aim of the commentary is not to provide a post-hoc justification of the segments, but rather to draw attention to some features of the segmentation which are not self-evident by referring to the performance rates or chart such as that in figure 6.10 on the preceding page, or a table such as 6.9 on page 319, and also to point out where the segmentation was not correct. The post-hoc analysis cannot be a procedure which the segmentation procedures leans upon, because the segmentation is meant to be carried out without intervention. As such, the post-hoc analysis is justifiable in view of goal of the present investigation which was to develop a procedure which could enable some claims about the relationship between lexical cohesion and text organisation to be made (see p.296).

In the actual text, the segments are the following. The first segment boundary does not start until sentence 8, therefore sentences 1 to sentence 7 are not part of a segment and hence do not enter in the computation of recall and precision. The section 'Land and Resources' beginning at sentence 2 occurs within this segment but does not count as a match because no boundary zones appear nearby. Overall, this part of the text presents details of the geography of Equatorial Guinea:

[0001] Equatorial Guinea, independent republic in western Africa, consisting of a mainland section (Río Muni) on the western coast and the coastal islets of Corisco, Elobey Grande, and Elobey Chico as well as the islands of Bioko (formerly Macías Nguema Biyogo and previously Fernando Po), and Annobón (Pagalu) in the Gulf of Guinea; total area, 28,051 sq km (10,831 sq mi). **[0002] Land and Resources** Mainland Equatorial Guinea is bounded on the north by Cameroon, on the east and south by Gabon, and on the west by the Gulf of Guinea. *[0003]* The terrain is gently rolling and heavily forested; about 60 per cent of the area is drained by the Mbini (formerly Benito) River. *[0004]* With Corisco

and the Elobeys islands it comprises the continental (formerly Río Muni) region, an area of 26,017 sq km (10,045 sq mi). *[0005]* The main island of Equatorial Guinea is Bioko (2017 sq km/ 779 sq mi), which is located off the western coast of Africa in the Bight of Bonny (Biafra). *[0006]* The island, primarily of volcanic origin, is mountainous and thickly wooded, with a steep, rocky coast. *[0007]* Its highest peak is Pico de Santa Isabel (3008 m/9868 ft).

The first segment starts at sentence 8, because a boundary zone was placed at that sentence. No match is computed here because no section boundaries appear in the boundary zone. This segment is a continuation of the section about 'Land and resources':

[0008] The island has fertile volcanic soils and is watered by several streams, and lakes are found in the mountains. *[0009]* Together with the small island of Annobón, lying about 640 km (about 400 mi) to the southwest, it comprises the insular (formerly Bioko) region. *[0010]* The climate is tropical; the average annual temperature is about 25° C (about 77° F) and the annual rainfall is more than 2005 mm (more than 79 in) in most areas.

The second segment begins at sentence 11 since a boundary zone occurs between sentences 11 and 12. The section entitled 'Population' starts exactly at sentence 11, and therefore a match is counted. This segment describes demographic features of the population of Equatorial Guinea:

[0011] Population The population of Equatorial Guinea (1990 estimate) was 348,000. *[0012]* The overall population density was about 12 persons per sq km (about 32 per sq mi). *[0013]* The population is composed almost entirely of black Africans: the Bantu-speaking Bubis, most of whom live on Bioko; the Bengas on Elobey and Corisco; and the Fang (Spanish Pamúes) on the mainland. *[0014]* Persons of European descent and of mixed black and European descent make up the remainder.

The third segment is initiated at sentence 15, and runs up to sentence 21. A boundary zone spans sentences 15, 16 and 17. The section 'Economy and Government', which begins at sentence 17, occurs within the boundary zone, therefore a match is computed. This segment contains the end of the previous section about 'Population', but is best characterized by a presentation of the economy of the country. Two observations are in order here. Firstly, the two initial sentences of the segment could be seen as a colony (Hoey, 1986) in that they are very loosely connected and hence could be read in any order,

that is, it does not make a difference to the comprehension of the text if one reads sentence 16 first and then reads sentence 15. This colony quality would justify one-sentence segments, and the fact that LSM was not allowed to identify one-sentence segments may have rendered it inadequate to deal with colony texts. In retrospect, it might also be argued that encyclopedia articles such as 'Equatorial Guinea' are not perfect data, since they do not have much more than minimal coherence.

Secondly, the section beginning with sentence 17 was hardly wisely labelled, given that the author treated in a single section two themes that would logically deserve separate sections or subsections. Interestingly, LSM identified this division with some accuracy, breaking the section close to where it was most natural, that is, nearly after where the author finished talking about the economy and moved on to focussing on the government of Equatorial Guinea (sentence 23). In short, 'Economy and Government' is a hybrid section and as such it is unlikely to be picked up as a single section by any segmentation system. It may even be said that a hybrid section is the kind of section that a system designer would not want his/her segmentation algorithm to detect since there is very little linguistic justification for hybrid section divisions.

[0015] Spanish is the official language, and Roman Catholicism is the predominant religion. *[0016]* The capital of the continental region is Bata (1983 census, 24, 100), on the mainland, and the largest city, chief port, and capital of the republic is Malabo, formerly Santa Isabel (15,253), on the northern coast of Bioko. **[0017] Economy and Government** Agriculture is the main source of livelihood in Equatorial Guinea. *[0018]* The principal export is cacao, which is grown almost entirely on Bioko. *[0019]* Coffee is grown on the mainland, which also produces tropical hardwood timber. *[0020]* Rice, bananas, yams, and millet are the staple foods. *[0021]* Local manufacturing industries include the processing of oil and soap, cacao, yucca, coffee, and seafood.

The fourth segment comprises sentences 22 and 23. It is demarcated by virtue of the occurrence of a peak at sentence 22. No sections appear within this segment. This segment could be interpreted as a transition seg-

ment, since it contains both the end of the half of the previous section which dealt with the economy, and the beginning of the half which describes the government and political system of Equatorial Guinea:

[0022] The monetary system is based on the franc system (2864 CFA francs equal U S $1; 1990). *[0023]* Under the 1982 constitution, the president was elected by universal suffrage to a seven-year term, and members of the legislature were elected to five-year terms.

The fifth segment spans sentences 24 through 27. The first three sentences (24, 25, and 26) are part of a boundary zone, therefore the section about 'History', whose boundary is at sentence 26, counts as a match. This segment indicates the end of the discussion about the political system and government, and in a sense it is also hybrid since sentences 24 and 25 are as much about history as sentences 26 and 27. In this sense, LSM grouped together sentences which have a degree of coherence, and therefore belong in the same segment:

[0024] The Democratic Party of Equatorial Guinea was the sole legal political party. *[0025]* A new multiparty constitution was approved in 1991. **[0026] History** The island of Fernando Po was sighted in 1471 by Fernão do Po, a Portuguese navigator. *[0027]* Portugal ceded the island to Spain in 1778.

The sixth segment begins at sentence 28 and runs on until sentence 31. This segment was created by the solo peak at sentence 28. No section boundaries occur in this segment, so no matches are computed. It is best characterized by a continuation of the presentation of the history of the country:

[0028] From 1827 to 1844, with the permission of the Spanish government, Great Britain maintained a naval station at Fernando Po and also administered the island. *[0029]* In 1844 the Spanish settled in the area that became the province of Río Muni. *[0030]* In 1904 Fernando Po and Río Muni were organised into the Western African Territories, later known as Spanish Guinea. *[0031]* On October 12, 1968, the territory became the independent republic of Equatorial Guinea, with Francisco Macias Nguema as president.

The seventh segment comprises sentences 32 through 38 and also came about because of a solo peak, this time at sentence 32. Again, no new section boundaries appear in it. This segment presents a closing to the long section about the history of Equatorial Guinea, and in so doing captures the

discussion about the more recent events in the country's history:

[0032] In 1972 Nguema appointed himself president for life. *[0033]* Extreme dictatorial and repressive policies led to the flight of an estimated 100,000 refugees to neighbouring countries; at least 50,000 of those who remained were killed, and another 40,000 were sent into forced labor. *[0034]* In 1979 Nguema was overthrown in a military coup, tried for treason, and executed. *[0035]* Lieutenant Colonel Teodoro Obiang Nguema Mbasogo, who led the coup, then became president. *[0036]* Parliamentary elections, based on a single slate of candidates, were held in 1983 and 1988. *[0037]* Although the first multiparty elections took place in November 1993, they were internationally condemned and boycotted by approximately 80 per cent of the eligible voters. *[0038]* Opposition forces called for the boycott after the Obiang Nguema government refused to prepare an accurate electoral roll and guarantee the right to campaign without harassment.

The last segment runs from sentence 39 up to the end. It is caused by a boundary zone between sentences 39 and 40. A new section begins at sentence 39; since it falls within the boundary zone, it is counted as a match. Although the section is entitled 'Further Reading', it actually signals the end of the body of the text, therefore this segment is perhaps best characterized as an 'appendix':

[0039] Further Reading "Equatorial Guinea," Microsoft (R) Encarta. *[0040]* Copyright (c) 1994 Microsoft Corporation. *[0041]* Copyright (c) 1994 Funk & Wagnall's Corporation.

In conclusion, this section has presented a detailed view of the segmentation of one single text. A chart showing the segmentation of a text on which LSM achieved the highest recall and precision rates was offered to illustrate the various elements involved in segmentation by LSM. The eight final segment boundaries inserted by LSM were presented in the actual text, followed by a brief commentary on the contents of each of the resulting segments.

6.19 Achievement of goals

The major goal of the investigation presented in this chapter was to develop a new segmentation procedure, given the poor performance of the cluster analysis procedure developed in pilot study 3. The development of a new

procedure had to include the automatic computation of cohesion, the automatic placement of boundaries, and the capability to handle several texts. The LSM procedure attained all these three goals. In addition, LSM is consistent with the guidelines established for the research as a whole: extensive coverage, inductive orientation, and objective evaluation. For these reasons, LSM seems to be an adequate segmentation algorithm for applying to a large number of texts, which will in turn help answer the general question about the relationship between lexical cohesion and text internal divisions.

6.20 Improving LSM

One of the aspects of LSM which could be experimented with is the number of links necessary for inclusion in a link set. Currently all sentences which share a link at all are included in link sets. If a threshold were introduced which selected only those sentences which share a criterial number of links then the composition of the link sets would be altered, and consequently the median of those link sets would change as well giving rise to a different segmentation of the texts.

Initially, therefore, two thresholds were tested: 2 and 3 links. Only those sentences sharing at least two links (2-link threshold) and three links (3-

	Threshold	
	2 links	3 links
Texts	25	25
Sections	400	400
Provisional boundaries	477	112
Final boundaries	291	88
Matching boundaries	98	20
Recall	24.5%	5.0%
Precision	33.68%	22.73%

Table 6.10: LSM segmentation performance with two and three links

link threshold) were included in the link sets. No other modifications were introduced, so whatever changes there might be in the segmentation would be the result of the changes brought about in the composition of the link sets.

Table 6.10 on the preceding page sets out the results of the application of thresholds to the segmentation. The 2-link threshold produced the better performance of the two thresholds, notably with regard to recall, which was very low for the 3-link threshold (5%). A measure of the adequacy of the performance is the level of segmentation achieved at random: 23.25% recall and 17.32% precision. Compared to random segmentation, the 3-link threshold option performed below random which is unacceptable. With a threshold of 2 links recall was only slightly above random (24.5%). In contrast to recall, precision rates stayed well above random, even for the 3-link threshold. The scores for individual texts appear in appendices 15 and 16 on pp.483 and 484.

Figure 6.11 presents a comparison of the performance of LSM with three different link thresholds. The 1-link threshold option yielded the best recall rates. Recall was lower the higher the threshold. This is related to the reduction in the total number of sentences available for placing a segment

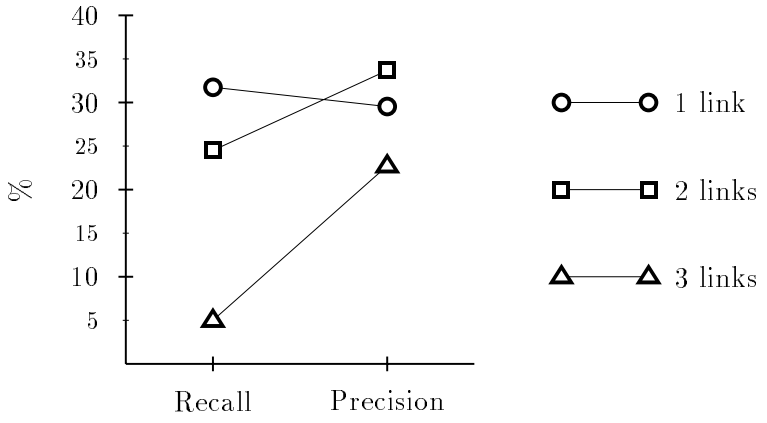


Figure 6.11: Performance of LSM by threshold

boundary at as the threshold grows larger. The higher the threshold, the fewer sentences there will be which qualify for inclusion in link sets. As a result, there are fewer sentences with link sets, fewer medians, fewer median differences and ultimately fewer potential segment boundaries. Obviously, the number of section boundaries stays the same, and so the chances of a segment boundary matching a section boundary are lower.

In contrast to recall levels, precision rates are higher for 2 links than for 1 link. This suggests that the reduced number of potential segment boundaries does not in itself affect the ability of the procedure to place matching boundaries, despite the fact that the drop relative to the 3-link threshold may suggest that higher thresholds may affect the precision of LSM more dramatically. Nevertheless, the fact that the difference among the link thresholds for precision is not as large as that for recall indicates that the LSM precision is not greatly affected by the sheer number of sentences available for it to place boundaries at.

This concludes the development of the Link Set Median procedure. LSM proved satisfactory in that it included the automatic computation of cohesion, and enabled the automatic placement of boundaries, while being capable of handling several texts. LSM is also consistent with the guidelines for the whole investigation, namely extensive coverage, inductive orientation, and objective evaluation. LSM will therefore be used in the main large-scale investigation which is reported in the following chapter..

Chapter 7

Main study: Large-scale application of the Link Set Median procedure

In this chapter the main study on segmentation is presented. The focus of the chapter is on reporting the application of the Link Set Median procedure to a corpus of 300 texts. The chapter begins with a presentation of its specific aims, followed by a description of the data, and a report on four different analyses of the data. The chapter ends with a summary of the results.

7.1 Aims

The aims of the study were to:

1. Find out whether LSM segmentation performs better than random segmentation on a wider range of texts than used so far;
2. Find out whether performance is affected by link levels;
3. Find out whether performance is affected by text type;

LSM segmentation was carried out by applying the same procedures as described in the previous chapter (see section 6.5, p. 279 ff.). Random segmentation of the data used in this study was also implemented as described in the previous chapter (see p. 295).

7.2 Data

The data used for this study consist of a corpus totalling 300 texts. The data are made up of three independent corpora, one for each of the following genres: research articles, business reports, and encyclopedia articles. These three genres were selected because they represent three types of texts widely used by three different discourse communities, respectively academics, business executives and shareholders, and students/readers in general. Each corpus contains 100 texts. A sample size of 100 texts is above the 60 units suggested by Sibson (1972) on statistical grounds as a convenient sample. It is also much higher than the 10-text sample suggested by Biber (1995a, p.133). The 100-text samples can therefore be claimed to be representative of each genre. Table 7.1 presents the dimensions of the three corpora. Altogether the whole corpus totals over one and a quarter million running words, beyond the traditional 1-million-word benchmark for corpus analysis. The research article corpus is the longest, with more than half a million words,

Corpus	Research Articles	Business Reports	Encyclopedia Articles	Grand Total
Texts	100	100	100	300
Sections	940	1,741	956	3,637
Sentences	20,090	14,631	9,743	44,464
Total running words	577,026	429,728	255,956	1,262,710
Total different words	23,903	13,263	19,073	—

Table 7.1: Data for the main study

and the encyclopedia corpus is the shortest, with about a quarter of a million words. However, the corpus was planned based on a criterial number of texts rather than words, since the unit around which the analysis centres is the text. Care was taken during text selection so that the texts were randomly selected from a larger subset of the population of each text type.

The research articles were drawn from two sources, the main one of which was the electronic library of research articles available from the University of Liverpool Sydney Jones library homepage[1]; the other source was the collection of printed articles kept on the shelves in Sydney Jones library itself. The electronic depository was preferred because the articles could be more easily rendered in the format required for inputting into `words`, the computer software designed for the analysis of lexical cohesion (see chapter 5, section 5.4.6, p.245 ff.). The electronic articles were originally saved as 'Adobe Acrobat' (`.pdf`), a format incompatible with `words`, which requires plain ASCII or ANSI. A simple macro was created which copied each `.pdf` file into Microsoft Word and saved them in MS Word's native format (`.doc`). Because this conversion method is primitive, many formatting features were lost, which meant the texts had to be manually checked for spelling and punctuation. Also, several formatting characters were left in the files which could ruin the analysis. To correct these problems, filters were applied to the texts by using Text Converter, a search-and-replace utility available in WordSmith Tools (Scott, 1996).

Once the corrective measures had been applied, the research articles were apparently ready for running through `words`. Preliminary runs of the files, though, identified problems with formatting characters which escaped the filters created for Text Converter. Surprisingly, these late corrections proved more difficult since it was only a handful of characters which were causing

[1] The web address is `http://www.liv.ac.uk`.

damage that could easily have escaped attention. The most serious problem had to do with the loss both of certain end-of-sentence markers and of several words at the onset of the subsequent sentences. This problem was quite serious since the sums for sentence totals always added up and it was only when an individual check on each file was carried out that the problem was perceived. After these nuisance characters were removed the texts were ready for running through `words`.

The printed research articles were scanned into electronic format and checked manually against the original. This transfer method proved more laborious throughout, since the OCR software ('TextBridge') misinterpreted several characters, tables and figures. For this reason, the electronic source was preferred, thus making up the majority of the texts in the corpus.

The business reports were chosen from the electronic depository for the United States Securities and Exchange Commission in Washington, DC on WWW[2]. The site has thousands of reports of various kinds. The type of report chosen for this study is the 10-K form, an annual report which most American firms are required to issue by law, within 90 days after the end of the company's fiscal year. The 10-K form is important because it 'provides a comprehensive overview of the registrant's business' (Guide to Corporate filings, 1997). The reports are made up of several more-or-less independent parts. The initial part is obligatory and it is there where the companies describe their dealings in running text. The middle parts are mostly tables and figures followed by legal text. At the end the reports include other texts or text extracts such as the Annual Report for Shareholders as an appendix. Figure 7.2 on the following page gives a typical sample table of contents of a 10-K form. The headings listed are only the main sections; there were several

[2]The address is `http://www.sec.gov`

Table 7.2: Typical 10-K contents page

other divisions which were taken into account and counted as sections. Since the 10-K form as a whole contained other text portions from other sources, as well as large sections of numeric data, using the whole report did not seem a good choice. The choice was made to select the first part of each report, because that was largely self-contained and independent of the other parts. Unlike articles, the business reports did not pose problems during conversion for use with `words` as the files were already in ASCII format on the Internet. The only difficulty was eliminating the unwanted parts of each form, which was done manually.

It might seem that choosing part of a 10-K report would be a violation of the principle of investigating full texts only. However, what was chosen was not an incomplete text but an incomplete *document*. The boundaries of a text are determined to an appreciable extent by theoretical considerations, whereas the limits of a document are determined by practical considerations which have to do with tradition, the medium on which the material is published, and the publishers themselves. Hence, a magazine is a document, but the articles, advertisements, and letters in it can be treated as individual texts; likewise, an issue of a scholarly journal is a document typically comprising several texts such as research articles, reviews, and conference announcements. The boundaries of a document published in print are typically its front and back cover. By contrast, the boundaries of documents published electronically are the first and last lines of code in the computer file carrying the document; the 10-K report is document of this kind. It is argued here that the 10-K report is a document containing a number of texts, one of which is Part I, and so it is legitimate to treat Part I as a separate text.

The encyclopedia articles were selected from the 1995 edition of Encarta, published by Microsoft on CD-ROM. The texts were chosen from the main

menu and copied into a word processor, from where they were saved as ASCII. The encyclopedia articles were the least problematic, as they could be input to `words` straight away.

All texts were marked up for textual features before being processed in `words`. Section headings were identified by hand, surrounded by open-close tags, and eliminated from the texts prior to analysis. The sentence following a section heading was then considered a section boundary and identified by a section boundary marker. Not all texts had section headings, though; some were simply identified by a sequential number (e.g. I, II, III ...). In cases such as these, the very first sentence of each section was considered a section boundary and tagged as such.

By far, the textual feature which deserved the most attention was sentence boundaries. The sentence final period appears in computer files as a dot, which can be mistaken for the dot used in figures, acronyms, and abbreviations of all kinds (e.g. 9.99, U.K., km., etc). Searching for a string beginning with a capital letter and ending with a dot improves results somewhat but it is not by any measure a reliable approach. Correctly disambiguating end of sentences is not a trivial task in automatic text processing (Atwell, 1986; Grefenstette and Tapainen, 1994). For the present study, the correct identification of sentence ends is a top priority; wrong sentence divisions would invalidate the conclusions about lexical cohesion between sentences. The best approach to the data given the computational resources and skills available to the project was to search for numbers with decimals ('9.99', for instance) and delete the decimal point ('9 99'), and search for common acronyms such as 'e.g.', 'Mr.', and 'U.S.A' and replace them with forms without dots ('e g', 'Mr ', 'USA'). A manual check was then carried out to verify that the remaining full stops were indicating sentence ends. These were marked up with a unique 'end-of-sentence' tag.

As explained above (see section 5.4.6 on page 245), the way `words` handles a text prior to the identification of repeated items is determined by a number of control files which handle such tasks as removal of stop words, lemmatisation, and identification of synonyms and multi-word items. Given the large size of the three corpora, it was felt that it was unrealistic to try to finely tune the control files so that all different word forms were lemmatised, all different synonyms were correctly matched, and all multi-word groups were adequately tokenized, and therefore it was decided that no further effort would be invested in updating the control files. Thus, the control files used in the analysis contained but a subset of the instructions which would be necessary to handle the texts in full.

The 100 texts for each corpus were selected at random in different ways. The research articles were chosen from an initial pool of about 140 files, 120 of which were from the electronic depository. These were all picked at random. The twenty articles in print were chosen by first picking assorted issues of journals off the central shelves on the first floor in Sydney Jones Library, and then scanning twenty random articles from the journals. With respect to the collection of online articles, the initial intention was to obtain a random sample of 120 articles from 60 different journals by collecting two articles from the last issue of each journal. The selection of the online articles proceeded as follows. First, the list of journal titles from the electronic library of research articles was accessed on the Internet, then one journal title was chosen arbitrarily from this list. The list of issues available from each journal was then brought onto the computer screen, from which the latest issue of each journal was chosen. From this list of articles in the latest issue two articles were chosen arbitrarily and downloaded.

A problem which occurred was that sometimes there was only one article available online in the latest issue; in these cases only one article was selected.

Another frequent problem was that on many occasions the downloading was not completed successfully because of problems on the remote server; the solution in such cases was to select another article and start downloading again. The effect of these problems on the collection of the data was that it was not possible to collect two different articles from the last issue of 60 different articles from the electronic library of research articles as initially planned. To reach the target number of 120 articles, several journals had to contribute more than two articles.

The 120-article sample of online articles was then joined with the 20-article sample scanned from printed articles and stored in a directory on the computer. At this point, a random selection of the final 100-article sample took place, which consisted of simply choosing the top 100 articles from the directory.

The collection of business reports was gathered by trying to download about ten texts for each letter of the alphabet from the alphabetical index on the web page. The initial intention was to obtain about 250 texts in this way. When more than one report was available for any one company, the version for the most recent year was chosen. The initial plan was marred because on several days the Internet connection was either too slow or the site was down; after several days without successfully downloading any whole texts, the collection was stopped before all the letters of the alphabet had been worked through. At this point about 150 full business reports had been collected. The first 100 reports in the hard disk directory were chosen.

The encyclopedia articles were chosen by randomly selecting about 15 articles from each of the options on the initial menu in Encarta 95, namely Physical Science and Technology; Life Sciences; Geography; History; Social Sciences; Religion and Philosophy; Art, Language and Literature; Performing Arts; and Sports, Hobbies and Pets. The only selection criterion applied at

this stage was that articles had to have section divisions. This resulted in over 130 articles. From those, 100 were chosen at random to make up the final corpus.

Appendices 17, 18, and 19 present the titles of each text and the reference numbers assigned to them.

7.3 Methods

The texts were segmented using the LSM procedure described in the previous chapter (see section 6.5, pp.279 ff.).

7.4 Analysis of variance

In order to answer the first question posed by the main study, namely whether LSM segmentation performs better than random segmentation at matching segments with section boundaries in all three sub-corpora, mean recall and precision rates were compared statistically. The comparison was carried out by means of a one-way repeated-measures MANOVA (Girden, 1992; SAS Institute Inc, 1989a) with type of segmentation (LSM or random segmentation) as the between-subjects independent variable and the link levels as six dependent variables. The validity of repeated measures analysis of variance rests on homogeneity of covariance, or sphericity (Girden, 1992, pp. 15-18). A multivariate analysis of variance (MANOVA) was chosen because, unlike repeated-measures ANOVA, it does not require homogeneity of covariance, an assumption which is commonly violated in repeated measures designs (Tabachnick and Fidell, 1989, pp.470-471; Girden, 1992, p.18). In the case of the present research, by homogeneity of covariance is meant, for example, equivalent correlations between recall rates across different levels of linkage

(e.g. ≥1 links vs ≥2 links, ≥1 links vs ≥3 links, etc).

Results are presented in tables 7.3 through 7.8 (see pp.344–349). The tables are organised as follows. The top of each table states the individual mean performance rates (with recall and precision in separate tables) for each corpus. The mean rates indicate *mean percentages* taken by dividing the percentage rate for individual texts by the total number of texts and multiplying by 100. Therefore, the mean rates do not reflect simply the rate of matches for the corpus as a whole but the average percentage per text. To illustrate, across the research article corpus there were 943 sections, and LSM matched 342 of those; if recall were calculated for the whole of the corpus, the rate would be 36.27% (342÷943×100). By contrast, by first calculating individual recall rates for each text and then averaging out the percentage recall, the mean recall rate for the corpus is 36.93%, as shown in table 7.3 on page 344 against links ≥1. The difference in this case is small, but it may not necessarily be so, which may affect the statistical comparison of the means. The mean percentage approach was preferred since what is important for the current study is how the segmentation procedure performs on a text by text basis, and not across the corpus as a whole with no regard for boundaries between texts. Similarly, the overall rate for each type of segmentation was obtained by averaging out the individual 600 rates (i.e. 100 texts × 6 levels of linkage) for each corpus.

In analysing the variation among precision and recall rates, three types of comparison needed to be made. Firstly, it was necessary to compare the mean performance rates across all link levels for LSM with the mean performance rates across all link levels for random segmentation. The values compared here are those which appear in the top table in the rows marked 'All'. The results of the comparison appear in the table entitled 'Between segmentations'. A p-value in that table which was less than 0.05 indicates a

significant difference between the mean value for LSM segmentation and the mean value for random segmentation.

Secondly, it was necessary to compare the mean performance rates by link level for each kind of segmentation. More specifically, it was necessary to contrast the LSM mean for links ≥ 1 against the LSM for links ≥ 2, the LSM mean for links ≥ 2 against the LSM for links ≥ 3, and so on until the contrast between the LSM mean for links ≥ 5 and the LSM for links ≥ 6. As part of the same set of comparisons, it was also necessary to carry out the same type of comparison for random segmentation, beginning with the random segmentation mean for links ≥ 1 versus the random segmentation mean for links ≥ 2, and so on up to the random segmentation mean for links ≥ 5 versus the random segmentation mean for links ≥ 6. The data for this set of comparison are the individual mean values for each link level within the table with LSM values, and within the table with random segmentation values. The results of this comparison are shown in the table labelled 'Within segmentations' in the row which says 'Links'. A p-value less than 0.05 indicates that there was a significant difference among link levels for each type of segmentation (LSM and random segmentation).

Finally, the last kind of comparison that needed to be carried out was with respect to the mean performance rates by linkage level between each kind of segmentation. In other words, it was necessary to compare the *LSM* mean for links ≥ 1 with the *random* segmentation mean for links ≥ 1, then the *LSM* mean for links ≥ 2 with the *random* segmentation mean for links ≥ 2, and so on, until the comparison between the *LSM* mean for links ≥ 6 and the *random* segmentation mean for links ≥ 6. The data for this set of comparison are also the individual mean values for each link level, but unlike in the previous comparison, one is not restricted to comparing within each segmentation table, rather the comparison is between each table. The

results of the comparison appear in the 'Within segmentation' table in the row marked 'Link × Segmentation'. In the statistical literature, this type of comparison is concerned with finding whether there was *interaction* between the two kinds of segmentation (LSM and random) and the various link levels. Interaction means that two variables are not moving in parallel, that is, a rise in one of them may imply a fall in another. Thus, a p-value that is less than 0.05 indicated that there was interaction between LSM and random segmentation across link levels. The presence of interaction would indicate the fact that differences between LSM and random segmentation were not constant, that is, a significant difference between LSM and random segmentation would cease to exist at a certain link level. Where there was significant interaction (i.e. $p \leq 0.05$ for Link × Segmentation), a table entitled 'Pairwise comparisons' was added in order to find out at which link level a difference between types of segmentation ceased to be statistically significant. A value of p that is less than 0.05 in the 'Pairwise comparisons' table indicates a link level for which there was a statistically significant difference between LSM and random; the acronym 'NS' ('not significant'), in turn, indicates a difference that was not statistically significant.

The mean recall rates for the research article corpus are shown in table 7.3 on page 344. The overall rate for LSM is 22.96%, whereas for random it is lower at 13.64%; the difference is statistically significant (between segmentations $F=38.77$, $p<0.0001$)). The within-segmentations MANOVA shows significant effects for both links ($F(5,194) = 84.82$, $p<0.0001$) and for links by segmentations ($F(5,194) = 7.09$, $p<0.0001$). The pairwise comparisons indicate that both LSM and random rates decrease as the number of links increase, but only up to links ≥ 5; at links ≥ 6 there are no significant differences any more between segmentations. The precision rates for the research article corpus are shown in table 7.4 on page 345. The overall rate is 10.09%

for LSM and 4.92% for random, which is significant (between segmentations $F = 28.19$, $p<0.0001$). The within-segmentations effects are not significant, indicating that there was no difference in precision across link levels for either LSM or random; in other words, the mean percentage rates for individual linkage levels are not statistically different for LSM or for random segmentation.

The mean recall rates for the business report corpus are depicted in table 7.5 on page 346. The overall rate is significantly higher for LSM (21.38%) than for random (15.92%), as indicated by the between segmentations value of F (25.61, $p<0.0001$). The within-segmentations analysis shows significant effects for links ($F(5,194) = 119.2580$, $p<0.0001$) as well as for the interaction between links and segmentation ($F(5,194) = 10.2595$, $p<0.0001$). This suggests that the recall rates vary across link levels, more specifically the rates for lower link levels seem to be higher. However, the presence of interaction indicates that when LSM and random segmentation are compared, these differences cease to exist; more specifically, according to the pairwise comparisons, it is at links $\geq$4 that the differences between LSM and random segmentation are no longer statistically significant. The LSM mean is significantly higher than the random mean for links $\geq$1 (29.30 against 21.80), links $\geq$2 (36.08 against 21.75), and links $\geq$3 (26.93 against 19.72); however, for links $\geq$4, the difference between LSM is no longer statistically significant (18.66 against 15.49), neither is it for links $\geq$5 (10.37 against 10.35) nor for links $\geq$6 (6.96 against 6.44).

The mean precision rates are shown in table 7.6 on page 347. As with recall, the LSM rates are significantly higher than random (22.37% vs 13.40%, $F = 35.99$, $p<0.0001$). The within-segmentations analysis also yield significant results ($F(5,194) = 3.05$, $p<0.0113$ for the links effect, and $F(5,194) = 3.91$, $p<0.0021$, for the links by segmentation). This suggests that precision

rates vary as the number of links also varies, and that LSM and random interact. The pairwise comparisons reveal the nature of the interaction, namely that up to links $\geq$5 LSM rates seem to be higher than random, but at links $\geq$6 their rates are no longer statistically different.

Finally, the results for the encyclopedia article corpus are shown in tables 7.7 and 7.8 (pp.348–349). Mean recall rates (see table 7.7) are significantly higher for LSM (17.26%) than for random (11.11%), as indicated by the between segmentations value of F (27.35, $p < 0.0001$). The within-segmentations effect of links is also significant ($F(5,194) = 117.07$, $p<0.0001$), and so is the interaction between links and segmentation ($F(5,194) = 8.057$, $p<0.0001$). This indicates that the rates vary across link levels, more specifically with LSM rates tending to be higher than random only up to links $\geq$3; after that, as revealed by the pairwise comparisons, the means cease to be statistically different. Precision rates present a similar picture (see table 7.8). LSM rates (12.95%) are significantly higher than random (8.01%) as indicated by the outcome of the between segmentations comparison ($F = 15.00$, $p<0.0001$). The within-segmentations effects of both links and link by segmentation are also significant ($F(5,194) = 19.4027$, $p<0.0001$ for links, and $F(5,194) = 6.4560$, $p<0.0001$ for link by segmentation). Pairwise comparisons indicate that the interaction between links and segmentation lies in the fact that LSM rates are significantly higher than random only up to links $\geq$3.

In summary, the most important result is that the main effect of segmentation was statistically significant in all three corpora for both recall and precision. More precisely, segmentation by LSM yielded higher performance rates than random segmentation. Within-segmentation analyses also

LSM					
Links	N	Mean %	sd	Min %	Max %
≥1	100	36.93	21.51	0	100.00
≥2	100	38.32	18.77	0	100.00
≥3	100	27.63	17.82	0	100.00
≥4	100	18.59	16.28	0	75.00
≥5	100	11.89	13.87	0	50.00
≥6	100	4.42	7.96	0	40.00
All	600	22.96	20.73	0	100.00
Random					
Links	N	Mean %	sd	Min %	Max %
≥1	100	22.92	18.66	0	100.00
≥2	100	22.27	19.31	0	100.00
≥3	100	16.86	17.71	0	100.00
≥4	100	10.44	14.17	0	100.00
≥5	100	5.83	10.03	0	50.00
≥6	100	3.51	7.97	0	50.00
All	600	13.64	17.00	0	100.00

Repeated Measures Analysis of Variance

Between segmentations: *Tests of Hypotheses for Between Subjects Effects*					
Source	df	SS	Mean square	F	p
LSM × Random	1	26073.26	26073.26	38.77	0.0001
Error	198	133168.74	672.56		

Within segmentations: *Manova Test Criteria and Exact F Statistics*				
	Hotelling-Lawley Trace	F	df (Num/Den)	p
Links	2.186	84.8201	5 194	0.0001
Link × Segmentation	0.183	7.0942	5 194	0.0001

Pairwise comparisons *Bonferroni (Dunn) T tests*					
Links	LSM %	Random %	df	MSE	p
≥1	36.93	22.92	198	405.5415	<0.05
≥2	38.32	22.27	198	362.7177	<0.05
≥3	27.63	16.86	198	315.808	<0.05
≥4	18.59	10.44	198	233.1364	<0.05
≥5	11.89	5.83	198	146.5594	<0.05
≥6	4.42	3.51	198	63.56281	NS

Table 7.3: Recall rates for research article corpus

LSM					
Links	N	Mean %	sd	Min %	Max %
≥1	100	11.29	9.55	0	57.14
≥2	100	10.85	7.90	0	45.45
≥3	100	10.01	8.21	0	36.84
≥4	100	9.89	10.03	0	50.00
≥5	100	11.05	14.91	0	100.00
≥6	100	7.45	15.85	0	100.00
All	600	10.09	11.54	0	100.00
Random					
Links	N	Mean %	sd	Min %	Max %
≥1	100	5.03	4.88	0	30.00
≥2	100	5.10	5.18	0	30.00
≥3	100	4.44	4.64	0	23.08
≥4	100	4.09	5.26	0	23.53
≥5	100	4.35	8.21	0	57.14
≥6	100	6.52	16.63	0	100.00
All	600	4.92	8.60	0	100.00

Repeated Measures Analysis of Variance

Between segmentations: *Tests of Hypotheses for Between Subjects Effects*					
Source	df	SS	Mean square	F	p
LSM × Random	1	8010.40	8010.40	28.19	0.0001
Error	198	56267.76	284.18		

Within segmentations: *Manova Test Criteria and Exact F Statistics*				
	Hotelling-Lawley Trace	F	df (Num/Den)	p
Links	0.037	1.4455	5 194	0.2097
Link × Segmentation	0.042	1.6378	5 194	0.1517

Table 7.4: Precision rates for research article corpus

LSM					
Links	N	Mean %	sd	Min %	Max %
≥1	100	29.30	14.91	0	83.33
≥2	100	36.08	14.63	0	100.00
≥3	100	26.93	15.11	0	60.00
≥4	100	18.66	12.82	0	50.00
≥5	100	10.37	8.89	0	33.33
≥6	100	6.96	7.38	0	31.25
All	600	21.38	16.34	0	100.00
Random					
Links	N	Mean %	sd	Min %	Max %
≥1	100	21.80	11.47	0	58.82
≥2	100	21.75	11.69	0	58.82
≥3	100	19.72	11.26	0	58.82
≥4	100	15.49	9.89	0	47.06
≥5	100	10.35	8.54	0	35.29
≥6	100	6.44	6.89	0	25.00
All	600	15.92	11.64	0	58.82

Repeated Measures Analysis of Variance

Between segmentations: *Tests of Hypotheses for Between Subjects Effects*					
Source	df	SS	Mean square	F	p
LSM × Random	1	8935.89	8935.89	25.61	0.0001
Error	198	69088.77	348.93		

Within segmentations: *Manova Test Criteria and Exact F Statistics*				
	Hotelling-Lawley Trace	F	df (Num/Den)	p
Links	3.073	119.2580	5 194	0.0001
Link × Segmentation	0.264	10.2595	5 194	0.0001

Pairwise comparisons *Bonferroni (Dunn) T tests*					
Links	LSM %	Random %	df	MSE	p
≥1	29.30	21.80	198	177.1292	<0.05
≥2	36.08	21.75	198	175.579	<0.05
≥3	26.94	19.72	198	177.677	<0.05
≥4	18.66	15.49	198	131.2057	NS
≥5	10.37	10.35	198	76.1522	NS
≥6	6.96	6.44	198	51.0790	NS

Table 7.5: Recall rates for business report corpus

LSM					
Links	N	Mean %	sd	Min %	Max %
≥1	100	23.61	14.92	0	75.00
≥2	100	27.18	14.69	0	80.00
≥3	100	23.32	14.72	0	75.00
≥4	100	21.59	17.01	0	100.00
≥5	100	18.94	18.53	0	100.00
≥6	100	19.59	24.43	0	100.00
All	600	22.37	17.86	0	100.00
Random					
Links	N	Mean %	sd	Min %	Max %
≥1	100	13.01	7.83	0	36.36
≥2	100	12.99	8.06	0	36.36
≥3	100	13.10	9.36	0	55.56
≥4	100	13.31	9.56	0	37.50
≥5	100	13.71	12.92	0	50.00
≥6	100	14.26	18.86	0	100.00
All	600	13.40	11.71	0	100.00

Repeated Measures Analysis of Variance

Between segmentations: *Tests of Hypotheses for Between Subjects Effects*					
Source	df	SS	Mean square	F	p
LSM × Random	1	24173.44	24173.44	35.99	0.0001
Error	198	132989.79	671.6656		

Within segmentations: *Manova Test Criteria and Exact F Statistics*				
	Hotelling-Lawley Trace	F	df (Num/Den)	p
Links	0.078	3.05	5 194	0.0113
Link × Segmentation	0.110	3.91	5 194	0.0021

Pairwise comparisons *Bonferroni (Dunn) T tests*					
Links	LSM %	Random %	df	MSE	p
≥1	23.61	13.01	198	142.1602	<0.05
≥2	27.18	12.99	198	140.4874	<0.05
≥3	23.32	13.10	198	152.2481	<0.05
≥4	21.59	13.31	198	190.517	<0.05
≥5	18.95	13.71	198	255.3252	<0.05
≥6	19.59	14.26	198	476.6925	NS

Table 7.6: Precision rates for business report corpus

LSM					
Links	N	Mean %	sd	Min %	Max %
≥1	100	35.35	23.67	0	100.00
≥2	100	38.70	22.25	0	100.00
≥3	100	22.66	20.48	0	100.00
≥4	100	4.90	9.95	0	50.00
≥5	100	1.63	5.18	0	25.00
≥6	100	0.31	2.58	0	25.00
All	600	17.26	22.72	0	100.00
Random					
Links	N	Mean %	sd	Min %	Max %
≥1	100	23.25	19.74	0	75.00
≥2	100	21.40	18.20	0	75.00
≥3	100	15.14	16.71	0	66.67
≥4	100	5.35	10.16	0	50.00
≥5	100	1.25	4.45	0	33.33
≥6	100	0.28	1.85	0	16.67
All	600	11.11	16.51	0	75.00

Repeated Measures Analysis of Variance

Between segmentations: *Tests of Hypotheses for Between Subjects Effects*					
Source	df	SS	Mean square	F	p
LSM × Random	1	11337.52	11337.52	27.35	0.0001
Error	198	82092.87	414.6104		

Within segmentations: *Manova Test Criteria and Exact F Statistics*				
	Hotelling-Lawley Trace	F	df (Num/Den)	p
Links	3.017	117.07	5 194	0.0001
Link × Segmentation	0.207	8.057	5 194	0.0001

Pairwise comparisons *Bonferroni (Dunn) T tests*					
Links	LSM %	Random %	df	MSE	p
≥1	35.35	23.25	198	475.1116	<0.05
≥2	38.70	21.40	198	413.366	<0.05
≥3	22.66	15.14	198	349.4668	<0.05
≥4	5.35	4.90	198	101.2802	NS
≥5	1.63	1.25	198	23.3667	NS
≥6	0.31	0.28	198	5.0463	NS

Table 7.7: Recall rates for encyclopedia article corpus

LSM					
Links	N	Mean %	sd	Min %	Max %
≥1	100	19.01	16.95	0	100.00
≥2	100	21.66	14.44	0	66.67
≥3	100	19.76	20.03	0	100.00
≥4	100	8.77	17.69	0	100.00
≥5	100	6.50	20.30	0	100.00
≥6	100	2.00	14.07	0	100.00
All	600	12.95	18.90	0	100.00
Random					
Links	N	Mean %	sd	Min %	Max %
≥1	100	9.70	9.50	0	50.00
≥2	100	9.28	8.05	0	33.33
≥3	100	9.93	10.94	0	50.00
≥4	100	9.45	17.06	0	100.00
≥5	100	6.70	20.82	0	100.00
≥6	100	3.00	17.14	0	100.00
All	600	8.01	14.83	0	100.00

Repeated Measures Analysis of Variance

Between segmentations: *Tests of Hypotheses for Between Subjects Effects*					
Source	df	SS	Mean square	F	p
LSM × Random	1	7322.31	7322.31	15.00	0.0001
Error	198	96676.57	488.26		

Within segmentations: *Manova Test Criteria and Exact F Statistics*				
	Hotelling-Lawley Trace	F	df (Num/Den)	p
Links	0.500	19.4027	5 194	0.0001
Link × Segmentation	0.166	6.4560	5 194	0.0001

Pairwise comparisons *Bonferroni (Dunn) T tests*					
Links	LSM %	Random %	df	MSE	p
≥1	19.01	9.70	198	188.9244	<0.05
≥2	21.66	9.28	198	136.8331	<0.05
≥3	19.76	9.93	198	260.5506	<0.05
≥4	9.45	8.77	198	302.1319	NS
≥5	6.70	6.50	198	423.0009	NS
≥6	3.00	2.00	198	245.9596	NS

Table 7.8: Precision rates for encyclopedia article corpus

revealed that LSM did not outperform random at all six link levels, though. Looking at all three corpora at once, there was a tendency for LSM and random to yield comparable segmentation rates at higher links levels, particularly six links or more. However, the overall picture in this respect is not as clear-cut as for between segmentations alone. The only situation in which link levels did not influence segmentation was in respect to precision rates in the research article corpus; here both segmentations produced equivalent levels of precision regardless of the number of links. In all other situations link levels had an effect on performance, notably at six links or more, since in five of the six situations (the exception being research articles precision) LSM and random rates did not differ statistically.

As mentioned in the previous chapter, one explanation for the effect of links is that as the number of links required for obtaining the median increases, fewer sentences exist which share as many links with any other sentences, therefore failing to produce link sets and ultimately a median. Segment boundaries cannot be placed at these sentences, hence the number of possible matches will be lower. In short, higher link levels seem to affect results because they eliminate section boundary sentences from the analysis.

The analysis provided so far did not answer the question of at which link level LSM performs best. To answer this question a one-way analysis of variance was conducted with link as the independent variable and either recall or precision as the dependent variable, each with six levels of linkage ($\geq$1 through $\geq$6). Since the previous multivariate analyses provided omnibus F tests which indicated significant effect of links and link by segmentation, another F test is not needed from the one-way analysis, the goal of which is simply to compare the means for LSM segmentation between different link levels. There are a number of tests available for testing mean differences in analysis of variance. Tukey-Kramer's method is reportedly more powerful

than other tests such as Bonferroni, Sindak or Scheffe (SAS Institute Inc, 1989b, p.944); however it has a high type II error rate, that is, it tends to retain the null hypothesis when it should be rejected.

A better alternative seems to be the multiple F test developed by Ryan, Einot, Gabriel and Welsch ('REGWF'); together with its variant, 'REGWQ', it is reported to be the most powerful mean comparison tests in the current literature (SAS Institute Inc, 1989b, p.947). The results of the analysis by REGWF are presented in the following manner. Individual mean values are assigned to a 'group', and means which are part of the same group are given the same letter of the alphabet (beginning with 'A'). If two or more means have the same grouping letter, the differences among them are considered by REGWF not to be statistically different. To take an example from a table which follows, more precisely table 7.9 on page 353, the mean values 38.32 and 36.93 were both given the grouping letter 'A', which means that according to REGWF the difference between these two values was not statistically significant. Still in the same table, the mean value 27.63 was given a 'B' for grouping letter, and this indicated that REGWF considered that 27.63 was statistically different from *both* 38.32 *and* 36.93. Moving to the precision values in the same table 7.9, REGWF assigned all six mean values to the same grouping, namely grouping 'A'. This indicated that according to REGWF there was no statistical difference between any pair of mean values, that is, 11.29 was not statistically higher than 11.05, and 11.05 was not statistically higher than 10.85, and so on.

The results of the comparison of means between link levels for LSM segmentation are presented in tables 7.9 to 7.11 (pp.353–355). The results for research articles appear in table 7.9 on page 353. For recall, the REGWF groupings indicate no difference between links $\geq$1 and $\geq$2, even though the mean percentage for links $\geq$2 seems higher than for links $\geq$1 (38.32% vs

36.93%). The remaining means are statistically different, as they are placed in separate groups by REGWF in ascending order. There is no single highest recall mean. For precision, all means are placed in a single REGWF group, which indicates that there is no statistical difference between them. As with recall, there is no single highest precision mean.

The results for the business report corpus are shown in table 7.10 on page 354. For recall, the REGWF groupings indicate that the highest mean is for links $\geq$2, with links $\geq$1 and links $\geq$3 tied in second place. For precision, REGWF indicates two overlapping groups: the first runs between links $\geq$1 and $\geq$4, and the second between links $\geq$2 and the rest. There is, therefore, no single highest precision mean, the general trend being again that higher link levels are associated with lower means.

Finally, the results for the encyclopedia articles are set forth in table 7.11 on page 355. For recall, the REGWF indicate that there is no difference between the two highest means, namely for links $\geq$1 and $\geq$2, indicating that there is no highest recall mean. For precision, there is a tie between the means for links $\geq$1, $\geq$2, and $\geq$3, also suggesting that there is no single highest precision mean.

In conclusion, in only one of the six situations was a single mean significantly higher than all the others, namely for business reports recall. On the whole, the results suggest that the highest rates are found between link levels $\geq$1 and $\geq$2. It appears, therefore, that there is no single link level which favours LSM, but LSM performance seems to be nonetheless aided by using link sets formed with lower linkage constraints.

In relation to the aims declared at the beginning of the analysis, there is firstly evidence that LSM segmentation performs better than random. It appears that LSM is a principled method for segmenting texts which takes

Ryan-Einot-Gabriel-Welsch Multiple F Test

Recall			
df=594, MSE=275.7007			
REGWF Grouping	Mean %	N	Link
A	38.32	100	≥2
A	36.93	100	≥1
B	27.63	100	≥3
C	18.59	100	≥4
D	11.89	100	≥5
E	4.42	100	≥6
Precision			
df=594, MSE=132.6509			
REGWF Grouping	Mean %	N	Link
A	11.29	100	≥1
A	11.05	100	≥5
A	10.85	100	≥2
A	10.01	100	≥3
A	9.89	100	≥4
A	7.45	100	≥6

Table 7.9: Comparison of means for LSM segmentation of research article corpus

Ryan-Einot-Gabriel-Welsch Multiple F Test

Recall			
df=594, MSE=160.5959			
REGWF Grouping	Mean %	N	Link
A	36.08	100	≥2
B	29.30	100	≥1
B	26.94	100	≥3
C	18.66	100	≥4
D	10.37	100	≥5
D	6.96	100	≥6
Precision			
df=594, MSE=314.324			
REGWF Grouping	Mean %	N	Link
A	27.18	100	≥2
B A	23.61	100	≥1
B A	23.32	100	≥3
B A	21.56	100	≥4
B	19.59	100	≥6
B	18.95	100	≥5

Table 7.10: Comparison of means for LSM segmentation of business report corpus

Ryan-Einot-Gabriel-Welsch Multiple F Test

Recall			
df=594, MSE=268.0451			
REGWF Grouping	Mean %	N	Link
A	38.70	100	≥2
A	35.35	100	≥1
B	22.67	100	≥3
C	4.90	100	≥4
C	1.63	100	≥5
C	0.31	100	≥6
Precision			
df=594, MSE=303.4363			
REGWF Grouping	Mean %	N	Link
A	21.66	100	≥2
A	19.76	100	≥3
A	19.01	100	≥1
B	8.77	100	≥4
C B	6.50	100	≥5
C	2.00	100	≥6

Table 7.11: Comparison of means for LSM segmentation of encyclopedia article corpus

into account the lexical cohesion across sentences, and whose results are better than might be expected by chance. Second, performance seems to be affected by link levels. Higher levels tend to blur the distinction between LSM and random segmentation, and LSM seems to perform best at the lower link levels. Finally, the advantage of LSM over random segmentation does not change in relation to text type, since LSM outperforms random segmentation in all three corpora. The next step will be to try to explain why LSM works. This requires a different kind of analysis than that provided by analysis of variance, an issue which will be tackled in the next section.

7.5 Multiple Regression

In trying to explain why LSM works, it is first necessary to identify which textual characteristics are related to higher segmentation performance. A statistical technique which is suited to this task is multiple regression, since it allows for the identification of linear relationships between a dependent variable and several independent variables. Regression analysis is particularly informative because it involves explanation as well as prediction; that is, through regression it becomes possible to locate those independent variables which can lead to a good prediction of the dependent variable, as well as identify the independent variables which cause the dependent variable (Lewis-Beck, 1980, p.20).

A few technical terms must be introduced at this stage. The dependent variable is also called the response variable, and it is the variable which one tries to predict; in our case, the interest lies in segmentation performance measures, and therefore the response variables will be recall and precision rates separately. The regressor variables are the independent variables which are used to try to predict the response variable. In our case, the initial

regressor variables relate to quantitative measures for each individual text, as listed below:

Sections Total sections;

Sentences Total sentences;

Boundaries Total boundary zones;

Links Total links;

Avg. Median Difference Average difference between subsequent link set medians;

Tokens Total running lexical words;

Types Total unique lexical words;

Links/Sentence Average links per sentence;

Tokens/Sentence Average tokens per sentence;

Types/Sentence Average types per sentence;

Sentence/Section Average sentences per section;

Links/Section Average links per section;

Tokens/Section Average tokens per section;

Types/Section Average types per section.

The above independent variables were selected because intuitively they seem to influence the performance of LSM segmentation in various ways. For example, texts with more sections might influence precision positively by offering more chances for the correct segment boundaries to be placed; conversely, texts with fewer sections might yield higher recall rates since fewer

section boundaries need to be recovered. Longer texts, as measured by the total number of sentences, might also influence performance differently. It might be posited that longer texts would induce lower precision, since as they may have more boundaries, there would be more chances for incorrect boundary decisions to be made. The same conditions might favour recall, since there would be more chances for finding more of the existing boundaries. As may have become apparent, there are multiple dependencies between the variables: the total number of boundaries seem to be influenced by the length of the text, which in turn influences the total sentences, which in turn influences the total tokens, which in turn influences the total types, which in turn influences the total links, and so on. One of the roles of regression analysis will be to identify only those variables which are significantly associated with performance.

The regressor variables are fitted in an equation such as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n + \varepsilon$$

The term to the left of the equal sign (Y) is the response variable. The first term to the right of the equal sign (β_0) is known as the intercept, and it indicates the average value of Y when each X equals zero (Lewis-Beck, 1980, p.19). The various other β indicate unknown parameters which will be estimated through regression analysis. Each X stands for an independent variable. The error term, ε, indicates the value not predicted by the model but needed in order for the value of the response variable to be predicted exactly.

Realistically, exact predictions in language are rare, and therefore an amount of discrepancy is expected between the actual and the predicted

values. That is why the regression model is normally represented as:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n$$

where $\hat{Y}$ represents the predicted value of the response variable, without the error term.

The overall fit between the actual and predicted values is indicated by the R^2 statistic, or 'coefficient of multiple determination', which measures 'the percentage of the variation in the dependent variable which is explained by variations in the independent variables taken together' (Schroeder et al., 1986, p.33). In our case, given that the dependent variables are recall and precision rates, and that the independent variables are quantitative textual features, R^2 measures the amount of variation in performance rates explained by variations in the quantitative textual features. The calculation of R^2 is independent of any of the terms explained so far, and it is carried out by dividing a measure of the total variation which the regression equation explains by a measure of the total variation in the dependent variable (Schroeder et al., 1986, pp.26-27). For a demonstration of how R^2 is calculated the reader is referred to standard references on regression analysis (Schroeder et al., 1986; Tabachnick and Fidell, 1989). What is more relevant for the purposes of the present investigation is the interpretation of R^2, which is straightforward: as said above, R^2 denotes the percentage of explained variation, and therefore one simply has to convert the value of R^2 to a percentage by shifting the decimal point to the right. Thus, to take an example from table 7.13 on page 363, the value of R^2=0.5304 for precision indicates that 53.04% of the variation in precision was explained by the regression equation.

The value of R^2 can be improved by removing outliers, or extreme cases which introduce bias. Outliers can be detected by a number of statistics, the

most common one being *Cook's distance*. For the analysis of the present data, outliers were eliminated when they scored higher than 1 on Cook's distance (Tabachnick and Fidell, 1989, p.130). Other measures employed to locate cases which influence the parameter estimates are 'DFFITS' and Student's R (SAS Institute Inc, 1989b, pp. 1371-1372); a cutoff point of 2 (≥ 2 or $\leq$-2) was set for these two statistics (SAS Institute Inc, 1989b, pp. 1418–1419). Once outliers were detected and removed, a final model was produced. The equations presented below as part of tables 7.13 to 7.15 (pp.363–365) are therefore based on data without outliers. In the data for the present study, there were 33 outliers, distributed as in table 7.12.

An improvement on the value of R^2 could also have been achieved by transforming the data so that they become normally distributed (Tabachnick and Fidell, 1989, pp. 72-86). If the data are ill-conditioned due to skewness (that is, the distribution is concentrated on end points) or kurtosis (that is, the distribution is too peaked or too flat), the fit will tend to be loose. The distribution of data could have been corrected by one of the traditional methods described in the literature, such as taking the square root of individual values or applying log transformations to the values (Tabachnick and Fidell, 1989); however, transformed data have the serious disadvantage of being 'robbed of the straightforward interpretation possible when the variable

Corpus	Rate	Total	Texts
Research	Recall	6	16, 47, 48, 51, 59, 74
Articles	Precision	4	11, 12, 76, 77
Business	Recall	4	35, 62, 76, 81
Reports	Precision	8	6, 24, 32, 34, 46, 47, 76, 81
Encyclopedia	Recall	5	10, 49, 64, 71, 79
Articles	Precision	6	19, 30, 44, 46, 49, 56
Total		33	

Table 7.12: Outliers

was measured in the original units' (Lewis-Beck, 1980, p.41), and therefore no transformations were applied to the data. As will become evident in the discussion of the results further below, this was a wise decision, since transformed data would have made the task of interpreting the multitude of parameter estimates considerably harder than it already was.

One important consideration in multiple regression is the selection of a final model. After the fact, not all independent variables will be found to be good regressors. A decision must then be made as to which subset of independent variables will form part of the final model. There are a number of selection methods for multiple regression (SAS Institute Inc, 1989b, pp. 1397-1399). Two of the most popular are backward elimination and maximum R^2 improvement. In backward selection, first all variables are fitted and then variables are extracted one at a time, and the resulting model is re-evaluated. The elimination stops when in the resulting model all regressors satisfy the criterion of being significant at a given α (usually 0.10). In maximum R^2 improvement, models are produced for all combinations of variables, ranging from one to as many independent variables as had been specified, in such a way that the resulting value of R^2 is as high as possible for that number of variables. The advantage of maximum R^2 improvement is that the resulting model explains the highest amount of variation; the disadvantage, however, is that the regressors are not necessarily significant, and therefore they do not help explain the response variable. In backward elimination all variables in the final model are significant, hence important for explaining the response variable, but the value of R^2 is not necessarily the highest possible. Since in theory it is always possible to increase R^2 by adding more variables, without necessarily adding to the validity of the model, maximum R^2 improvement does not seem to be as good a method as backward elimination for the present study. Thus, the selection method adopted for the present study is backward

elimination, which means that the final models only contain the smallest possible combination of regressors significant at α=0.10.

The results of the regression analysis of the individual corpora are presented in tables 7.13 to 7.15 (pp.363–365). Each table gives separate analyses for each response variable (a performance measure, either recall and precision). For each response variable two separate sets of results are presented. The first refers to the analysis of variance of the regression model; the value of F and the significance attributed to it are shown there, and they indicate whether the regression model as a whole is significant. The model itself is presented below the analysis of variance, where each regressor is listed (down the column marked 'variable') together with its estimate, F value and respective significance. Because a backward model selection method was employed only the final optimal model is presented. In each model, only those variables whose significance is p<0.1 are included.

A fitted equation follows the variables in each model, which is based on the parameter estimates listed in the table. The equation was used to produce predicted values for each response variable. The predicted values are reproduced in appendix 20 (p.497 ff.). To illustrate how the predicted values were obtained, let us take the recall value for text 28 from the research article corpus. The model for predicted recall of research articles in table 7.13 on the next page is 'Intercept + (Sections × -1.24051145) + (Avg.Median Difference × 0.97627129) + (Tokens/Sentence × 1.95466140) + (Tokens/Section × -0.03837662)'. The intercept equals 16.20936260 (see table 7.13 on the following page); the values for the other variables for research article 28 are given in appendix 20 on p.498: Sections = 6, Average Median Difference = 26.3, Tokens/Sentence = 17.8089, and Tokens/Section = 932.00. By substituting these values in the model, the following equation is obtained: 16.20936260

Recall					
R^2=0.1477					
	DF	Sum of Squares	Mean Square	F	p
Regression	4	4287.53591537	1071.88397884	3.86	0.0062
Error	89	24738.01065485	277.95517590		
Total	93	29025.5465702			
Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	p
Intercept	16.20936260	15.14098364	318.56529426	1.15	0.2873
Sections	-1.24051145	0.43579634	2252.21149840	8.10	0.0055
Avg.Median Diff.	0.97627129	0.32717931	2474.82302347	8.90	0.0037
Tokens/Sentence	1.95466140	0.93782430	1207.46591023	4.34	0.0400
Tokens/Section	-0.03837662	0.00984862	4220.43779496	15.18	0.0002

Predicted Recall=
Intercept + (Sections × -1.24051145) + (Avg.Median Difference × 0.97627129) + (Tokens/Sentence × 1.95466140) + (Tokens/Section × -0.03837662)

Precision					
R^2=0.5304					
	DF	Sum of Squares	Mean Square	F	p
Regression	9	2536.51414754	281.83490528	10.79	0.0001
Error	86	2245.73982642	26.11325380		
Total	95	4782.25397396			
Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	p
Intercept	6.25605816	4.83990124	43.63043531	1.67	0.1996
Sections	0.41895367	0.20112832	113.30432992	4.34	0.0402
Sentences	-0.05746629	0.02256301	169.39215752	6.49	0.0126
Avg.Median Diff.	0.40944807	0.20698848	102.18009075	3.91	0.0511
Links/Sentence	-0.17562089	0.08036505	124.70356236	4.78	0.0316
Tokens/Sentence	2.40006373	0.86455499	201.24323465	7.71	0.0068
Types/Sentence	-4.17530514	1.73100635	151.92880107	5.82	0.0180
Links/Section	0.00698532	0.00281422	160.88599652	6.16	0.0150
Tokens/Section	-0.07599469	0.02525566	236.43406808	9.05	0.0034
Types/Section	0.14180833	0.05208069	193.60248177	7.41	0.0078

Predicted Precision=
Intercept + (Sections × 0.41895367) + (Sentences × -0.05746629) + (Avg.Median Difference × 0.40944807) + (Links/Sentence × -0.17562089) + (Tokens/Sentence × 2.40006373) + (Types/Sentence × -4.17530514) + (Links/Section × 0.00698532) + (Tokens/Section × -0.07599469) + (Types/Section × 0.14180833)

Table 7.13: Multiple regression analysis of segmentation of research article corpus

Recall					
R^2=0.2080					
	DF	Sum of Squares	Mean Square	F	p
Regression	4	3433.82669296	858.45667324	5.97	0.0003
Error	91	13077.20775600	143.70557974		
Total	95	16511.03444896			
Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	p
Intercept	24.78348240	11.18027122	706.14455237	4.91	0.0291
Sentences	-0.34094337	0.07976298	2625.64126306	18.27	0.0001
Boundaries	0.89116918	0.19650590	2955.58029757	20.57	0.0001
Tokens/Sentence	1.76508922	0.88520254	571.37554810	3.98	0.0491
Types/Sentence	-4.00483352	2.06543412	540.28122113	3.76	0.0556

Predicted Recall=
Intercept + (Sentences × -0.34094337) + (Boundaries × 0.89116918) +
(Tokens/Sentence × 1.76508922) + (Types/Sentence × -4.00483352)

Precision					
R^2=0.4210					
	DF	Sum of Squares	Mean Square	F	p
Regression	3	5317.13015698	1772.37671899	21.33	0.0001
Error	88	7311.13155063	83.08104035		
Total	91	12628.26170761			
Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	p
Intercept	8.08316231	9.51906167	59.90681695	0.72	0.3981
Links/Sentence	-0.11755869	0.02743415	1525.55692487	18.36	0.0001
Tokens/Sentence	2.16410700	0.61135500	1041.05207436	12.53	0.0006
Types/Section	-0.33197056	0.04424542	4676.96729280	56.29	0.0001

Predicted Precision=
Intercept + (Links/Sentence × -0.11755869) + (Tokens/Sentence × 2.16410700) +
(Types/Section × -0.33197056)

Table 7.14: Multiple regression analysis of segmentation of business report corpus

Recall					
R^2=0.2768					
	DF	Sum of Squares	Mean Square	F	p
Regression	6	11684.17227652	1947.36204609	5.61	0.0001
Error	88	30526.94149191	346.89706241		
Total	94	42211.11376842			
Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	p
Intercept	-26.79169675	17.45695431	817.07911401	2.36	0.1284
Avg.Median Diff.	0.53104587	0.28911731	1170.35201404	3.37	0.0696
Links/Sentence	0.80556235	0.41933437	1280.20055925	3.69	0.0580
Types/Sentence	3.58460452	2.14655856	967.38118020	2.79	0.0985
Links/Section	-0.12737450	0.04576499	2687.19312806	7.75	0.0066
Tokens/Section	0.58576459	0.20139921	2934.48116842	8.46	0.0046
Types/Section	-0.73708833	0.36628032	1404.79362153	4.05	0.0472

Predicted Recall=

Intercept + (Avg.Median Difference × 0.53104587) + (Links/Sentence × 0.80556235) + (Types/Sentence × 3.58460452) + (Links/Section × -0.12737450) + (Tokens/Section × 0.58576459) + (Types/Section × -0.73708833)

Precision					
R^2=0.2978					
	DF	Sum of Squares	Mean Square	F	p
Regression	6	2943.49870941	490.58311824	6.15	0.0001
Error	87	6940.08249591	79.77106317		
Total	93	9883.58120532			
Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	p
Intercept	-10.57126135	8.27455466	130.19962747	1.63	0.2048
Sections	1.02746685	0.30260603	919.65696275	11.53	0.0010
Boundaries	-0.17669149	0.09885263	254.85911287	3.19	0.0774
Types/Sentence	3.80774237	1.09218759	969.58529987	12.15	0.0008
Links/Section	-0.02405868	0.00975453	485.26185485	6.08	0.0156
Tokens/Section	0.23567821	0.08680604	588.00951060	7.37	0.0080
Types/Section	-0.47857969	0.16943255	636.44406518	7.98	0.0059

Predicted Precision=

Intercept + (Sections × 1.02746685) + (Boundaries × -0.17669149) + (Types/Sentence × 3.80774237) + (Links/Section × -0.02405868) + (Tokens/Section × 0.23567821) + (Types/Section × -0.47857969)

Table 7.15: Multiple regression analysis of segmentation of encyclopedia article corpus

+ (6 × -1.24051145) + (26.3 × 0.97627129) + (17.8089 × 1.95466140) + (932.00 × -0.03837662), which is 16.20936260 - 7.4430684 + 25.675935 + 34.810369 - 35.766991, or 33.485607. The predicted recall value for research article 28 is thus 33.485607, or 33.4856 (shortened to four decimal places), which is the value given on p.498. The actual recall value for this text was 33.33, therefore the predicted value is only 0.15562 point away from the observed recall value. The difference between the observed and the predicted values is called 'residual', so the residual for research article 28 is 0.15562, which suggests a good fit between the prediction and the actual value. This residual is the lowest for recall in the research articles corpus.

The equations can be used to predict the likely segmentation outcome of other texts, although in reality they are only suitable for the samples from which they derive. To apply them to a different sample would be an exercise in extrapolation, which might result in a wrong forecast. The predicted values are merely illustrative, and are not in themselves important to the discussion.

In the presentation of the results, first a brief commentary will be provided on recall and precision regressors for each corpus. The focus of the initial interpretation will be on the *direction* of the influence of specific variables as revealed by the sign of the parameter estimate. Table 7.16 on the next page presents the interpretation associated with positive and negative values of individual parameter estimates. The labels indicating the intended interpretation of individual variables were meant to be self-explanatory. Thus, a positive value for 'Sections' indicates a 'sectionally dense text', that is, a text with a large number of sections, and a negative value for sections indicates a 'sectionally sparse text', or a text with few sections; a positive value for 'Sentences' indicates a text that is sententially long, whereas a negative value indicates a text that is 'sententially short'. The variable whose interpretation

Variable	Indication	
	Positive	Negative
Sections	sectionally dense text	sectionally sparse text
Sentences	sententially long text	short text
Boundaries	segmentally dense text	segmentally sparse text
Links	cohesively dense text	cohesively sparse text
Avg.Median Difference	cohesively unsettled text	cohesively settled text
Tokens	lexically long text	lexically short text
Types	lexically rich text	lexically poor text
Links/Sentence	cohesively dense sentence	cohesively sparse sentence
Tokens/Sentence	lexically long sentence	lexically short sentence
Types/Sentence	lexically rich sentence	lexically poor sentence
Sentences/Section	sententially long section	sententially short section
Links/Section	cohesively dense section	cohesively sparse section
Tokens/Section	lexically long section	lexically short section
Types/Section	lexically rich section	lexically poor section

Type-token interaction

Positive	Negative	Indication
tokens	types	lexically sparse text
types	tokens	lexically dense text
tokens/section	types/section	lexically sparse section
types/section	tokens/section	lexically dense section
tokens/sentence	types/sentence	lexically sparse sentence
types/sentence	tokens/sentence	lexically dense sentence

Table 7.16: Interpretation of sign of parameter estimates

is perhaps less straightforward is 'Avg. Median Difference', which is interpreted to indicate texts that are 'cohesively settled' or 'cohesively unsettled'. This variable, when positive, signals a text whose adjacent link set medians differ greatly, and when negative, it signals a text whose adjacent link set medians have nearly similar values. When adjacent link set medians are different, this suggests that adjacent sentences are to an appreciable extent cohesively dissimilar, thus creating what might be described as 'unsettled cohesion'. By contrast, a text whose adjacent link set medians are largely similar is a text whose adjacent sentences are largely similar to one another, a situation which might be seen to give rise to a more 'settled cohesion'. A pair of variables whose interpretation might be potentially misleading are 'tokens' and 'types'. Both refer to the number of lexical words only, that is, grammatical words were excluded from these counts. 'Tokens' refers to the number of running lexical words in each text, and indicates 'lexically long' or 'lexically short' texts, whereas 'types' refers to the total of different lexical words in each text, and indicates the size of the lexical vocabulary in each text, thus signalling 'lexically rich' and 'lexically poor' texts.

Note that the variables relating to counts of types and tokens by themselves indicate lexical richness and length respectively, but when they both appear in the same model they interact and are interpreted in terms of lexical density. After a brief interpretation of the parameters for each individual corpus, an interpretation of the overall trend of influence will be offered. In the following discussion of the contribution of individual variables to the models, variables will be referred to in small capitals (e.g. SENTENCES).

The values of R^2 for all corpora and performance measures are low, ranging from 0.1477 (research articles recall) to 0.5304 (research articles precision). These values indicate that the total variation explained by the models never exceeds 53% at best, and can be as low as 15%. A simple arithmetic

average of the six values is 0.3136, which means that on average 31.36% of the variation in segmentation is accounted for. These rates indicate that there is a large amount of performance unaccounted for by the models. This must be borne in mind during the interpretation of the models, since there are yet several influences on segmentation performance which have not been detected by the variables entered in the models. Realistically, this is not surprising, since it would be naive to expect that any one quantitative model would be able to account for the vast majority of a textual phenomenon as specific as segmentation. For larger scale phenomena, higher levels of variation have been reported, notably Biber (1988), whose analyses reportedly account for upwards of 80% of register variation (1988, p.127). For more specific phenomena, however, more modest values have been reported, such as Parsons (1990), whose analysis accounts for about 30% of the relationship between cohesion and coherence in student writing.

The regression analysis of the research article corpus is presented in table 7.13 on page 363. Higher recall seems to occur in texts that are sectionally sparse and cohesively unsettled, with sentences that are lexically long and sections that are lexically short. Higher precision seems to be more frequent in texts that are short, sectionally dense and cohesively unsettled. The sentences in these texts are cohesively sparse and lexically sparse (a combination of lexically long and lexically poor). And the sections in these texts tend to be cohesively dense and also lexically dense (a combination of lexically short and lexically rich).

The regressor and parameter estimates for the business report corpus is presented in table 7.14 on page 364. Higher recall seems to be associated with texts that are short, segmentally dense, and which have sentences that are lexically sparse (a combination of lexically long and lexically poor). Higher precision seems to be related to texts whose sentences are cohesively sparse

and lexically long, and whose sections are lexically poor.

Finally, the regression analysis of the encyclopedia article corpus is shown in table 7.15. Higher recall seems to be found in texts that are cohesively unsettled, whose sentences are cohesively dense and lexically rich, and whose sections are cohesively sparse and lexically sparse (a combination of lexically long and lexically poor). Higher precision seems to be associated with texts that are sectionally dense and segmentally sparse. As with recall, the sentences in these texts tend to be lexically rich, and the sections tend to be cohesively sparse and lexically sparse (a combination of lexically long and lexically poor).

The proliferation of regressors and the variation in the parameter estimates makes it difficult to identify a trend in the analysis. For example, sententially shorter texts (negative SENTENCES) seem to influence both higher precision (for articles) and higher recall (for business reports); lexically short sections (negative TOKENS/SECTION) seem to be associated with lower performance (lower recall and precision) of articles but also with higher performance (higher recall and precision) on encyclopedia articles; and precision on business reports seems to be affected by lexically poor sections (negative TYPES/SECTION) while precision on articles seems to be influenced by the opposite, lexically rich sections (positive TYPES/SECTION).

In order to avoid an *embarras de richesse*, attention will have to be focused on the major trends signalled by the regressors. Such trends can be observed by tabulating the frequency of each positive and negative influence on the individual variables across the corpora. Table 7.17 presents the frequency totals of negative and positive parameter estimates for each corpus and broken down by performance measure (recall and precision). Importantly, substantial frequencies are displayed in bold; these are cells which contain totals equal to or greater than the average. For the whole corpus,

Totals

	Recall		Precision		Grand Total	
Variable	+	−	+	−	+	−
Sections	0	1	**2**	0	2	1
Sentences	0	1	0	1	0	2
Boundaries	1	0	0	1	1	1
Links	0	0	0	0	0	0
Avg.Median Difference	**2**	0	1	0	**3**	0
Tokens	0	0	0	0	0	0
Types	0	0	0	0	0	0
Links/Sentence	1	0	0	**2**	1	2
Tokens/Sentence	**2**	0	**2**	0	**4**	0
Types/Sentence	0	1	1	1	1	2
Sentence/Section	0	0	0	0	0	0
Links/Section	0	1	1	1	1	2
Tokens/Section	1	1	1	1	2	2
Types/Section	0	1	1	**2**	1	**3**

The figures above are based on the tabulation below:

Sign of parameter estimates

	Research Articles		Business Reports		Encyclopedia Articles	
Variable	Recall	Precision	Recall	Precision	Recall	Precision
Sections	—	+				+
Sentences		—	—			
Boundaries			+			—
Links						
Avg.Median Difference	+	+			+	
Tokens						
Types						
Links/Sentence		—		—	+	
Tokens/Sentence	+	+	+	+		
Types/Sentence		—	—	—		+
Sentence/Section						
Links/Section		+			—	—
Tokens/Section	—	—			+	+
Types/Section		+		—	—	—

Table 7.17: Counts of positive and negative parameter estimates

the maximum value of any one row is six (i.e. three corpora × two performance measures, viz. recall and precision), hence the average is three (6 ÷ 2 cells); a value of three or higher therefore indicates a substantial frequency for the corpus. For individual measures, the cutoff for substantial frequencies is 1.5, since the highest value that any one row can achieve is three, as there is only one recall and precision count for each of the three corpora.

The substantial frequencies displayed in bold in table 7.17 on the preceding page indicate that there is no one single variable which influences all corpora on both measures. The only variable that comes close to being predominant is TOKENS / SENTENCE which is positively associated with four of the possible six conditions. Hence, if a single major influence were to be detected, that would probably be long sentences. It appears that the longer the sentences, the closer to the original sectioning the segmentation gets. However, it must be stressed that no single variable explains either recall or precision on its own. Quite the contrary, as the tables show numerous variables are needed to obtain approximations of recall and precision rates. Even so, as pointed out above, the total variation explained by the models never exceeds 53%, and therefore there is still a considerable amount of variation left unaccounted for. While this is far from ideal in statistical terms, in linguistic terms it is realistic, since linguistic and textual characteristics exhibit an enormous degree of variation, much of which is unknown.

Despite these problems, a few tentative trends can be observed. Across the three corpora, besides long sentences (positive TOKENS / SENTENCE), the other major influences are cohesively unsettled texts (positive AVG.MEDIAN DIFFERENCE) and lexically poor sections (negative TYPES/SECTION). Under these circumstances, it seems more likely that the segmentation will resemble the original sectioning of the text, either because a greater proportion of the total sections will be matched, or because more of the segment boundaries

will match the section boundaries. It can only be speculated on why these three characteristics seem to have an influence on higher performance.

Apart from long sentences, which influence recall and precision across the board, the other two major influences are associated with separate performance measures. Cohesively unsettled texts (positive AVG.MEDIAN DIFFERENCE) are influential on recall (frequency of 2), while lexically poor sections (negative TYPES/SECTION) are influential on precision (frequency of 2). Cohesively unsettled texts result from higher average median differences, which in turn reflect the degree of variation between adjacent sentences with respect to their lexical cohesion; in this manner, it would perhaps appear that cohesively unsettled texts create the conditions for more segmentation opportunities, thus increasing the likelihood of recovering more of the total sections.

Lexically poor sections, on the other hand, are associated with the higher probability of making a matching segmentation decision (precision). It is difficult to interpret this variable on its own. Nevertheless, it would appear that it is indirectly reflecting section length, since the ratio of types to section will be lower if the total of types is held constant while the number of sections rises. Such a scenario can be envisaged for longer texts, whose type count will not rise linearly with text length, whereas the section count may. In support of this interpretation is the fact that a major positive influence of SECTIONS is found on recall (frequency of 2), denoting that precision is aided by the presence of more section divisions. In short, the odds of making a higher proportion of correct segmentation guesses is greater if the number of sections is high. For instance, in an extreme case, if there are ten sections in a text and only one segment boundary is placed, that one segmentation decision may hit any one of the ten section boundaries and thus produce 100% precision. This is not the case with recall, though, since if there are

more sections, more matches will be needed for a high score to be achieved. Using the same extreme example, only 10% recall would be produced; if the situation were reversed though, say there was only one section but ten segments, with only one match the recall rate would then be 100%.

In summary, the analysis offered in this section has presented one attempt to explain why LSM segmentation works. This was done by singling out those textual characteristics which seem to result in higher correspondence between the original sectioning and the proposed segmentation. Six models deriving from multiple regression have been offered which on average account for about 31% of the variation in segmentation. The usefulness of the models themselves is that they can be utilized to predict the likely segmentation outcome of texts. More importantly, the variables associated with each model can be used to explain the textual conditions which give rise to higher segmentation performance. The major trends of influence on segmentation have been identified and discussed.

So far, the analysis has tried to explain segmentation success by identifying textual characteristics. In other words, the focus has been on predicting the segmentation success given certain characteristics in texts, be they intrinsic to the segmentation model (e.g. unsettled cohesion) or not (e.g. text length). Now it is necessary to focus on predicting key textual characteristics given the information that the segmentation procedure can provide. The priorities need to be reversed, so to speak. Among the host of potentially key textual characteristics, the most central would certainly be the position of sections in the text. It would be desirable to be able to predict section boundaries, that is, to be able to know which sentences are more likely to be section initiators. In other words, the main question asked so far in this chapter was 'here are the divisions that LSM made, are they section boundaries?', but it would also be legitimate to ask the question from the other

direction, that is, 'here are the section boundaries, would they be picked up?' If the answer is affirmative, then a strong relationship between lexical cohesion and textual organisation will have been found. In order to try to answer this question a different kind of analysis will need to be performed, a task which requires a section of its own.

7.6 Logistic regression

In the previous section multiple regression was applied to the problem of predicting performance scores and thus helping to understand which factors favour segmentation. The task now is to predict whether a sentence is a section boundary or not. In other words, the goal of this part of the analysis is to use information about lexical cohesion to find ways of estimating the probability that a sentence will be a section boundary. Since the problem, as in the previous section, is one of prediction, the statistical technique which best suits this problem is regression.

Linear regression, as applied in the previous section, requires that the variables be measured on a ratio or interval scale; in other words, the numeric values utilized to code them must be meaningful. For instance, if the variable text length is measured in total lexical tokens, a one-unit change in text length will mean a difference of one word; a two-unit change will mean a difference of two words, and so on. However, the variable needed for coding whether a sentence is a section boundary or not is different, since it can only take two values: yes or no. This type of variable is referred to as a 'dummy' variable, and the information it represents is discrete. Thus, linear regression cannot be used to predict section boundaries because SECTION is not a ratio or interval variable.

Logistic regression is a statistical procedure used for predicting values of

dependent variables measured on a binary scale, such as SECTION. In logistic regression one estimates the *probability* that an event will occur, for instance that a sentence will be a section boundary. The probability of an event is calculated by first obtaining the value of z:

$$z = \alpha + \beta_1 X + \beta_2 X + \ldots \beta_n X$$

The symbols in the formula refer to the parameter estimates, as in linear regression (see p.358): α is the intercept, and the various β refer to the estimates for the variables.

Once z is known, its value is substituted in the formula:

$$Prob(event) = \frac{e^z}{1 + e^z}$$

The resulting value is the probability of the event occurring. The method of calculating probabilities by applying this formula will be exemplified below after the variables are presented.

A number of variables present themselves as possibly influencing whether a sentence will be a section or not. Unlike in the previous investigation, where the variables were related to text characteristics but not necessarily reflecting knowledge about segmentation (e.g. text length), in this part of the main study the variables must reflect information that the segmentation procedure can provide. In this manner, if section boundaries are successfully predicted, this will indicate that LSM taps into vital information about how texts are organised to such an extent that it can make powerful predictions about the likelihood of a sentence being a section boundary given knowledge of the lexical cohesion of the text. If the prediction is not successful, then it could be concluded that LSM is not adaptable to the task of predicting section boundaries, despite being suited to producing segmentations which

resemble the original sectioning of texts. Among the variables that enter into the LSM algorithm the most crucial is the median, or the midpoint of a sentence's link set. By knowing the median, it becomes possible to calculate the median difference from a particular sentence to its predecessor or successor, which in turn leads to the average median difference, which is then used to assess whether each median deviates from the average, at which point the sentence is categorized as a segment candidate in a boundary zone or not. Therefore, the best choice of a variable for the current investigation is the median, since the other measures in LSM depend on it. The other LSM measures deriving from the median are not needed since they are derived in order to suggest possible segmentation points, which are superfluous to the current investigation.

The rationale behind LSM is based on contrasting medians, and so for the present investigation medians must be contrasted as well. However, LSM was restricted to one comparison at a time, namely between the current sentence and its successor. A more powerful contrast could be made if more medians were considered, for example by comparing the current median to a subset of its predecessors and to its successors. This would be particularly interesting since it would become possible to observe *patterns* instead of simple individual changes. Operationally, the patterns would be assessed within a *window* formed by a certain number of medians, for instance, three on either side (i.e. before and after) of each median. Three medians on either side seem a good number since it does not complicate the calculations or the interpretations too much. Logistic regression is expected to find those median patterns within the window which are indicative of section boundaries.

The window must be understood as a moving interval running along the length of the corpus covering seven sentences at any one time. The window moves one sentence at a time, so the very first window for each text will

extend from sentence 1 up to and including sentence 7; the second window will cover sentences 2 to 8; the third window will extend from sentence 3 to 9, and so on till the end of the text.

The data for the present study are the same as for the previous study, namely three corpora of 100 texts each (see previous chapter for descriptive statistics of the samples). The variables for this study and their respective labels are listed below; all eight variables were coded for each sentence.

- SECTION: Status of the sentence, dummy variable: section boundary (value 1) or non-section boundary (value 0);
- MEDIAN: The midpoint of the link set for each sentence;
- MEDIAN_{-1}: The previous median for each sentence;
- MEDIAN_{-2}: The previous median but one for each sentence;
- MEDIAN_{-3}: The previous median but two for each sentence;
- MEDIAN_{+1}: The following median for each sentence;
- MEDIAN_{+2}: The following median but one for each sentence;
- MEDIAN_{+3}: The following median but two for each sentence;

Since the windows are stepped up one sentence at a time, they overlap in all but one sentence. The same information is therefore encoded under a different variable label. What is MEDIAN_{-2} for window 1 is MEDIAN_{-3} for window 2; what is MEDIAN_{-1} for window 2 is MEDIAN_{-2} for window 3, and so on. This must be borne in mind when the sign of the parameter estimates are interpreted further below.

The boundaries between texts were maintained, since texts are independent of each other and their order in the corpus is random. This resulted in

missing values for window variables in several occasions. Missing values were generated for MEDIAN_1, MEDIAN_2 and MEDIAN_3 for the first sentence of every text, for MEDIAN_2 and MEDIAN_3 for the second sentence, and for MEDIAN_3 for the third sentence. A similar procedure was used for the last three sentences of each text.

Once these values had been computed for each sentence, logistic regression was run through the data using `SAS proc logistic`. As in the previous study, the backward model selection method was used (see p.361). This method chooses only those variables which significantly contribute to the model by shedding variables one at a time and computing the significance of the whole model. In the end, a final optimal model is returned. The results discussed below refer to the final models selected by backward elimination.

The results of logistic regression of the three corpora are presented in tables 7.18, 7.19, and 7.20 (pp.386–388). Each table presents three kinds of information. The first at the top ('sample size') refers to the size of each corpus in sentences and section boundaries. The middle table ('analysis of maximum likelihood estimates') presents the model itself, and includes values of the variables which significantly contribute to the prediction of section boundaries. Each variable is accompanied by its parameter estimate, standard error, Wald Chi-Square statistic and respective significance. At the bottom of the middle table appears the significance for the model as a whole, which is indicative of the joint significance of the variables. Since all variables in the models are significant, the only value which is of interest to the remainder of the chapter is the parameter estimates. They will be used in order to compute z (see p.376) and ultimately the probability of section boundaries. The bottom table ('association of predicted probabilities and observed responses') presents a comparison between the probabilities of a sentence being a section according to the model and its actual status. The

comparison is carried out based on checking whether the predicted probability of sections is greater than the predicted probability of non-sections for all pairs of sections vs non-sections. For instance, suppose the expected section probabilities of a short text had been computed as follows:

Sentence	Is it a Section?	Probability
1	1	0.04
2	1	0.03
3	1	0.02
4	0	0.01
5	0	0.03

The relevant section versus non-section comparisons would then be between sentences 1 and 4, 1 and 5, 2 and 4, 2 and 5, 3 and 4, and 3 and 5. If the predicted probability for the section sentence is greater than for the non-section sentence, the pair is counted as 'concordant'; if the probability is not greater for a section, the pair is counted as 'discordant'; if the probabilities are equal, the pair is counted as 'tied'. The pairs and respective comparisons for the data in the previous table would be:

Pairs	Concordant?
Sentence 1 (section) vs Sentence 4 (non-section)	Yes ($0.04 > 0.01$)
Sentence 1 (section) vs Sentence 5 (non-section)	Yes ($0.04 > 0.03$)
Sentence 2 (section) vs Sentence 4 (non-section)	Yes ($0.03 > 0.01$)
Sentence 2 (section) vs Sentence 5 (non-section)	Tied ($0.03 = 0.03$)
Sentence 3 (section) vs Sentence 4 (non-section)	Yes ($0.02 > 0.01$)
Sentence 3 (section) vs Sentence 5 (non-section)	No ($0.02 < 0.03$)

The percentage concordance rate of the above table can be calculated by dividing the number of concordant pairs (4) by the total number of pairs (6)

and multiplying by 100, which gives 66.67%. For the actual data, a non-parametric statistic (c, or 'coefficient of concordance', Siegel, 1975, p.223) is reported which indicates the predictive ability of the model. If c is significant, the model has successfully predicted which sentences are section boundaries. The concordant and discordant values in tables 7.18, 7.19, and 7.20 are displayed in two ways; the top half of the table shows the percentages including ties, while in the bottom half the count of ties has been split in two and each half added to the concordant and discordant totals.

The probabilities themselves are calculated as follows. For example, for articles the final model shown in table 7.18 (p.386) has the following intercept[3], variables and respective parameter estimates:

Variable	Parameter estimate
Intercept	-2.9437
Median_{-3}	0.00238
Median_{-1}	0.00238
Median	-0.00303
Median_{+1}	-0.00232

Sentences 34 to 38 of research article 68 have the following medians:

Sentence	Is it a Section?	Median
33	0	158.5
34	0	174.0
35	0	184.0
36	0	194.5
37	1	140.0
38	0	122.0

[3] As a reminder, the intercept indicates the average predicted value when each parameter estimate equals zero; see previous discussion on p.358.

Taking sentence 37 as an illustration, the values of the model variables for it are:

SECTION	1
MEDIAN_{-3}	174
MEDIAN_{-1}	194.5
MEDIAN	140
MEDIAN_{+1}	122

First the value of z is calculated by multiplying the parameter estimates by the values of each variable in the formula:

$$\begin{aligned} z &= Intercept + Median_{-3} \times 0.00238 + Median_{-1} \times 0.00238 + \\ &\quad Median \times -0.00303 + Median_{+1} \times -0.00232 \\ z &= -2.9437 + 174 \times 0.00238 + 194.5 \times 0.00238 + 140 \times -0.00303 + \\ &\quad 122 \times -0.00232 \\ z &= -2.77391 \end{aligned}$$

Next the value of z is substituted in the formula:

$$\begin{aligned} Prob(section_{sent37,article68}) &= \frac{e^z}{1+e^z} \\ Prob(section_{sent37,article68}) &= \frac{e^{-2.77391}}{1+e^{-2.77391}} \\ Prob(section_{sent37,article68}) &= 0.0588 \end{aligned}$$

The probability of sentence 37 being a section boundary is estimated at 0.0588. By contrast, the previous sentence (36), a non-section sentence, has the following variable values:

SECTION	0
MEDIAN_{-3}	158.5
MEDIAN_{-1}	184
MEDIAN	194.5
MEDIAN_{+1}	140

These values yield z=-3.04269, as:

$$\begin{aligned} z &= Intercept + Median_{-3} \times 0.00238 + Median_{-1} \times 0.00238 + \\ & Median \times -0.00303 + Median_{+1} \times -0.00232 \\ z &= -2.9437 + 158.5 \times 0.00238 + 184 \times 0.00238 + 194.5 \times -0.00303 + \\ & 140 \times -0.00232 \end{aligned}$$

Plugged into the formula for predicted probability, it yields a probability of 0.0455:

$$\begin{aligned} Prob(section_{sent36,article68}) &= \frac{e^z}{1+e^z} \\ Prob(section_{sent36,article68}) &= \frac{e^{-3.04269}}{1+e^{-3.04269}} \end{aligned}$$

The probability of being a section boundary is lower for sentence 36 (0.0455) than for sentence 37 (0.0588). In this case, the prediction is correct since sentence 37 is a section boundary and sentence 36 is not. In order to understand why this happened, it is necessary to introduce the concepts of *odds*. The odds of a section boundary occurring are defined as the ratio of the probability that a section boundary will occur to the probability that it will not occur (Norusis, 1990, p.49). The parameter estimates indicate the change in the *log odds* associated with a one-unit change in one of

the variables. For example, for research articles the parameter estimate for MEDIAN$_{-1}$ is 0.00238; this indicates that if the value of the previous median increases by 1 (and the other variables do not change between the two sentences) the log odds of that sentence being a section boundary will increase by an amount equal to the parameter estimate (i.e. 0.00238). In other words, if the median of the preceding sentence is higher than the current median, the log odds of that sentence being a section boundary increases. For MEDIAN, the parameter estimate is -0.00303, therefore a drop in the value of the median (as compared to the previous median) indicates an increase in log odds by the same amount, that is, 0.00303; phrased in another way, a rise in the value of the median indicates a decrease in log odds by 0.00303. For MEDIAN$_{-3}$, the parameter estimate is 0.00238, hence a rise in the median for the preceding sentence but two will raise the log odds of the current sentence being a section boundary by 0.00238. Finally, for MEDIAN$_{+1}$, the parameter estimate is -0.00232, which indicates a rise in log odds by 0.00232 if the following median is lower than the current one. In short, by interpreting the parameter estimates in terms of log odds, it is possible to estimate the pattern which would be indicative of section boundaries in research articles as:

- higher previous median but two MEDIAN$_{-3}$.
- higher previous median MEDIAN$_{-1}$;
- lower median (MEDIAN);
- lower following median MEDIAN$_{+1}$.

In the case of sentence 37 of research article 68, all of these conditions obtained:

Variable	Sentence		Sentence 37 is:
	36	37	
	Non-section	Section	
MEDIAN_{-3}	158.5	174.0	higher
MEDIAN_{-1}	184.0	194.5	higher
MEDIAN	194.5	140.0	lower
MEDIAN_{+1}	140.0	122.0	lower

Of course this is the result of one single comparison of the millions needed to assess the predictive power across each corpus. For the research article corpus, for example, over 18 million paired comparisons were made. From these, it was found that 48.5% of the time the probabilities associated with section boundaries were higher than for non-sections, while 38.8% of the time the probabilities of non-section boundaries were higher; for a further 12.7% of pairs the probabilities were tied.

Now that the central concepts in logistic regression needed for predicting the probability of section boundaries have been presented and illustrated, attention can be turned to the presentation and interpretation of the results for each corpus. The results for individual corpora appear in tables 7.18 to 7.20 (pp.386 to 388).

The results for research articles appear in table 7.18 on the next page. The sign of the parameter estimates in the 'analysis of maximum likelihood estimates' table indicates that a sentence has a greater probability of being a section if the previous median (MEDIAN_{-1}) and the previous median but two (MEDIAN_{+1}) are higher while the current median (MEDIAN) and the following median (MEDIAN_{+1}) are lower. The 'association of predicted probabilities and observed responses' table shows a significant association between actual

Sample size

Sentences	20,090
Sections	940
Non-sections	19,150

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	p
Intercept	1	-2.9437	0.0636	2139.5825	0.0001
Median_{-3}	1	0.00238	0.00108	4.8785	0.0272
Median_{-1}	1	0.00238	0.00111	4.6130	0.0317
Median	1	-0.00303	0.00113	7.1324	0.0076
Median_{+1}	1	-0.00232	0.00111	4.3607	0.0368
Chi-squared score = 17.108 with 4 DF (p=0.0018)					

Association of Predicted Probabilities and Observed Responses

Concordant	48.5%
Discordant	38.8%
Tied	12.7%
Pairs	18,001,000
c	0.549
p<0.0001	
Split Ties	
Concordant	54.85%
Discordant	45.15%

Table 7.18: Logistic regression analysis of research article corpus

Sample size

Sentences	14,631
Sections	1,741
Non-sections	12,890

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	p
Intercept	1	-1.7973	0.0489	1351.7469	0.0001
Median_{-1}	1	0.00283	0.000986	8.2189	0.0041
Median_{+1}	1	-0.00234	0.00105	4.9774	0.0257
Median_{+3}	1	-0.00244	0.00101	5.8616	0.0155
Chi-squared score = 33.490 with 3 DF (p=0.0001)					

Association of Predicted Probabilities and Observed Responses

Concordant	52.3%
Discordant	43.5%
Tied	4.2%
Pairs	22,441,490
c	0.544
p<0.0001	
Split Ties	
Concordant	54.4%
Discordant	45.6%

Table 7.19: Logistic regression analysis of business report corpus

Sample size

Sentences	9,743
Sections	956
Non-sections	8,787

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	p
Intercept	1	-2.2429	0.0529	1798.2826	0.0001
Median_{-1}	1	0.00243	0.00095	6.5130	0.0107
Median_{+1}	1	-0.00220	0.000974	5.1281	0.0235
Chi-squared score = 6.531 with 2 DF (p=0.0382)					

Association of Predicted Probabilities and Observed Responses

Concordant	44.3%
Discordant	42.2%
Tied	13.6%
Pairs	8,400,372
c	0.511
p<0.0001	
Split Ties	
Concordant	51.1%
Discordant	48.9%

Table 7.20: Logistic regression analysis of encyclopedia article corpus

and predicted sections (c=0.549, p<0.0001), with the majority of pairs being concordant (48.5% including ties, or 54.85% excluding ties).

The results for business reports are presented in table 7.19 on page 387. The table 'analysis of maximum likelihood estimates' indicates that a sentence has a greater probability of being a section if the previous median (MEDIAN$_{-1}$) is higher while the current median (MEDIAN) and the following median (MEDIAN$_{+1}$) are lower. The 'association of predicted probabilities and observed responses' table suggests a significant association between actual and predicted sections (c=0.544, p<0.0001), with the majority of pairs being concordant (52.3% including ties, or 54.4% excluding ties).

Finally, table 7.20 on the preceding page sets forth the results for encyclopedia articles. According to the parameter estimates in the 'analysis of maximum likelihood estimates', a sentence has a greater probability of being a section if the previous median (MEDIAN$_{-1}$) is higher and the following median (MEDIAN$_{+1}$) is lower. There is a significant association between actual and predicted sections (c=0.511, p<0.0001) according to the 'association of predicted probabilities and observed responses' table; the majority of pairs are concordant (44.3% including ties, or 51.1% excluding ties).

A pattern can be induced from the parameter estimates by noting the sign of each parameter estimate and interpreting it as a *change in relation to the subsequent position* in the text. As noted above (p.378), the windows are overlapping, therefore the parameter estimates for a given sentence must be interpreted in relation to adjacent sentences. A positive sign indicates a rise from one median to the next; a negative sign represents a drop between the medians of adjacent sentences. So, if the parameter estimate is negative for, say, MEDIAN$_{+1}$, this means a drop between positions relative to MEDIAN$_{+1}$ and MEDIAN$_{+2}$. Note that in this case the variable MEDIAN$_{+2}$ does not need to be listed among the significant variables since it is presupposed by the fact

that it is the position that follows MEDIAN_{+1}. In terms of the sequence of sentences in the text, this would mean that a sentence has a greater likelihood of being a section boundary if the sentence immediately ahead of it has a lower mean than the next sentence but one.

Research articles
$\underbrace{-3 > -2}_{0.00238}\ \underbrace{-1 > M}_{0.00238}\ \underbrace{M < +1}_{-0.00303}\ \underbrace{+1 < +2}_{-0.00232}$
Business reports
$\underbrace{-1 > M}_{0.00283}\ \underbrace{+1 < +2}_{-0.00234}\ \underbrace{+3 < +4}_{-0.00244}$
Encyclopedia articles
$\underbrace{-1 > M}_{0.00243}\ \underbrace{+1 < +2}_{-0.00220}$

Table 7.21: Patterns for section boundaries

Table 7.21 lists the patterns which increase the odds of a sentence being a section for each corpus. The variables are arranged in sequential order with their respective parameter estimates in subscript. The patterns are similar for the three corpora: a continuous rise in median values before the section boundary followed by a drop after the boundary.

The interpretation of these patterns needs to be careful since the absolute values of the medians are meaningless. When speaking of a median being higher than another it does not matter which value is assigned to the higher or the lower median so long as these values reflect this difference. It must be recalled that the median is simply a measure of the central tendency of the link sets, in other words, they are a shortcut for comparing link sets. Similar medians are taken to be indicators of similar link sets, which in turn

indicate similarity between the lexical cohesive profiles of two sentences. On the whole, the patterns suggest that there is a shift in the lexical cohesive pattern of sentences as they approach a section boundary. Sentences immediately before the section boundary seem to be characterised by linking further forward in the text than sentences immediately after the section boundary (i.e. the medians for pre-boundary sentences are higher than for boundary or post-boundary sentences).

In summary, in this part of the chapter a logistic regression analysis was applied to the problem of predicting section boundaries. At the end of the previous section, the question that was asked was whether it would be possible to predict section boundaries beyond what is expected by chance. The results of the logistic regression offer an affirmative answer, indicating that there is a higher significant probability for section breaks to occur if the flow of medians across the text follows a certain pattern. Since the information used to predict the section boundaries was derived from LSM, there is evidence that LSM taps into vital information about how texts are organised, which partly explains why it is able to provide segmentation of texts which resemble the original author's divisions. The immediate implication of the successful prediction of sections is that there is some evidence that sections are not arbitrary. Rather they seem to correlate with linguistic features. The implications of the results reported in this chapter will be discussed in full in the next chapter.

7.7 Conclusion

This chapter has presented the results of the main large-scale study into segmentation. Three corpora were investigated, each with 100 texts of three different genres. The need for the large-scale investigation was justified on

the grounds that it would be desirable to test LSM on a larger set of texts than the sample used in the previous chapter (25 texts). In this manner, a more comprehensive investigation into the power of the segmentation procedure was carried out. Furthermore, by applying LSM to different genres it was possible to assess the performance of LSM in segmenting three different genres. The results of the first study reported in this chapter indicated that LSM performed better than random equally for the three genres. The second study found that higher linkage levels for link set formation tended to worsen the performance of LSM. Having established that LSM worked, the third study went on to investigate some of the possible causes of successful segmentation by examining a number of textual characteristics and their correlation with performance. It was found that certain characteristics such as sentence length, section length, and section density tended to be associated with higher performance across the three genres. The last study looked at the issue of predicting section boundaries from lexical cohesion using information provided by LSM. The results indicated that section breaks seem to be associated with a particular pattern of lexical cohesion alternation across sentences. The next chapter will provide a discussion of the findings so far by placing the results of the investigation in the context of previous research.

Chapter 8

Discussion

The implications of the results obtained in the main study of segmentation presented in this thesis will be discussed in this section. The discussion will centre on the impact of the findings on previous research in discourse analysis, segmentation by computer, and lexical cohesion, and on the status of sections as linguistic units.

8.1 Discourse analysis and segmentation

The present study has implications for the way texts are analysed in discourse analysis. The major finding is the confirmation that texts are segmentable by computer. Models of discourse have been built on the tacit assumption that texts are divisible; other procedures have also been proposed for segmenting texts manually on the understanding that texts are segmentable and segments relate to linguistic cues (e.g. Cloran, 1995; Passonneau and Litman, 1993). Therefore, in a sense the present investigation adds nothing new to discourse analysis. However, the possibility of segmenting texts reliably on the computer by drawing on linguistic expression has important implications for the way text organisation is seen in the literature. These implications

will be discussed in what follows.

Before addressing the implications that segmentability has for discourse analysis, it is useful to present a summary of the main points put forward before in chapter 2. This chapter presented a review of major models for analysis of discourse and interpreted them as suggesting ways of segmenting texts into discrete units. The chapter concluded that although there was wide variation across the discourse analytical approaches, some trends could be observed across the various approaches which have a bearing on their application to textual segmentation. The first trend was described as 'labelling'. Discourse units obtained by using these models are given a label which suggests their function or place in the model. A key aspect of labels is that they help the analyst find similar units in other individual texts. The second trend was that models are deductive; that is, they make *predictive* statements about the recurrence of segments in other texts. In other words, models are not meant to be applicable only to the source data from which they derived. Rather, they are expected 'to work' with other texts as well. The third trend was referred to as the lack of validation of segments, as the results of the application of models are not checked against a reference. It was claimed that validation could be achieved by checking the results of the analysis with other possible analyses of the same data. In practice, this would be rather difficult to implement, since there is no agreed 'correct' analysis which could serve as a reference. The fourth trend noted was that the majority seemed to have been tried and tested on a restricted number of actual texts, which has implications for the validity of the models. Finally, the fifth trend which was observed across the various approaches was that most models are derived from analysis of short texts. This seems to reflect the subjective and non-computational orientation of discourse analysis. The remark by Phillips (1985) that a blackboard-full of data is the most that

traditional linguistics can cope with also holds true for discourse analysis.

Discourse analytical approaches are model-based, that is, they are centred around the notion of a finite set of categories which are meant to explain how certain text types work. To sum up, discourse analytical models have five characteristics: labelled constituents, deductive orientation, unvalidated analysis, restricted coverage (model derives from examination of little amount of data) and restricted applicability (model is meant to be used on short texts).

The computer-based approach experimented with in this thesis offers another perspective from which to analyse texts. Firstly, the use of computers opened up the possibility of investigating a large quantity of data, which, according to Phillips (1985), is an element too often overlooked in discourse analysis. Secondly, the LSM procedure allowed for a bottom-up analysis of texts, which in turn met the recommendation made by Phillips (1985) that the analysis be as free as possible from *a priori* categorization. The fact that the analysis was model-free, that is, it did not depend on a finite pre-defined set of text constituents, made it more versatile in that it was not restricted by those constituents predicted by the model. In this sense, analysis by LSM segmentation goes some way towards meeting Sinclair's recommendation that analysts should 'refrain from imposing analytical categories from the outside' (Sinclair, 1991, p.29). The disadvantage is that computerized analysis of segmentation cannot explain what each segment means, and therefore cannot present a unified picture of the text as model-based analyses can. Segments need to be interpreted by the analyst, as attempted in the sample analysis provided in section 6.18 (p.317 ff.). The analysis showed that there are a number of possibly interesting issues relating to sectioning by humans versus segmenting by computer which could be explored further.

The opposition between model-based and model-free approaches to dis-

course analysis referred to here has a parallel in the distinction between discourse structure and discourse organisation proposed by Hoey (1991b). According to Hoey (1991b, p. 13), several approaches to discourse (e.g. Graustein and Thiele, 1983; Longacre, 1983; van Dijk, 1980) centre around the notion that texts can fully described in terms of structures, the main characteristics of which being their inviolability and 'ability to specify impossible combinations' (p. 201). The sort of picture of text provided by the analysis of patterns of repetition amongst sentences is of a different kind: instead of making predictions, it accounts for probabilities (p.194).

Instead of hierarchies, texts can be seen as patterns, as those extracted for the probabilities of sentences being section boundaries (see p.389). Given that these patterns are probabilistic, they fit in with an organisational view of text (Hoey, 1991b), and not with a structural description of text, since the patterns do not predict with certainty which segments are sections. As such, the view of text subscribed to by LSM segmentation is essentially organisational, since it does not 'risk declaring which combinations of elements could not occur together in a text' (Hoey, 1991b, p.204).

Segmentation is essentially of a probabilistic nature, an assumption which was taken into account when the decision was taken to investigate the probability of section breaks by means of logistic regression (see section 7.6, pp.375 ff.). The decision lies with the author, who takes into account, among other aspects, the interaction with the reader, generic constraints, and layout conventions as fundamental in deciding whether to segment. Writers have a choice of segmenting at certain places; some of these opportunities will be taken up, while others will not (Hoey, personal communication). The probabilistic nature of LSM segmentation is in agreement with Goutsos's (1996a) argument that segmentation must not be understood in absolute yes or no terms.

The implication of segmentation being an organisational description is that it does not support a macrostructural view of discourse organisation. As Goutsos (1996a) explains, a macrostructural view is based on the notion that discourse is organised at a 'deep' or 'schematic' level (e.g. Graustein and Thiele, 1983; van Dijk, 1980). Macrostructural approaches are 'schematic' or 'propositional', that is, they provide analyses based on 'the semantic relations between constitutive units of predications or propositions' (Goutsos, 1996a, p.502). Van Dijk (1980), for example, sees discourse as made up of micro-propositions which fuse into larger macropropositions which in turn fit into a superstructure. Similarly, Pitkin (1969, p.142) proposes discourse blocs as 'the smallest unit to have a discrete function in the discourse', and sees discourse blocs as joining each other into larger units which eventually comprise the whole text. One of the reasons why such views have held out for so long is that their rationale derives from sentence grammar. Incidentally, the early efforts by van Dijk (1972) were labelled 'text grammar', and although his approach has changed over time, the hook-up to sentence grammar is still evident. These models date back to a time when a debate was going on about whether sentences and texts were essentially the same entities in different lengths or whether they differed on other significant dimensions (cf. Petöfi, 1979; Petöfi, 1982), so that a 'grammar' of texts could be developed (cf. Petöfi and Rieser, 1973). These models take grammar as a metaphor for text (Hoey 1991b, p.203). However, as Halliday and Hasan (1976, pp. 7-9) had already suggested, the kinds of grammatical relations which occur within the sentence are not found beyond the sentence, with the implication that a grammatical metaphor cannot adequately account for text. Halliday and Hasan (1976) propose 'texture' as a construct which accounts for the relationships found beyond the sentence within text.

Sequential models, by contrast, are based on the analysis of the role

of surface features of text as indicators of textual relations (Goutsos, 1996a, p.529). Sequential relations are not subordinated to topic; instead perception of topic results from the functioning of sequentiality (p.529). Importantly, Goutsos (1996a) builds these insights into a description of text in terms of segments. In this sense, the results of the study presented here can be seen as another example of studies which seek to describe the sequential, rather than the macrostructural, organisation of text.

The kind of macrostructural description provided by Phillips (1985, 1989) is of a different kind. Phillips (1985) terms 'macrostructure' the description of the 'global pattern of textual organisation' (p.4) or the 'overall pattern of connectivity among chapters' (p.165) formed by the linkage between the individual collocational networks found in separate chapters of textbooks. The approach followed by Phillips (1985, 1989) is in one way the opposite of that adhered to in the present investigation. Phillips first computed lexical cohesion in intervals smaller than the text (chapters) and then looked for links between individual intervals to see whether groupings, or *segments* of chapters as he called them, were formed. In the present study, the web of cohesion was first computed for the whole text and then ways of optimally breaking it up were sought, resulting in *segments*. In another way, though, the two approaches are similar, in that in the present study the cohesion was computed between sentences and then groupings of sentences were spotted based on their cohesive similarity, much in the same way as Phillips did. Therefore, according to the latter interpretation at least, the current study has also looked at macrostructure in the way Phillips (1985, 1989) did.

Discourse models and computer segmentations can be seen as providing complementary perspectives on the way texts are organised. Segments provide a means of knowing where the major natural divisions of the texts are, whereas discourse constituents can provide an interpretation of what the

divisions mean. It would be potentially interesting to see, for example, what kind of clause relations (Hoey, 1983), macrostructures (van Dijk, 1980), generic structure elements (Hasan, 1989), or rhetorical structure relations (Mann and Thompson, 1986b, 1987a; Mann et al., 1989) are related to the segments found by computer. Hoey (personal communication, 1996) suggests that segments might be related to sequence relations, for instance Problem–Solution. Alternatively, it would also be interesting to see how closely segments match the divisions produced by application of discourse models.

An approach to discourse organisation which embodies segmentation is Grosz and Sidner's (1986) theory of discourse. Their theory holds a central place for segments, which are seen as natural components of discourses: 'just as the words in a single sentence form constituent phrases, the utterances in a discourse are naturally aggregated into *discourse segments*' (Grosz and Sidner, 1986, p.177, original emphasis). Their theory has had an appeal to computational linguistics because it is formulated in computational terms (Mann and Thompson, 1987a, p.42). While the present research did not follow their model, it was influenced indirectly by their concern with the computational aspects involved in investigating discourse segmentation.

The emphasis placed throughout the pilot and main study on achieving reliability in segmentation through the use of computers might be taken to mean that the approach followed in the present investigation is entirely incompatible with subjective approaches to discourse organisation. Genre analysis, for example, is an influential methodology for doing what is essentially the segmentation of texts (Hopkins and Dudley-Evans, 1988; Hyland, 1990; Marshall, 1991; Nwogu, 1991; Salager-Meyer, 1989, 1990; Swales, 1981, 1990; Tinberg, 1988). Although informed by linguistic evidence[1], genre ana-

[1] A concession with which Paltridge (1994) disagrees, arguing instead that genre analysis is essentially driven by content (see section 2.3.4 on p.37).

lysis is largely based on intuitive judgement. The main units in genre analysis are 'moves', which are delimited by interpreting the communicative function of a portion of discourse. In proposing moves, the analyst is guided by the understanding that discourse is 'reader's discourse', that is, discourse is created 'as a result of the reader's interpretation of the text' (Bhatia, 1993, p.8). Nevertheless, there is an important point of contact between computational segmentation and genre analysis, namely the very fact that in both approaches texts are seen as being constituted by discrete contiguous blocks; in segmentation these are called 'segments', and in genre analysis they are named 'moves'. Of course, there too many differences between moves and segments to allow them to be considered a reflection of each other. One crucial difference is that moves are seen as an element with a clear function in the whole generic structure, normally expressed by a label attached to them (e.g. 'introducing purpose' in research articles, or 'ending politely' in a sales promotion letter). Segments do not need to be seen as playing a functional part in the whole in order to be located. Nevertheless, the two approaches are not conflicting; rather they present complementary views on what is basically the same phenomenon, namely the principled division of texts into componential units.

No attempt has been made in this thesis to explore to what extent current models of discourse are reflected in the segmentations proposed by LSM. To have done so would have implied a departure from the original aims of the studies reported here which were concerned with developing a segmentation procedure based on lexical cohesion. Now that LSM has proved a successful segmentation procedure, future research can build on that and explore the relationship between segmentation and existing models of discourse organisation.

8.2 Computers and segmentation

With the development of a specific segmentation procedure ('Link Set Median', or 'LSM'), the main study reported in this thesis makes a contribution to previous research in computational segmentation. Chapter 3 reviewed the most important segmentation algorithms developed to date. It was argued that the studies on segmentation shared three features in common: use of computers, reliance on lexical cohesion, and application of mathematical models. Most studies make use of computers in order to segment texts. This suggests that those areas which have taken a greater interest in segmentation are exactly those which are occupied with developing computer applications and which see segmentation as just one more task which the computer can be programmed to carry out with relative ease. In this sense, it is not surprising that advances in computer segmentation have not fed into discourse analysis; previous research in computer segmentation has largely ignored what segmentation means to the issue of how texts are organised, having focused on how segmentation can be programmed. In the process of making the segmentation algorithm work, many decisions are taken which are arbitrary to the discourse analyst, such as the use of even-sized portions of text (Hearst, 1993; Salton and Buckley, 1991; Salton et al., 1994) instead of meaningful units such as the sentence or the clause. The chapter concluded that this state of affairs is unfortunate, as finding out more about segmentation should also provide ways of understanding how texts work, and presumably also the segmentation is more likely to be useful if it approximates to some reality in the texts. The chapter recommended that ways should be sought to reconcile computer-aided segmentation and discourse studies; one of the ways would be to devise a segmentation procedure which avoided arbitrary decisions and was informed by discourse considerations.

The LSM procedure proved successful as a segmentation algorithm, since

it managed to find more target segments (sections) than expected by chance. The performance rates of LSM are lower than the top segmentation algorithm available in the literature (TextTile); however rates *per se* are not a preoccupation of the present investigation, as these can be improved by tweaking certain parameters of the procedure, notably by forcing segment boundaries to fall only between paragraph breaks, which is where section breaks will always occur. This is as we have noted the strategy used by TextTile. Significantly, LSM and TextTile tend to find different segments; that is, there are many target segments that only LSM finds, and equally there are many others that only TextTile recovers. Therefore, in a sense there is no 'better' procedure, as they can be seen to complement each other.

The fact that the results achieved by LSM did not depend on using a semantic database indicates that relying on repetition alone may suffice for segmentation. This corroborates previous research by Hearst (1994a) who found no substantial increase in performance of TextTiling by adding a thesaurus. The present research also corroborates previous research by Youmans (1991) who provided segmentation of texts by computing repetition alone; Youmans (1991) found no performance gains after lemmatizing the words in his texts. LSM offers an alternative to previous research by Kozima (1993b), Morris (1988, 1991) and Okumura and Honda (1994) who resorted to a thesaurus and a dictionary for assessing similarity between words. Although a dictionary and a thesaurus would help pick up complex repetition (e.g. politician (N) – political (Adj)), as well as simple and complex paraphrase (e.g. sedating (V) – tranquilized (V)), and hot (Adj) – cold (Adj), respectively), or even hyponymy (e.g. 'bear' and 'animals'), the gain in performance would not necessarily be proportional to the effort needed to create and fine-tune a database. As Benbrahim (1996) notes, on average the majority of links (about $\frac{3}{4}$) are formed by simple repetition; in his research, the addition of

complex repetition links only increased coverage by 16%, while simple mutual paraphrase contributed 6% (see table 4.2 on page 167). Considering that certain complex repetition links can be picked up without a database by simply stemming suffixes (e.g. human*s* (N) = human (N) – human (Adj)), it is only a minority of the links which truly necessitate a thesaurus.

The success of LSM partly depends on the fact that lexical cohesion is a suitable measure for use in computer-aided segmentation. As previous research has suggested, from a computational point of view, lexical cohesion can be successfully accounted for on the computer (Kozima and Furugori, 1993; Kozima, 1993a,b; Morris and Hirst, 1991; Morris, 1988), and from a discourse point of view, it is a relevant descriptor of how texts and parts of texts hang together (Halliday and Hasan, 1976; Hasan, 1989, 1984; Hoey, 1988, 1991a). Therefore, LSM provides further evidence that lexical cohesion can be profitably explored as a means of describing the internal divisions within written texts.

One important contribution of LSM to the literature on computer-aided segmentation is that it shows that a computational methodology does not necessarily have to part with certain key principles suggested by non-computational research into text organisation. As noted above, previous research in segmentation has on several occasions made arbitrary choices purely on the grounds that they would make the segmentation algorithm work better. For instance, Hearst and Plaunt (1993) use even intervals based on average sentence and paragraph lengths. Youmans (1991) uses several fixed word intervals (e.g. 35 words) in Vocabulary Management Profile (VMP). Kozima (1993a) employs 51-word windows in 'Lexical Cohesion Profile' (LCP). Reynar (1994) trains his segmentation procedure ('dotplot') by finding boundaries between texts; in other words, text internal boundaries are treated as no different from divisions between whole texts, which may make

sense to the computer but is certainly highly questionable from a discourse point of view. The development of LSM, on the other hand, centred around the notion that lexical cohesion between sentences is indicative of meaning sharing (Hoey, 1983, 1991b; Winter, 1974); therefore finding those sentences which are more similar in lexico-cohesive terms ought to bring out those text parts whose meaning is shared internally. LSM never departed from this principle, but acknowledging that comparing the lexical cohesion between sentences is a complex task, I decided on comparing sentences on the basis of a central element in their lexical cohesion makeup, namely the median of the set of sentences with which they link.

8.3 Lexical cohesion

The present investigation also contributes to current research on lexical cohesion. Chapter 4 reviewed the main approaches to the study of lexical cohesion. Two main strands were identified: studies which viewed lexical cohesion as lexical chains or strings (e.g. Eggins, 1994; Halliday and Hasan, 1976; Hasan, 1989; Halliday, 1985; Parsons, 1990), and studies which approached lexical cohesion as clusters (e.g. Hoey, 1991b,a). The lexical chains approach has been used in computer segmentation procedures (Morris, 1988; Morris and Hirst, 1991); two problems associated with it are the need for assistance from a thesaurus (e.g. Stairmand, 1996a; Stairmand and Black, 1996), and the increased difficulty in finding boundaries in longer texts (Hearst and Plaunt, 1993). The cluster approach has also been shown to be computerisable (Benbrahim, 1996; Benbrahim and Ahmad, 1994; Berber Sardinha, 1995e; Collier, 1994) but it had not been applied to segmentation in studies independent of the present thesis. In a sense, Hoey's methodology was designed to demonstrate the opposite of segmentation, that is, that texts form

an integrated whole. The successful application of his method to segmentation suggests that it can be extended to investigate the ways in which texts are divided into parts. The fact that the method proposed by Hoey (1991b) for the analysis of text unity was utilized successfully for the analysis of text divisions can be explained if one considers that both unity and division are but two sides of the same coin. All texts are about difference and sameness (Hoey, personal communication, 1996); difference surfaces as segments, while sameness is made evident by the existence of bonding.

The relationship between lexical cohesion and segmentation indicated by the results of the present investigation suggests a possible alternative to a particular problem facing genre analysis as identified by Paltridge (1994). According to him, most work in genre analysis 'draws essentially on categories based on *content* to determine textual boundaries, rather than on the way in which the content is expressed *linguistically*' (Paltridge, 1994, p.295). The investigation presented here suggests that it is possible to determine textual boundaries using linguistic information rather than content information.

The present study has also emphasized the importance of repetition in text. Lexical cohesion was identified on the computer by pattern-matching, that is, by comparing strings of characters without the aid of a database (thesaurus, dictionary, etc). In this manner, the majority of lexical cohesive links detected by the segmentation analysis were simple repetition (Hoey, 1991b, pp. 52-55) (e.g. 'country'⇒'country'), the kind which the computer can correctly identify. Hoey (1991b) had already shown how repetition creates text unity. Winter (1974) had called attention to the 'meaning sharing' role of repetition. Repetition also formed the basis of Pêcheux's (1969/1995) approach to discourse analysis, and Winburne's (1962) analysis of sentence attachment. All of these studies had already capitalized on repetition as a major device for showing aspects of text organisation. However, none had

shown how repetition could be drawn on for segmentation. On the computer side of the literature, though, repetition had already been used for segmentation purposes (Hearst, 1994a; Youmans, 1991), but there was no reflection on the linguistic theory underpinning it.

The claims made by Halliday and Hasan (1976) that cohesion may indicate 'transitions' within texts such as between 'different stages in a complex transaction, or between narration and description in a passage of prose fiction' (p. 295) appear to have been substantiated by the present investigation. Although the present study did not look exactly at the issues Halliday and Hasan (1976) mention, it did indicate that segment boundaries match section boundaries to a considerable degree; since section boundaries indicate shifts introduced by the writer (Goutsos, 1996a), segments may be seen to signal a certain kind of 'transition', and as such the findings reported in this thesis seem to indicate that Halliday and Hasan's predictions are borne out. Halliday and Hasan (1976) speak further of 'discontinuities' (p.295) as another phenomenon which correlates with cohesion. In similar vein, Petöfi and Sözer (1987, p.453) describe the lack of 'continuity' in text revealed by the presence of islands in the constitution of the text. In so far as a new segment can be interpreted as a break in the continuation of the previous segment, the present research provides some evidence for the relationship between cohesion and discontinuity suggested by Halliday and Hasan (1976).

Discontinuity was specifically looked at by Goutsos (1996a), who sees the indication of discontinuity as one of the major tasks writers must manage. Goutsos (1996a) observes that discontinuity is made apparent at segment boundaries: 'the writer is faced with the tasks to manage the interaction through discourse in sequential terms and to segment discourse into chunks and indicate their boundaries, i.e. the discontinuity between one another.' (p .504). Goutsos (1996a) argues that one of the ways whereby discontinuity

is made apparent to the reader is by typographical conventions such as headings, like those found at section boundaries. The fact that segments match section boundaries to an appreciable extent indicates that LSM segments seem to relate to the 'discontinuities' introduced by writers to manage the text flow which Goutsos discusses.

8.4 Sections

The results presented in this investigation also have implications for the status of sections as linguistic units. Judging by the lack of interest in sections in the linguistic literature (cf. Berber Sardinha, 1995a, 1996b), the natural conclusion would be that sections are not important to the way texts are organised. Yet a quick glance at a number of different text types would indicate that many of them have section divisions, and that therefore sections must serve a purpose. A research article, for instance, which did not have a single demarcated section would be a very odd exemplar; readers normally expect sections such as 'Introduction' or 'Methodology' to appear in articles. In a sense, then, sections are constitutive of certain genres, and are not superfluous as the lack of interest in the linguistic literature would lead one to suppose. The present study has indicated that lexical cohesion is related to section boundaries, indicating that sections have a linguistic underpinning. This is important in that it suggests that understanding how sections work is not 'beyond' text linguistics, as it were. As the latter part of the main study has suggested (see section 7.6, pp.375 ff), it seems possible now to account for the appearance of sections in certain parts of texts and not in others by observing certain aspects of the flow of lexical cohesion in text.

In addition to perhaps adding more weight to sections as linguistic units, the match of sections and segments also indicates that segments are in a

sense related to the sequential organisation of texts. Section headings indicate the reason for segmenting the text; sections labelled 'Introduction' and 'Conclusion', for instance, are indicative of what may be called 'sequential organisation', in that the contents of such sections are defined only in relation to the other sections, not in absolute terms. Thus, the *raison d'être* of an 'Introduction' is tautological, namely to 'introduce' elements relevant to the article, such as the goals, major themes, questions, and previous research, among others. The present investigation has indicated that the sequential organisation of texts as revealed by their sections can be investigated by computer.

The relationship between sections and segments also suggests that segments may be seen as a means whereby readers keep track of the unfolding of the contents of the text. Lorch and Lorch (1996) have provided evidence that headings aid the comprehension of written texts. Since segment boundaries coincide with headings to an appreciable extent, the functions of segments may be seen as that of signalling to readers how the text is organised. This would be in agreement with Goutsos (1996a) whose own research found that segments could be interpreted in terms of strategies used by writers in managing the interaction with the reader.

8.4.1 Topicality

The fact that sections and segments have been found to match to an appreciable extent can be taken to be indicative that segments reflect the topicality or 'aboutness' (Collins and Scott, 1996; Phillips, 1985, 1989; Scott, 1997) of texts. Sections are customarily prefaced by headings indicating what they are about, for example, 'Properties' or 'Legal Proceedings' as found in business reports (see figure 7.2 on page 333).This is in accordance with Phillips (1985, p.124) who noted that sections provide a window onto text contents, thus

reflecting the 'aboutness' of the text. As many sections reflect topics, and many segments match sections, there is some evidence that many segments reflect topics. To extend the syllogism, sections have a certain linguistic underpinning as they relate to lexical cohesion, therefore lexical cohesion seems related to the expression of topicality in texts. The implication of this link between topicality and linguistic expression is that the present investigation offers some evidence that the notion of topic, albeit vague and debatable (Georgakopoulou and Goutsos, 1997, p.74), can be approached from a linguistic point of view.

The tentative link between lexical cohesion and topicality serves to substantiate to a certain degree the claims made by van Dijk (1980) about the relationship between topic and discourse unit. In his model, units are defined on the basis that they express a certain topic. By contrast, Brown and Yule (1983, p.73) had suggested that topic, being a pre-theoretical notion, cannot serve as the basis for linguistic analysis of discourse as it does not find expression in linguistic categories. The present investigation suggests that texts, by virtue of the association between segments and sections, appear to be organised in respect to topics, and this organisation is revealed by a linguistic characteristic, namely lexical cohesion. By the same token, this investigation is in agreement with Giora's (1983) claims about the relationship between manual segmentation and topic introduction.

The relationship between topicality and the internal organisation of texts via lexical cohesion had already been suggested by Hoey (1991b, p.91) in terms of the existence of topic opening and topic closing sentences. The former bond mostly with later sentences while the latter do so mostly with earlier sentences (Hoey, 1991b, pp.118-119). Hoey (1991b) found that texts can be successfully abridged by choosing portions bound by topic opening and topic closing sentences. This suggests that the major topical units can

be picked out by relying on the lexical cohesion among sentences.

8.5 Conclusion

The results of the present investigation into segmentation have several implications for research in discourse analysis, segmentation by computer, and lexical cohesion.

The major finding in the present investigation is that texts are segmentable and to research in discourse analysis this suggests that an alternative perspective to model-based text analysis is possible. The use of computers in language analysis has become a reality (Barnbrook, 1996; Butler, 1992a; Hockey and Ide, 1994b,a; Lancashire, 1991; Landow and Delany, 1993; Stubbs, 1996), and discourse analysts can no longer ignore the impact of computer-aided research on the way texts are seen to be organised. The view of text organisation offered by segmentation is one that is model-free, unlike most of discourse analysis research (e.g. Bhatia, 1993; Hasan, 1996a; Longacre, 1983; Sinclair and Coulthard, 1975; Swales, 1990; van Dijk and Kintsch, 1983; van Dijk, 1980, 1983). In this manner, segmentation analysis describes text organisation rather than text structure (Hoey, 1991b) in that no predictive statements are made about which segments may or may not occur. The present investigation also disfavours the traditional macrostructural view of discourse (van Dijk, 1980) since segmentation is not based on a pre-defined hierarchical set of text constituents. It was argued that discourse models and computer segmentations can complement each other, rather than simply oppose each other. Segmentation can show where the major natural divisions of the texts are, while through discourse analysis one can provide an interpretation of what the segments mean *vis-à-vis* a theory of discourse. Future research can explore the possible relationship between segmentation

and existing models of discourse organisation.

The fact that the computer can segment texts raises the question of what status this segmentation has. Firstly, those who advocate that text is hierarchically structured (e.g. Mann and Thompson, 1986b; van Dijk, 1980) would probably expect the segmentation to represent in two-dimensional shadow format the major divisions of the hierarchy. As such, the segmentation would have the status of being part of the hierarchical organisation of the text. Secondly, those who hold the view that text is staged (e.g. Goutsos, 1996a; Hasan, 1984) might expect the segments to roughly correspond with the boundaries of the stages. Finally, those who see text as a web of relationships (Hoey and Winter, 1986; Hoey, 1983) would consider the segmentation to have the status of provisional pause points which are amongst a cluster of clues which the reader makes use of in order to interpret the web. Unlike in the hierarchical and stage views, these points would not be absolute division points, though. The graphic representations of discourse organisation supported by these different views are illustrated in figure 8.1.

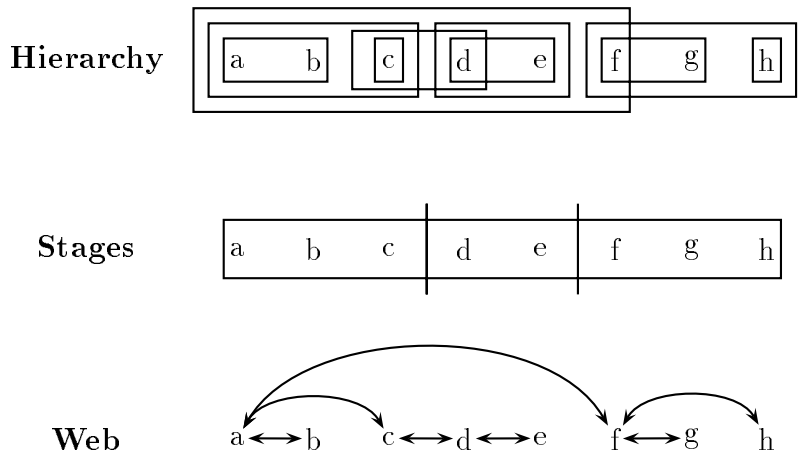


Figure 8.1: Representations of discourse

The discourse view which more closely corresponds with the segmentational division of discourse is the staged position. Segments resemble stages in that they are contiguous and sequentially arranged. By contrast, segments are radically different from hierarchical constituents since segments do not subdvide, overlap, or subsume others. Segments are also unlike the nodes of a web of relationships given that segments do not presuppose multiple connections; rather, the basic relation among segments operates between adjacent segments, in that a segment boundary indicates a transition point or a 'discontinuity' (Goutsos, 1996a; see previous discussion in section 2.4.4 on page 64) between the current segment and its immediate neighbours.

There is an important difference, though, between segments and stages in that, unlike segments, there are constraints which regulate the order of realization of stages. For example, the Generic Structure Potential (GSP) for a service encounter prescribes that a 'greeting' cannot follow a 'sales enquiry', and that the 'purchase' has to precede the 'purchase closure' (Hasan, 1989, p.64). Another difference is that in a staged view there may be obligatory elements. In a service encounter, the obligatory elements are 'sale request', 'sale compliance', 'sale', and 'purchase' (Hasan, 1989, p.60), and it is the presence of these elements which characterises the genre known as 'service encounter'. Neither of these constraints are built into the segmentational description of discourse.

The current investigation also adds to the research in computational segmentation. The performance of the procedure developed for segmenting the texts (LSM) was lower than another existing algorithm (TextTile). However, the interest of the present investigation does not lie in competing with other procedures, not least because the performance of LSM could be improved mechanically by fiddling with parameters. The results of the current investigation suggest that segmentation can be adequately carried out without having

recourse to thesauri or electronic dictionaries, unlike previous research (Kozima and Furugori, 1993; Kozima, 1993a,b; Morris and Hirst, 1991; Morris, 1988; Okumura and Honda, 1994). In addition, LSM suggests that research in manual text analysis can provide the basis for computer-aided segmentation research (e.g. Hoey) without too many modifications. By contrast, certain aspects of previous computational segmentation algorithms have been based on arbitrary rather than research-informed decisions.

Previously, lexical cohesion had been used for segmentation in the form of lexical chains or strings (Morris, 1988; Morris and Hirst, 1991). The current investigation suggests that lexical cohesion described in terms of clusters (Hoey, 1988, 1991b) can also be utilized to indicate textual segments. Particularly, the present investigation stresses the importance of repetition in text (Hoey, 1983, 1991b; Winter, 1974).

Another contribution of the present investigation is that it suggests that sections can be viewed as being linguistically motivated. To date, text research had avoided explaining sections on linguistic grounds, partly because of the tendency of discourse analysis to handle small amounts of data (Phillips, 1985, 1989). The current investigation indicates that sections have linguistic realization which is reflected in the lexical cohesion of the text. And as segments matched sections to an appreciable extent, the final major finding in the present investigation is that segments may be seen to reflect the topicality or 'aboutness' (Collins and Scott, 1996; Phillips, 1985, 1989; Scott, 1997) of texts, just as sections do.

Chapter 9

Conclusion

In this chapter a summary of the main findings of the study presented in this thesis will be provided. The main findings will be discussed and checked against the aims declared in the Introduction. Finally, a few suggestions for further research are made.

9.1 Summary

In chapter 1, the argument was presented that although several disciplines make use of computers for the analysis of large quantities of linguistic data, little interest has been devoted to the large-scale analysis of individual texts from a discourse analysis perspective. In a discourse analysis perspective, the text is the basic unit of analysis (Georgakopoulou and Goutsos, 1997; Scott, 1997), and questions relating to text organization are central. It was also argued that the kind of computer-based analysis which could address text organization was *segmentation*, or the division of texts into discrete units. The major aim of the investigation was also declared, namely the development of a computer-assisted segmentation procedure.

In chapter 2, it was argued that although segmentation is a common

task in discourse analysis, existing approaches provided only a restricted framework for segmentation. An alternative framework was proposed which would allow for extensive coverage, inductive data treatment, and independent validation. These desiderata could only therefore be achieved by using computers.

In chapter 3, a description was provided of the major approaches to segmentation by computer. It was remarked that lexical cohesion had been commonly used in segmentation tasks, in addition to having an intuitive appeal for characterizing segments. The most common operationalisation of lexical cohesion was through lexical chains, but they are problematic in that they need databases in order to be reliably identified.

In chapter 4, a review of major approaches to lexical cohesion was presented which indicated that the proposal by Hoey (1991b), based on the clustering of lexical cohesion among sentences, provided a viable alternative for the segmentation task. A key insight in Hoey's approach is that of the central role of repetition in creating cohesion between sentences. Because of this stress on repetition, his approach can be implemented on the computer.

Chapter 5 included three studies aimed at developing the final methodology for segmenting texts by computer. Generally, the pilot studies were concerned with developing ways for the computation of lexical cohesion, placement of segment boundaries, and evaluation of the performance of the segmentation. Before attempting to develop a computer-based procedure, it was decided that manual analyses should be tried out and evaluated.

The first pilot study (see section 5.2, p.191 ff.) concerned itself with the development of ways for computing lexical cohesion by computer, carrying out a manual segmentation of the text by examining the lexical base, and checking to what extent the segments matched the existing section divisions. A segmentation procedure was implemented which drew an 'exclusion line'

in a lexical cohesion matrix and then looked for segments within a narrow strip within the matrix. The segmentation of one single text was provided in this way. It was felt that several aspects needed improving, and therefore a second pilot study was conducted.

The second pilot study (see section 5.3, p.206 ff.) concentrated on developing another procedure for segmentation and trying it out by hand. Specifically, the pilot study took upon itself the tasks of developing a new procedure for segmenting a matrix of lexical cohesive links, and trying out new schemes for measuring performance of segmentation. A procedure was devised which searched a lexical cohesive matrix for triangle shaped clusters which would lead to segments in the text. The analysis of one text was provided in this manner. The pilot study concluded that a full automatic segmentation should be pursued.

The third pilot study (see section 5.4, p.218 ff.) was the first which experimented with a computer-assisted procedure for segmentation. The goal of the study was to develop a segmentation procedure which could place segment boundaries without human assistance. A statistical procedure known as 'cluster analysis' was used for the segmentation. A corpus of 25 texts was segmented. Although in principle cluster analysis appeared suitable for segmentation, the actual implementation resulted in low performance. This suggested that a new procedure should be developed, a task which required yet another study. In the third pilot, though, the goal of computing lexical cohesion automatically was achieved through the development of a specific computer program.

Chapter 6 reported the development of another computer-assisted segmentation procedure. It was based on the comparison of the lexical cohesive profile of each sentence in the texts. The procedure was labelled 'Link Set Median', or 'LSM' because it worked by computing differences between the

median sentence in each sentence's link set. The same 25-text corpus as used for pilot study 3 was segmented. A better performance was achieved with LSM than with cluster analysis. Moreover, in this study other goals were attained, namely the automatic computation of cohesion, the automatic placement of boundaries, and the ability to handle several texts. Significantly, LSM permits extensive coverage, inductive orientation, and objective evaluation, that is, the major characteristics which a segmentation procedure should have. Therefore, LSM was settled on as the best alternative for segmenting texts in a larger-scale study.

Chapter 7 reported the results of the application of LSM to a corpus of 300 texts, distributed in three separate corpora of 100 texts each, and representing three distinct text types: research articles, business reports, and encyclopedia articles. The goals of this stage of the study were to find out whether LSM segmentation performed better than random, and if performance was affected by link levels and by text type. The first most important set of results were that the main effect of segmentation was statistically significant in all three corpora for both recall and precision, with LSM yielding higher performance rates than random segmentation. This suggested that LSM is a principled method for segmenting texts which builds on lexical cohesion across sentences. Second, performance seemed to be affected by link levels, as higher levels tended to make LSM and random performances not statistically different; LSM seemed to perform better at the lowest link levels. Finally, the advantage of LSM over random segmentation does not change in relation to text type.

The first of two analyses tried to explain why the segmentation worked. Several textual characteristics were identified which seemed to influence the performance of the segmentation. The results indicated that statistically these characteristics explained about 31% of the variation in segmentation.

The analysis also presented ways of predicting the segmentation of a text by knowing certain of its characteristics. Hence, part of the explanation of why the segmentation works is that it relates to certain specific textual characteristics.

The second analysis tried to explain why the segmentation worked by looking at the extent to which the segmentation could predict the probability that section boundaries could occur in certain places in the texts more than in others. The results indicated that the majority of sections could be predicted probabilistically, as section breaks could be identified by a specific pattern in the flow of medians across the text. This was interpreted as evidence that the segmentation worked because it tapped into a major characteristic of the texts, namely the flow of lexical cohesion which underlies the choice of sentences to be section initiators.

9.2 Contributions

The main contribution of the present study is that texts are not only segmentable by hand, but also by computer. Several studies have proposed models of discourse which allegedly work, but fewer are replicable and applicable to large texts, and even fewer can be used over many texts. A second contribution is that the segmentation need not be based on arbitrary criteria; rather it can draw on principles derived from discourse considerations. A third contribution is that the present study shows the importance of lexical cohesion in general, and repetition in particular, in segmenting and hence in organising texts. A fourth contribution is that sections do not appear to be arbitrarily imposed on texts, rather they seem to relate to the lexical cohesion of the text. Finally, as sections relate to the organisation of the subject-matter of texts, another contribution of the present study is that segments seem to be

related to the topicality or 'aboutness' of texts.

9.3 Attainment of aims

As stated in the introduction (p.16), the major aim of the study reported in this thesis was to develop a computer-assisted segmentation procedure whose fundamental characteristic should be that it would borrow insights from research in discourse analysis and text linguistics, so that it could make a contribution to these fields. A major characteristic of the analysis was that it was carried out a large number of texts and each text was analysed independently of the others.

The study reported in this thesis has managed to develop a computational procedure for segmenting texts following insights from discourse analysis, and therefore the main aim of the thesis has been achieved. The three specific aims were also attained:

1. A range of discourse characteristics was considered, and one particular discourse feature, namely lexical cohesion, was chosen to serve as the basis for the segmentation procedure.

2. A variety of segmentation techniques was experimented with, both manual and automatic.

3. Specialized computer software was developed exclusively to help in the computation of lexical cohesion; another set of routines was written for carrying out the segmentation itself.

9.4 Further research

There are several pieces of research which could be conducted to answer questions that were not answered in the investigations reported in this thesis.

Firstly, and most obviously, although LSM achieved a good result, it is not yet a 100% result. It is entirely legitimate to doubt that a 100% result is ever to be expected or even desirable, given that at least some sections that were inserted by some writers were not as wise as others. There are still grounds for suspecting that a better hit rate could be achieved. One of the aspects which could be looked at is whether there are other linguistic features in the environment that might be used to identify segments in addition to lexical repetition. By 'in addition to' is meant either other forms of repetition or other linguistic features altogether. To take a simple example, if there is a choice between two sentences as to which one would be a better segment boundary, one of which starts with a pronoun and the other does not, one would presumably favour the sentence that does not begin with a pronoun over the one that does. As mentioned previously on p.124, this is part of the segmentation strategy employed by Hahn and Strube (1997). In short, it would be important to try to supplement the linguistic criteria currently employed by LSM with other linguistic criteria.

Secondly, as mentioned above (p.312), the negative performance of pilot study 3 is an interesting result. To actually have a segmentation procedure that regularly and systematically performed worse than random means that the procedure was identifying the opposite phenomenon from segmentation. Several possibilities connected to the way the cluster analysis was conducted were explored above which might explain why pilot study 3 did not work. Possible sources of error include the coding of the data, the inability of k-means clustering to group cases sensibly, and perhaps the inadequacy of the CCC statistic. Another interesting possibility would be that it was not that the statistics were 'wrong', but that they were picking up something different from segmentation. Subsequent research could look at these issues.

Thirdly, there are other genres to investigate. Narratives, in particular,

may be seen as a challenge for LSM segmentation, since they are said to operate in a significantly different way as far as lexical cohesion is concerned (Hoey, 1994). It would be worthwhile, for example, to use LSM to find chapter divisions in narrative. If LSM is found inadequate as a segmentation algorithm for narratives, there would be a case for arguing that another system should be developed to account for the segmentation that is done routinely and manually by writers of narratives.

Fourthly, since TextTiling and LSM identified a large proportion of different boundaries, another piece of research would be to try and construct a computer program that allowed for identification of the kinds of boundaries picked up by both TextTiling and LSM. In designing this new segmentation algorithm, consideration would have to be given to whether it would be legitimate to tweak the system to take account of existing boundaries such as paragraphs. At the moment, there is a strong case against such tweaking, but one would have to consider the potential uses of the systems; if the system were to be designed for research purposes (as LSM was), then it would not be legitimate to use such an adjustment, but if the system were meant as an information retrieval tool (as TextTiling was), then tweaking the system to achieve maximum performance would be a legitimate strategy. Of course, including paragraph boundaries does not in itself constitute tweaking, since it would be perfectly legitimate to use paragraph boundaries as one of the factors which raise the odds on a possible segment boundary being located there without restricting the segment boundaries to paragraph gaps.

Finally, with respect to comparisons with other procedures, TextTiling achieves a better result at the moment because of how it adjusts the segment boundaries to coincide with paragraph gaps. Thus, another piece of research could involve building in a similar tweak into LSM to see whether it improves on TextTiling. By making this change, there would be an absolute measure

of whether one system is more effective or equally effective or less effective than the other. At the moment this is not possible to assert because the comparison was between a tweaked system and an untweaked system, given that it was not possible to conveniently untweak TextTiling (see p.296).

9.5 Final comments

In this thesis a series of studies have been presented which were developed over four years of study. The research reported here is innovative and has implications for central areas of language analysis, especially with respect to applications of computers to the large-scale study of text organisation.

Bibliography

Aarts, J. and Meijs, W. (eds.), 1990. *Theory and Practice in Corpus Linguistics*. Rodopi, Amsterdam/Atlanta, Ga.

Aijmer, K. and Altenberg, B. (eds.), 1991. *English Corpus Linguistics - Studies in honour of Jan Svartvik*. Longman, London.

Alderfelder, M. S. and Blashfield, R. K., 1984. *Cluster Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage, Beverly Hills, CA.

American Heritage Dictionary, 1994. Third edition, Laurel, New York.

Atwell, E., 1986. Beyond the micro: Advanced software for research and teaching from computer science and artificial intelligence. In G. Leech and C. Candlin (eds.), *Computers in English Language Teaching and Research*. Longman, London, pp. 168–184.

Bales, R. F. and Strodtbeck, F., 1968. Phases in group problem solving. In D. Cartwright and A. Zander (eds.), *Group Dynamics*, third edition. Row,Peterson and Company, Evanston, pp. 624–638.

Barnbrook, G., 1996. *Language and Computers - A Practical Introduction to the Computer Analysis of Language*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press, Edinburgh.

Barthes, R., 1977. *Image, Music, Text*. London, Fontana.

Becker, A. L., 1965. 'A tagmemic approach to paragraph analysis'. *College Composition and Communication*, 16: 237–242.

Beeferman, D., Berger, A., and Lafferty, J., 1997. Text segmentation using exponential models. Unpublished manuscript, School of Computer Science, Carnegie Mellon University, Available from `http://xxx.lanl.gov/cmp-lg/`.

Benbrahim, M., 1996. Automatic text summarisation through lexical cohesion analysis. Unpublished Ph.D. thesis, Artificial Intelligence Group, Department of Mathematical and Computing Sciences, University of Surrey, Guilford.

Benbrahim, M. and Ahmad, K., 1994. Computer-aided lexical cohesion analysis and text abridgment. Technical Report Knowledge Processing 18, Computing Sciences Report CS-94-11, University of Surrey.

Benson, J. D. and Greaves, W. S., 1992. Collocation and field of discourse. In W. C. Mann and S. A. Thompson (eds.), *Discourse Description – Diverse Linguistic Analyses of a Fund-Raising Text*. John Benjamins, Amsterdam.

Berber Sardinha, A. P., 1991. 'A move analysis of an engineering project updating meeting'. *The ESPecialist*, 12: 1–18.

Berber Sardinha, A. P., 1993a. Lexis in annual reports: Paragraph linkage and cohesion distance. *DIRECT Papers*. Working Paper 5. CEPRIL, PUC-SP, Brazil, and AELSU, Liverpool University, England.

Berber Sardinha, A. P., 1993b. Lexis in annual reports: The cluster triangle technique. *DIRECT Papers*. Working Paper 2. CEPRIL, PUC-SP, Brazil, and AELSU, Liverpool University, England.

Berber Sardinha, A. P., 1995a. Annual business reports sections: key words. *DIRECT Papers*. Working Paper 25. CEPRIL, PUC-SP, Brazil, and AELSU, Liverpool University, England.

Berber Sardinha, A. P., 1995b. 'Collocations in a business text'. *Letras & Letras*, 11: 121–138.

Berber Sardinha, A. P., 1995c. 'Corpus choices in a short journalistic text'. *The ESPecialist*, 16: 1–20.

Berber Sardinha, A. P., 1995d. Intertextual lexical cohesion in newspaper reports. Paper presented at the 40th Annual Conference of the International Linguistic Association, Georgetown University, Washington, DC, USA, March 10, 1995.

Berber Sardinha, A. P., 1995e. A preliminary study into patterns of lexis of business texts. In B. Warvik, S.-K. Tanskanen, and R. Hiltunen (eds.), *Organization in Discourse. Proceedings from the Turku Conference*, Anglicana Turkuensia. Abo Akademi/University of Turku, Turku, pp. 157–166.

Berber Sardinha, A. P., 1996a. Business text organization: Segmentation and field choice. Paper presented at the American Association for Applied Linguistics Conference, 26 March 1996, Chicago, Ill, USA.

Berber Sardinha, A. P., 1996b. Sections as linguistic units: Key words. Paper presented at the 5th Annual Postgraduate Conference, 9 March 1996, University of Manchester, Manchester, UK.

Berry, M., 1989. Thematic options and success in writing. In C. S. Butler, R. A. Cardwell, and J. Channell (eds.), *Language and Literature - Theory and Practice: A Tribute to Walter Grauberg*. Nottingham University, Nottingham, pp. 62–80.

Bestgen, Y. and Costermans, J., 1997. Temporal markers of narrative structure: Studies in production. In J. Costermans and M. Fayol (eds.), *Processing Interclausal Relationships – Studies in the Production and Comprehension of Text*. Lawrence Earlbaum, Mahwah, NJ, pp. 201–217.

Bestgen, Y. and Vonk, W., 1995. 'The role of temporal segmentation markers in discourse processing'. *Discourse Processes*, 19: 385–406.

Bhatia, V. K., 1993. *Analysing Genre: Language Use in Professional Settings*. Longman, London.

Biber, D., 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Biber, D., 1993. 'Representativeness in corpus design'. *Literary and Linguistic Computing*, 8: 243–257.

Biber, D., 1995a. *Dimensions of Register Variation - A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Biber, D., 1995b. 'On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson'. *Text*, 15: 341–370.

Biber, D. and Finegan, E., 1988. 'Adverbial stance types in English'. *Discourse Processes*, 11: 1–34.

Biber, D. and Finegan, E., 1994. Intra-textual variation within medical research articles. In N. Oostdijk and P. de Haan (eds.), *Corpus-Based Research into Language*. Rodopi, Amsterdam, pp. 201–222.

Binsted, K., 1994. A symbolic description of punning riddles and its computer implementation. Unpublished manuscript, Available from `http://xxx.lanl.gov/cmp-lg/`, file 9406021.

Brown, G. and Yule, G., 1983. *Discourse Analysis*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Burke, P. J., 1991. Segmentation and control of a dissertation defense. In A. Grimshaw (ed.), *What's going on Here? Complementary Studies of Professional Talk*, Advances in Discourse Processes. Ablex, Norwood, NJ, pp. 95–123.

Butler, C. S. (ed.), 1992a. *Computers and Written Texts*. Applied Language Studies. Blackwell, Oxford.

Butler, C. S., 1992b. Editorial introduction. In C. S. Butler (ed.), *Computers and Written Texts*. Blackwell, Oxford, pp. vi–xii.

Cahn, J., 1996. An investigation into the correlation of cue phrases, unfilled pauses and the structuring of spoken discourse. Unpublished manuscript, Media Laboratory, Massachussets Institute of Technology, Cambridge, MA, USA.

Chen, K., 1995. Topic identification in discourse. Unpublished manuscript, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

Christensen, F., 1965. 'A generative rhetoric of the paragraph'. *College Composition and Communication*, 16: 144–156.

Church, K. W., 1993. Charalign: A program for aligning parallel texts at the character level. Paper presented at the 31st Annual Meeting of the Association for Computational Linguistics, 1993.

Clements, P., 1979. The effects of staging on recall from prose. In R. O. Freedle (ed.), *New Directions in Discourse Processing*. Ablex, Norwood, NJ, pp. 287–330.

Cloran, C., 1994. *Rhetorical Units and Decontextualization: An Enquiry into some Relations of Context, Meaning, and Grammar*. Monographs in Systemic Linguistics. Department of English Studies, University of Nottingham, Nottingham.

Cloran, C., 1995. Defining and relating text segments - Subject and theme in discourse. In R. Hasan and P. Fries (eds.), *On Subject and Theme: A Discourse Functional Perspective*. John Benjamins, Amsterdam, pp. 362–403.

Collier, A., 1994. 'A system for automating concordance line selection'. *Proceedings of NeMLaP*: 95–100.

Collins, H. and Scott, M., 1996. Lexical landscaping. *DIRECT Papers*. Working Paper 32. CEPRIL, PUC-SP, Brazil, and AELSU, Liverpool University, England.

Connor, U., 1996. *Contrastive Rhetoric - Cross-cultural Aspects of Second Language Writing*. Cambridge University Press, Cambridge.

Coulthard, M. and Brazil, D., 1992. Exchange structure. In M. Coulthard (ed.), *Advances in Spoken Discourse Analysis*. Routledge, London, pp. 50–78.

Crothers, E. J., 1979. *Paragraph Structure Inference*. Ablex, Norwood, NJ.

Crystal, D., 1991. *A Dictionary of Linguistics and Phonetics*. Third edition. Blackwell, London.

Daneš, F., 1974. Functional sentence perspective and the organization of the text. In F. Daneš (ed.), *Papers on Functional Sentence Perspective*. Academia/Mouton, Prague/The Hague, pp. 106–128.

Darnton, A., 1987. Episodes in the development of narrative awareness in children. M.litt. thesis, Department of English, University of Birmingham, UK.

Davies, F., 1994. From writer roles to elements of text: Interactive, organisational and topical. In L. Barbara and M. Scott (eds.), *Reflections on Language Learning – In honour of Antonieta Celani*. Multilingual Matters, Clevedon, pp. 170–183.

Davies, R., 1985. 'Q-Analysis: A methodology for librarianship and information science'. *Journal of Documentation*, 41: 221–246.

de Beaugrande, R., 1997. The story of Discourse Analysis. In T. A. van Dijk (ed.), *Discourse as Structure and Process*, Discourse Studies – A Multidisciplinary Introduction. Sage, London, pp. 35–62.

de Beaugrande, R.-A. and Dressler, W. U., 1981. *Introduction to Text Linguistics*. Longman, London.

Di Pietro, R. J., 1983. Form vs. function in discourse structures. In R. A. Hall (ed.), *The Ninth LACUS Forum*. Hornbeam, Columbia, SC.

Donaldson, T., Makuta, M., and Cohen, R., 1996. An integrated approach to evaluating text coherence and its application to the prevention of reader misconceptions (or the joy of detecting incoherence in texts). Unpublished manuscript, Department of Computer Science, University of Waterloo, Ontario, Canada.

Eggins, S., 1994. *An introduction to Systemic Functional Linguistics*. Pinter, London.

Ehrich, V. and Koster, C., 1983. 'Discourse organization and sentence form: the structure of room descriptions in Dutch'. *Discourse Processes*, 6: 169–195.

Everitt, B., 1974. *Cluster Analysis*. Social Science Research Council/Heinemann, London.

Firth, J. R., 1957. *Papers in Linguistics - 1934-1951*. Oxford University Press, Oxford.

Fisher, D. E., 1994. Topic characterization of full length texts using direct and indirect term evidence. Technical Report UCB/SCD 94-809, Computer Science Division (EECS), University of California, Berkeley, California, USA.

Frawley, W., 1987. 'Review of 'T A van Dijk (ed.) Handbook of Discourse Analysis". *Language*, 63: 361–397.

Fries, P. H., 1990. Toward a componential approach to text. In M. A. K. Halliday, J. Gibbons, and H. Nicholas (eds.), *Learning, Keeping and Using Language: Selected Papers from the Eighth World Congress of Applied Linguistics, Sydney, 16-21 August 1987*. John Benjamins, Amsterdam,Philadelphia, pp. 363–380.

Fries, P. H., 1995. Patterns of information in initial position in English. In P. H. Fries and M. Gregory (eds.), *Discourse in Society: Systemic Functional Perspectives (Meaning and Choice in Language: Studies for Michael Halliday)*. Ablex, Norwood,NJ, pp. 47–65.

Garcia-Berrio, A. and Mayordomo, T. A., 1987. Compositional structure: Macrostructure. In J. S. Petöfi (ed.), *Text and Discourse Constitution - Empirical Aspects, Theoretical Approaches*. De Gruyter, Berlin, pp. 170–213.

Georgakopoulou, A. and Goutsos, D., 1997. *Discourse Analysis - An Introduction*. Edinburgh University Press, Edinburgh.

Giora, R., 1983. 'Segmentation and segment cohesion: On the thematic organization of the text'. *Text*, 3: 155–181.

Girden, E. R., 1992. *ANOVA: Repeated Measures*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage, Newbury Park, CA.

Glass, A. L., 1983. 'The comprehension of idioms'. *Journal of Psycholinguistic Research*, 12 (4): 429–442.

Gledhill, C., 1995. 'Collocation and genre analysis – The phraseology of grammatical items in cancer research abstracts and articles'. *ZAA (Zeitschrift für Anglistik und Amerikanistik)*, 1: 11–36.

Good, I. J., 1977. The botryology of botryology. In J. van Ryzin (ed.), *Classification and Clustering*. Academic Press, New York.

Goutsos, D., 1996a. 'A model of sequential relations in expository text'. *Text*, 16: 501–533.

Goutsos, D., 1996b. *Modeling Discourse Topic: Sequential Relations and Strategies in Expository Text*. Ablex, New York.

Graustein, G. and Thiele, W., 1983. 'English monologues as complex entities'. *Linguistische Arbeitsberichte*, 41: 1–26.

Grefenstette, G. and Tapainen, P., 1994. What is a word, what is a sentence? Problems of tokenization. Unpublished manuscript, Rank Xerox Research Centre.

Gregory, M., 1985a. Discourse as the instantiation of message exchange. In J. Hall, Robert A. (ed.), *The Eleventh LACUS Forum*. Hornbeam Press, Columbia, SC, pp. 243–254.

Gregory, M., 1985b. Towards 'communication' linguistics: A framework. In J. D. Benson and W. S. Greaves (eds.), *Systemic Perspectives on Discourse*, volume 1 - Selected theoretical papers from the Ninth International Systemic Workshop. Ablex, Norwood,NJ, pp. 119–134.

Grimes, J. E., 1975. *The Thread of Discourse*. Janua Linguarum Series Minor. Mouton, The Hague.

Grimshaw, A. D. (ed.), 1991. *What's Going on Here? Complementary Studies of Professional Talk*. Ablex, Norwood,NJ.

Grosz, B., 1995. Discourse and dialogue. In R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue (eds.), *Survey of the State of the Art in Human Language Technology*. Joint publication by the National Science Foundation, Directorate XII-E of the Commission of the European Communities, and Center for Spoken Language Understanding, Oregon Graduate Institute, USA, Washington, DC, pp. 227–229.

Grosz, B. J., Pollack, M. E., and Sidner, C. L., 1989. Discourse. In M. I. Posner (ed.), *Foundations of cognitive science*. MIT Press, Cambridge,MA, pp. 437–468.

Grosz, B. J. and Sidner, C. L., 1986. 'Attention, intentions, the structure of discourse'. *Computational Linguistics*, 12: 175–204.

Guide to Corporate filings, 1997. Securities and Exchange Commission, Publications Unit, Washington, DC, USA.

Haas, S. W. and Losee, Robert M., J., 1994. 'Looking in text windows: Their size and composition'. *Information Processing and Management*, 30 (5): 619–624.

Hahn, U. and Strube, M., 1997. Centering in-the-large: Computing referential discourse segments. Unpublished manuscript, Computational Linguistics Research Group, Freiburg University, Germany.

Hak, T. and Helsloot, N. (eds.), 1995. *Michel Pêcheux. Automatic Discourse Analysis*. Rodopi, Amsterdam/Atlanta,GA.

Halliday, M. A. K., 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold, London.

Halliday, M. A. K., 1985. *An Introduction to Functional Grammar*. Arnold, London.

Halliday, M. A. K., 1994. *An Introduction to Functional Grammar*. Second edition. Edward Arnold, London.

Halliday, M. A. K. and Hasan, R., 1976. *Cohesion in English*. Longman, London.

Harris, M. D., 1989. 'Analysis of the discourse structure of lyric poetry'. *Computers and the Humanities*, 23: 423–428.

Harris, Z., 1951. *Methods in Structural Linguistics*. Chicago University Press, Chicago.

Hasan, R., 1977. Text in the systemic functional model. In W. Dressler (ed.), *Current Trends in Text Linguistics*. De Gruyter, Berlin, pp. 228–246.

Hasan, R., 1984. Coherence and cohesive harmony. In J. Flood (ed.), *Understanding Reading Comprehension: Cognition, Language and the Structure of Prose*. International Reading Association, Newark, Delaware, pp. 181–219.

Hasan, R., 1989. Part B. In M. A. K. Halliday and R. Hasan (eds.), *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*, second edition. Oxford University Press, Oxford, pp. 52–118.

Hasan, R., 1996a. The nursery tale as genre. In C. Cloran, D. Butt, and G. Williams (eds.), *Ways of Saying, Ways of Meaning: Selected Papers for Ruqaiya Hasan*. Cassell, London, pp. 51–71.

Hasan, R., 1996b. Semantic networks: a tool for the analysis of meaning. In C. Cloran, D. Butt, and G. Williams (eds.), *Ways of Saying: Ways of Meaning - Selected Papers for Ruqaiya Hasan*. Cassell, London, pp. 104–131.

Hearst, M., 1993. Texttiling: A quantitative approach to discourse segmentation. Technical Report Project Sequoia technical Report 24/93, University of California at Berkeley. Available via ftp from `cs-tr.cs.berkeley.edu`.

Hearst, M. A., 1994a. Context and structure in automated full-text information access. Unpublished Ph.D. thesis, Computer Science Division, Berkeley: University of California.

Hearst, M. A., 1994b. Multi-paragraph segmentation of expository texts. Technical Report Project Sequoia Technical Report 94/790, University of California at Berkeley. Available via ftp from `cs-tr.cs.berkeley.edu`.

Hearst, M. A. and Plaunt, C., 1993. Subtopic structuring for full-length document access. Technical Report Project Sequoia Technical Report 93/26, University of California at Berkeley. Available via ftp from `cs-tr.cs.berkeley.edu`.

Hinds, J., 1979. Organizational patterns in discourse. In T. Givon and J. Sadock (eds.), *Discourse and Syntax*, Syntax and Semantics. Ablex, New York, pp. 135–157.

Hirschberg, J. and Litman, D., 1993. 'Empirical studies on the disambiguation of cue phrases'. *Computational Linguistics*, 19.

Hockey, S. and Ide, N. (eds.), 1994a. *Research in Humanities Computing 2 - Selected Papers from the ALLC-ACH Conference, Siegen, June 1990*. Clarendon Press, Oxford.

Hockey, S. and Ide, N. (eds.), 1994b. *Research in Humanities Computing 3 - Selected papers from the ALLC/ACH Conference, Tempe, Arizona, March 1991*. Clarendon Press, Oxford.

Hoey, M., 1983. *On the Surface of Discourse*. George Allen & Unwin, London.

Hoey, M., 1985. 'The paragraph boundary as a marker of relation between the parts of a discourse'. *MALS Journal*, 10: 96–107.

Hoey, M., 1986. The discourse colony: A preliminary study of a neglected discourse type. In M. Coulthard (ed.), *Talking about Text - Studies Presented to David Brazil on his Retirement*. ELR/University of Birmingham, Birmingham, pp. 1–26.

Hoey, M., 1988. 'The clustering of lexical cohesion in non-narrative text'. *Trondheim Papers in Applied Linguistics*, 4: 154–180.

Hoey, M., 1991a. Another perspective on coherence and cohesive harmony. In E. Ventola (ed.), *Functional and Systemic Linguistics*. Mouton de Gruyter, Berlin, pp. 385–414.

Hoey, M., 1991b. *Patterns of Lexis in Text*. Oxford University Press, Oxford.

Hoey, M., 1993. Introduction. In M. Hoey (ed.), *Data, Description, Discourse – Papers on the English Language in Honour of John McH Sinclair on his Sixtieth Birthday*. HarperCollins, London, pp. v–ix.

Hoey, M., 1994. Patterns of lexis in narrative: A preliminary study. In S.-K. Tanskannen and B. Warvik (eds.), *Topics and Comments - Papers from the Discourse Project*. University of Turku, Turku, Finland, pp. 1–40.

Hoey, M., 1995a. The inseparability of word, grammar, and text. Inaugural lecture, 11 December 1995, University of Liverpool, UK.

Hoey, M., 1995b. The lexical nature of intertextuality: A preliminary study. In B. Warvik, S.-K. Tanskanen, and R. Hiltunen (eds.), *Organization in Discourse. Proceedings from the Turku Conference*. Abo Akademi, Turku, pp. 73–94.

Hoey, M., 1996. The interaction of textual and lexical factors in the identification of paragraph boundaries. ELG Papers, AELSU, University of Liverpool, UK (To appear in *Grammar and Text in Syncrony and Diachrony*).

Hoey, M. and Winter, E., 1986. Clause relations and the writer's communicative task. In B. Couture (ed.), *Functional Approaches to Writing - Research Perspectives*. Ablex, Norwood, NJ, pp. 120–141.

Hoey, M. and Wools, D., 1995. Abridge. Unpublished Software.

Hopkins, A. and Dudley-Evans, T., 1988. 'A genre-based investigation of the discussion sections in articles and dissertations'. *English for Specific Purposes*, 7: 113–120.

Hrebicek, L. and Altmann, G., 1993. Prospects of text linguistics. In L. Hrebicek and G. Altmann (eds.), *Quantitative Text Analysis*. Wissenschaftlicher Verlag Trier, Trier, pp. 1–28.

Hughes, J. and Atwell, E., 1994. The automated evaluation of inferred word classifications. In A. Cohn (ed.), *Proceedings of the 11th European Conference on Artificial Intelligence*. John Wiley & Sons, London, pp. 535–539.

Hughes, J. and Atwell, E., nd. Automatically acquiring and evaluating a classification of words. Unpublished manuscript, Centre for Computer Analysis of Language and Speech, University of Leeds.

Humphrey, T., 1996. Finding discourse boundaries in text – The marriage of two algorithms. Unpublished manuscript, Applied Research, Lexis-Nexis, Reed Elsevier plc.

Hwang, S. J. J., 1989. 'Recursion in the paragraph as a unit of discourse development'. *Discourse Processes*, 12: 461–477.

Hyland, K., 1990. 'A genre description of the argumentative essay'. *RELC Journal*, 21: 66–78.

Jordan, M. R., 1984. *Rhetoric of Everyday English Texts*. George Allen & Unwin, London.

Karlgren, J., Gamback, B., and Samuelsson, C., 1995. Clustering sentences - Making sense of synonymous sentences. Unpublished manuscript, NLP-group, Swedish Institute of Computer Science.

Kirk, J., 1994. Taking a byte at Corpus Linguistics. In L. Flowerdew and A. K. K. Tong (eds.), *Entering Text*. Language Centre, The Hong Kong University of Science and Technology, Hong Kong, pp. 18–49.

Knott, A. and Dale, R., 1993. Using linguistic phenomena to motivate a set of rhetorical relations. Unpublished manuscript, Department of Artificial Intelligence, Human Communication Centre, University of Edinburgh.

Kozima, H., 1993a. Computing lexical cohesion as a tool for text analysis. Unpublished Ph.D. thesis, University of Electro-Communications, Graduate School of Electro-Communications, Tokyo, Japan.

Kozima, H., 1993b. Text segmentation based on similarity between words. Unpublished manuscript, University of Electro-Communications, Tokyo, Japan.

Kozima, H. and Furugori, T., 1993. Segmenting narrative text into coherent scenes. In *Proceedings of EACL-93*. pp. 232–239.

Kukharenko, V., 1979. Some considerations about the properties of texts. In J. S. Petöfi (ed.), *Text vs Sentence*, volume 1. Helmut Buske Verlag, Hamburg, pp. 235–245.

Kyto, M., Ihalainen, O., and Rissanen, M. (eds.), 1988. *Corpus Linguistics Hard and Soft*. Rodopi, Amsterdam.

Labov, W., 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, PA.

Labov, W. and Waletzky, J., 1967. Narrative analysis. In J. Helm (ed.), *Essays on the Verbal and Visual Arts*. University of Washington Press, Seattle, pp. 12–44.

Lamprecht, R. R., 1988. 'Textsegmetierung in gegenstandlich-thematische Linien (Text segmentation on subject-thematic lines)'. *Wissenschaftliche Zeitschrift der Padagogischen Hochschule 'Karl Liebknecht' Potsdam*, 32 (2): 315–323.

Lancashire, I. (ed.), 1991. *Research in Humanities Computing I: Papers from the 1989 ACH-ALLC Conference*. Oxford University Press, Oxford.

Landow, G. P. and Delany, P. (eds.), 1993. *The Digital Word: Text-Based Computing in the Humanities*. Technical Communication and Information Systems Series. The MIT Press, Cambridge, Mass.

Langleben, M., 1979. On the triple opposition of a text to a sentence. In J. S. Petöfi (ed.), *Text vs Sentence*, volume 1. Helmut Buske Verlag, Hamburg, pp. 246–257.

Ledger, G., 1989. *Re-Counting Plato - A Computer Analysis of Plato's Style*. Clarendon Press, Oxford.

Leech, G. and Fligelstone, S., 1992. Computers and corpus analysis. In C. S. Butler (ed.), *Computers and Written Texts*. Blackwell, Oxford, pp. 115–140.

Lewis, R. D., 1996. *When Cultures Collide – Managing Successfully across Cultures*. Nicholas Brealey Publishing, London.

Lewis-Beck, M. S., 1980. *Applied Regression - An Introduction*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage, Beverly Hills.

Lohmann, P., 1988. Connectedness of texts: A bibliographical survey. In J. S. Petöfi (ed.), *Text and Discourse Constitution – Empirical Aspects, Theoretical Approaches*. De Gruyter, Berlin, pp. 478–502.

Longacre, R. E., 1976. *An Anatomy of Speech Notions*. Peter de Ridder Press, Lisse.

Longacre, R. E., 1979. The paragraph as a grammatical unit. In T. Givon (ed.), *Discourse and Syntax*. Ablex, New York, pp. 115–134.

Longacre, R. E., 1983. *The Grammar of Discourse*. Plenum Press, New York.

Longacre, R. E. and Levinsohn, S., 1978. Field analysis of discourse. In W. U. Dressler (ed.), *Current Trends in Textlinguistics*. De Gruyter, Berlin/New York, pp. 103–121.

Lorch, R. F. and Lorch, E. P., 1996. 'Effects of headings on text recall and summarization'. *Contemporary Educational Psychology*, 21: 261–278.

Mann, W. and Thompson, S. A., 1986a. 'Relational propositions in discourse'. *Discourse Processes*, 9: 57–90.

Mann, W. C., Mathiessen, C., and Thompson, S., 1989. Rhetorical Structure Theory and text analysis. Technical Report ISI Report 89, 242, University of Southern California.

Mann, W. C. and Thompson, S., 1987a. Rhetorical Structure Theory: A theory of text organization. Technical Report ISI Reprint Series 87/190, University of Southern California.

Mann, W. C. and Thompson, S. A., 1986b. Rhetorical Structure Theory: Description and construction of text structures. Technical Report ISI Reprint Series ISI/RS-86-174, University of Southern California.

Mann, W. C. and Thompson, S. A., 1987b. Rhetorical Structure Theory: A framework for the analysis of texts. Technical Report ISI Reprint Series 87, 185, University of Southern California.

Mann, W. C. and Thompson, S. A., 1988. 'Rhetorical Structure Theory: Toward a functional theory of text organization'. *Text*, 8: 243–281.

Mann, W. C. and Thompson, S. A. (eds.), 1992. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*. Pragmatics and Beyond Series. John Benjamins, Amsterdam.

Marcu, D., 1997. The automatic derivation of complex disjunctive structures. Paper presented at the 24th International Systemic Functional Congress, July 21-25, 1997, York University, Toronto, Canada.

Markels, R. B., 1983. 'Cohesion paradigms in paragraphs'. *College English*, 45: 450–464.

Marshall, S., 1991. 'A genre-based approach to the teaching of report-writing'. *English for Specific Purposes*, 10: 3–13.

Martin, J. R., 1989. *Factual Writing: Exploring and Challenging Social Reality*. Oxford University Press, Oxford.

Martin, J. R., 1992. *English Text*. John Benjamins, Philadelphia / Amsterdam.

Matthiessen, C. M. I. M., 1988. Representational issues in Systemic Functional Grammar. In J. D. Benson and W. S. Graves (eds.), *Systemic Functional Approaches to Discourse*. Ablex, Norwood, NJ, pp. 136–175.

McEnery, T. and Wilson, A., 1996. *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Meyer, B. J. F. and Rice, E., 1984. The structure of text. In P. D. e. a. Pearson (ed.), *Handbook of Reading Research*. Longman, New York.

Miall, D. S., 1992. 'Estimating changes in collocations of key words across a large text: A case study of Coleridge's notebooks'. *Computers and the Humanities*, 26: 1–12.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J., 1990. 'Introduction to WordNet: An on-line lexical database'. *Journal of Lexicography*, 3: 235–244.

Milligan, G. W. and Cooper, M. C., 1985. 'An examination of procedures for determining the number of clusters in a data set'. *Psychometrika*, 50 (2): 159–179.

Mitchell, T. F., 1957/1975. The language of buying and selling in Cyrenaica: A situational statement. In T. F. Mitchell (ed.), *Principles of Neo-Firthian Linguistics*. Longman, London, pp. 167–200.

Morris, J., 1988. Lexical cohesion, the thesaurus, and the structure of text. Technical Report Technical Report 219, University of Toronto, Toronto.

Morris, J. and Hirst, G., 1991. 'Lexical cohesion computed by thesaural relations as an indicator of the structure of text'. *Computational Linguistics*, 17: 21–48.

Norusis, M., 1990. *SPSS Advanced Statistics User's Guide*. SPSS Inc., Chicago, Ill.

Nwogu, K. N., 1991. 'Structure of science popularizations: a genre-analysis approach to the schema of popularized medical texts'. *English for Specific Purposes*, 10: 111–123.

Okumura, M. and Honda, T., 1994. Word sense disambiguation and text segmentation based on lexical cohesion. Paper presented at COLING 1994.

Ostler, S. E., 1987. English in parallels: A comparison of English and Arabic prose. In U. Connor and R. Kaplan (eds.), *Writing across Languages: Analysis of L2 Text*. Addison-Wesley, Reading,MA, pp. 169–185.

Paduceva, E. V., 1974. 'On the structure of the paragraph'. *Linguistics*, 13: 49–58.

Paltridge, B., 1994. 'Genre analysis and the identification of textual boundaries'. *Applied Linguistics*, 15: 288–299.

Parsons, G., 1990. *Cohesion and Coherence: Scientific Texts (A Comparative Study)*. Monographs in Systemic Linguistics. Department of English, Nottingham.

Parsons, G., 1996. The development of the concept of cohesive harmony. In M. Berry, R. Fawcett, C. Butler, and G. Huang (eds.), *Meaning and Form: Systemic Functional Interpretations (Meaning and Choice in Language: Studies for Michael Halliday)*. Ablex, Norwood,NJ, pp. 585–599.

Passonneau, R. J. and Litman, D., 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. pp. 148–155. Available from `http://xxx.lanl.gov/cmp-lg/`, file 9405015.

Passonneau, R. J. and Litman, D. J., 1995. Empirical analysis of three dimensions of spoken discourse: segmentation, coherence and linguistic devices. Technical report, Department of Computer Science, Columbia University, New York, NY, USA, and AT&T Bell Laboratories, Murray

Hill, NJ, USA (To appear in *Burning Issues in Discourse*, ed. by Hovy, Edward and Scott, Donia).

Pêcheux, M., 1969/1995. Automatic discourse analysis. In T. Hak and N. Helsloot (eds.), *Michel Pêcheux. Automatic Discourse Analysis*. Rodopi, Amsterdam/Atlanta,GA, pp. 63–121.

Petöfi, J. S., 1979. *Text vs Sentence: Basic Questions in Textlinguistics*. Papers in Textlinguistics. Helmut Buske Verlag, Hamburg.

Petöfi, J. S. (ed.), 1982. *Text vs Sentence 2: Basic Questions of Text Linguistics*. Papers in Textlinguistics. Helmut Buske Verlag, Hamburg.

Petöfi, J. S. and Rieser, H. (eds.), 1973. *Studies in Text Grammar*. Reidel, Dordrecht.

Petöfi, J. S. and Sözer, E., 1987. Static and dynamic aspects of text constitution. In J. S. Petöfi (ed.), *Text and Discourse Constitution - Empirical Aspects, Theoretical Approaches*. De Gruyter, Berlin, pp. 440–477.

Phillips, M., 1985. *Aspects of Text Structure - An Investigation of the Lexical Organisation of Text*. North-Holland Linguistic Series. North-Holland, Amsterdam.

Phillips, M., 1989. *Lexical Structure of Text*. Discourse analysis monographs. ELR, University of Birmingham, Birmingham.

Pike, K. and Pike, E. G., 1977. *Grammatical Analysis*. Summer Institute of Linguistics, Arlington, TX.

Pike, K. L., 1972. Grammar as wave. In R. M. Brend (ed.), *Kenneth L Pike - Selected Writings*. Mouton, The Hague, pp. 231–241.

Pike, K. L., 1982. *Linguistic Concepts: An Introduction to Tagmemics*. University of Nebraska Press, Lincoln.

Pitkin, W. L., 1969. 'Discourse blocs'. *College Composition and Communication*, 30: 138–148.

Pollard-Gott, L., McCloskey, M., and Todres, A. K., 1979. 'Subjective story structure'. *Discourse Processes*, 2: 251–281.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Raben, J., 1991. 'Humanities computing 25 years later'. *Computers and the Humanities*, 25: 341–350.

Renouf, A. and Collier, A., 1995. A system of automatic abridgment. Unpublished manuscript, Research and Development Unit for English Studies. University of Liverpool, UK.

Reynar, J. C., 1994. An automatic method of finding topic boundaries. Paper presented at the Association for Computational Linguistics Conference 1994, Student Session, available from `http://xxx.lanl.gov/cmp-lg/`.

Rietveld, T. and Van Hout, R., 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter, Berlin.

Rodgers, P. C., 1966. 'A discourse-centered rhetoric of the paragraph'. *College Composition and Communication*, 17: 2–11.

Rotondo, J., 1984. 'Clustering analysis of subject partitions of text'. *Discourse Processes*, 7: 69–88.

Rumelhart, D. E., 1975. Notes on a schema for stories. In D. G. Bobrow and A. Collins (eds.), *Representation and Understanding - Studies in Cognitive Science*. Academic Press, New York, pp. 211–236.

Sacks, E., Schegloff, E., and Jefferson, G., 1974. 'A simplest systematics for the organization of turn-taking for conversation'. *Language*, 50: 696–735.

Salager-Meyer, F., 1989. 'Principal components analysis and Medical English discourse: an investigation into genre analysis'. *System*, 17: 21–34.

Salager-Meyer, F., 1990. 'Discoursal flaws in medical English abstracts: a genre analysis per research- and text-type'. *Text*, 10: 365–384.

Salton, G., 1988. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

Salton, G. and Buckley, C., 1991. Automatic text structuring and retrieval: Experiments in automatic encyclopedia searching. Paper presented at the 14th Annual International ACM/SIGIR Conference.

Salton, G., Buckley, C., and Zhao, Z., 1990. Text linking and retrieval experiments for textbook components. Unpublished manuscript, Department of Computer Science, Cornell University, Ithaca, NY, USA.

Salton, G., Singhal, A., Buckley, C., and Mitra, M., 1994. Automatic text decomposition using text segments and text themes. Unpublished manuscript, Department of Computer Science, Cornell University.

Sarle, W. S., 1983. Cubic Clustering Criterion. Technical Report SAS Technical Report A-108, SAS Institute Inc, Cary, NC.

SAS Institute Inc, 1989a. *SAS/STAT User's Guide, Version 6*, volume 1. Fourth edition, SAS Institute Inc, Cary, NC.

SAS Institute Inc, 1989b. *SAS/STAT User's Guide, Version 6*, volume 2. Fourth edition, SAS Institute Inc, Cary, NC.

Schegloff, E. A. and Sacks, H., 1973. 'Opening up closings'. *Semiotica*, 7: 289–327.

Schiffrin, D., 1994. *Approaches to Discourse*. Blackwells Textbooks in Linguistics. Blackwell, Oxford.

Schroeder, L. D., Sjoquist, D. L., and Stephan, P. E., 1986. *Understanding Regression Analysis – An Introductory Guide*. Quantitative Applications in the Social Sciences. Sage, Beverly Hills, Cal.

Scinto, L. F. M., 1986. *Written Language and Psychological Development*. Academic Press, Orlando, Fla.

Scott, M., 1996. *WordSmith Tools*. Oxford University Press, Oxford. Computer Software.

Scott, M., 1997. 'PC Analysis of key words - and key key words'. *System*, 25: 233–245.

Sibson, R., 1972. 'Order invariant methods for data analysis'. *Journal of the Royal Statistical Society B (Methodological)*, 34: 311–337.

Siegel, S., 1975. *Estatística Não-Paramétrica*. McGraw Hill, São Paulo.

Sinclair, J., 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Sinclair, J. and Coulthard, M., 1992. Towards an analysis of discourse. In M. Coulthard (ed.), *Advances in Spoken Discourse Analysis*. Routledge, London, pp. 1–34.

Sinclair, J. M., 1966. Beginning the study of lexis. In C. E. Bazell (ed.), *In Memory of J R Firth*. Longman, London, pp. 410–430.

Sinclair, J. M., 1994. Trust the text. In M. Coulthard (ed.), *Advances in Written Text Analysis*. Routledge, London, pp. 12–25.

Sinclair, J. M. and Coulthard, R. M., 1975. *Towards and Analysis of Discourse - The English Used by Teachers and Pupils*. Oxford University Press, Oxford.

Skorochod'ko, E. F., 1972. 'Adaptive method of automatic abstracting and indexing'. *Information Processing*, 71: 1179–1182.

Smadja, F., 1992. Retrieving collocations from text: Xtract. Unpublished manuscript, Columbia University, New York City, USA.

Sparck Jones, K., 1996. How much has information technology contributed to linguistics? Technical report, Computer Laboratory, University of Cambridge, UK. (To appear in *Information Technology and Scholarly Disciplines*, ed. J T Coppock, Proceedings of the British Academy Symposium. London: The British Academy).

St-Onge, D., 1995. Detecting and correcting malapropisms with lexical chains. Unpublished master of Science Thesis, Department of Computer Science, University of Toronto, Canada.

Stairmand, M. A., 1996a. Generating lexical chains from free text. Technical Report CCL Report 96/2, Centre for Computational Linguistics, Department of Language Engineering, UMIST, Manchester, UK.

Stairmand, M. A., 1996b. An information retrieval system based on WordNet Synonym sets. Technical Report CCL Report 96/3, Centre for Computational Linguistics, Department of Language Engineering, UMIST, Manchester, UK.

Stairmand, M. A. and Black, W. J., 1996. Conceptual and contextual indexing using WordNet-derived lexical chains. In *Proceedings of 18th BCS IRSG Colloquium on Information Retrieval Research*. pp. 47–65.

Stoddard, S., 1991. *Text and Texture: Patterns of Cohesion*. Ablex, Norwood, NJ.

Stubbs, M., 1983. *Discourse Analysis - The Sociolinguistic Analysis of Natural Language*. Language in Society. Basil Blackwell, Oxford.

Stubbs, M., 1996. *Text and Corpus Analysis – Computer-Assisted Studies of Language and Culture*. Blackwell, Oxford.

Svartvik, J. (ed.), 1990. *The London-Lund Corpus of Spoken English: Description and Research*. Lund Studies in English. Lund University Press, Lund.

Svartvik, J. (ed.), 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 - Stockholm, 4-8 August 1991*. Trends in Linguistics - Studies and Monographs. Mouton De Gruyter, Berlin, New York.

Svartvik, J., 1996. Corpora are becoming mainstream. In J. Thomas and M. Short (eds.), *Using Corpora for Language Research – Studies in Honour of Geoffrey Leech*. Longman, London, pp. 3–13.

Swales, J., 1981. *Aspects of Article Introductions*. The Language Studies Unit, University of Aston, Birmingham.

Swales, J., 1990. *Genre Analysis - English in Academic and Research Settings*. Cambridge University Press, Cambridge.

Tabachnick, B. G. and Fidell, L. S., 1989. *Using Multivariate Statistics*. Second edition. Harper and Row, New York.

Thompson, G., 1996. *Introducing Functional Grammar*. Arnold, London.

Thorndyke, P. W., 1977. 'Cognitive structures in comprehension and memory of narrative discourse'. *Cognitive Psychology*, 9: 77–110.

Thury, E. M., 1988. 'A study of words relating to youth and old age in the plays of Euripides and its special implications for Euripides' *Suppliant Women*'. *Computers and the Humanities*, 22: 293–306.

Tinberg, R. J., 1988. 'The pH of a volatile genre'. *English for Specific Purposes*, 7: 205–212.

van Dijk, T. A., 1972. *Some Aspects of Text Grammars*. Mouton, The Hague.

van Dijk, T. A., 1980. *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Erlbaum, Hillsdale, NJ.

van Dijk, T. A., 1983. 'Discourse analysis: its development and application to the structure of news'. *Journal of Communication*, 33: 20–43.

van Dijk, T. A. (ed.), 1985. *Handbook of Discourse Analysis*. Academic Press, Orlando, Fla.

van Dijk, T. A. and Kintsch, W., 1983. *Strategies of Discourse Comprehension*. Academic Press, New York.

van Dijk, T. A. and Petöfi, J. S. (eds.), 1977. *Grammars and Descriptions - Study in Text Theory and Text Analysis*. Research in Text Theory. Mouton de Gruyter, Berlin.

van Rijsbergen, C. J., 1979. *Information Retrieval*. Buttersworth, London.

Ventola, E., 1979. 'The structure of casual conversation in English'. *Journal of Pragmatics*, 3: 267–298.

Ventola, E., 1986. *The Structure of Social Interaction - A Systemic Approach to the Semiotics of Service Encounters*. Frances Pinter, London.

Wessels, E. M., 1993a. Bonding and related measures of coherence in student academic writing. Unpublished ma dissertation, University of South Africa.

Wessels, E. M., 1993b. 'Lexical cohesion in student academic writing'. *South African Journal of Linguistics Supplement*, 15: 75–90.

Widdowson, H., 1978. *Teaching Language as Communication*. Oxford, Oxford University Press.

Wimmer, R. D. and Dominick, J. R., 1991. *Mass Media Research*. Third edition. Wadsworth Publishing Company, Belmont,CA.

Winburne, G., 1962. Sentence sequence in discourse. In *Proceedings of the IXth International congress of linguists, Cambridge, UK*. pp. 1094–1099.

Winter, E. O., 1971. Connection in science material: a proposition about the semantics of clause relations. In *Centre for Information on Language Teaching Papers and Reports*, no. 7. (London: Centre for Information on Language Teaching and Research for British Association for Applied Linguistics), pp. 41-52.

Winter, E. O., 1974. Replacement as a function of repetition: A study of some of its principal features in the clause relations of contemporary English. Unpublished Ph.D. thesis, University of London.

Winter, E. O., 1977. 'A clause-relational approach to English texts: a study of some predictive lexical items in written discourse'. *Instructional Science*, 6: 1–91.

Winter, E. O., 1979. 'Replacement as a fundamental function of the sentence in context'. *Forum Linguisticum*, 4: 95–133.

Woods, A., Fletcher, P., and Hughes, A., 1986. *Statistics in Language Studies*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Yarowsky, D., 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France*. pp. 454–460.

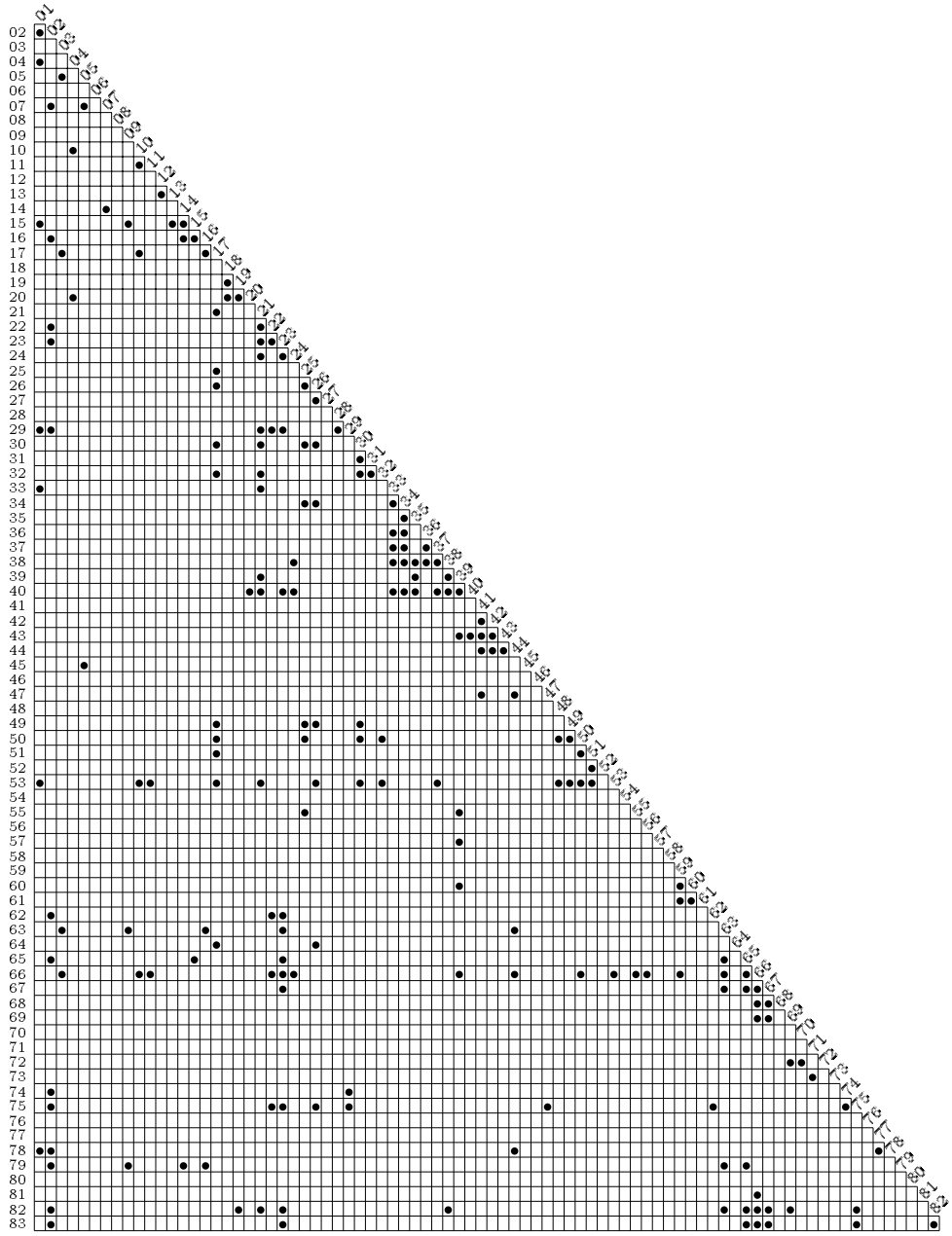
Youmans, G., 1991. 'A new tool for discourse analysis: The Vocabulary Management Profile'. *Language*, 67: 763–789.

Appendix 1

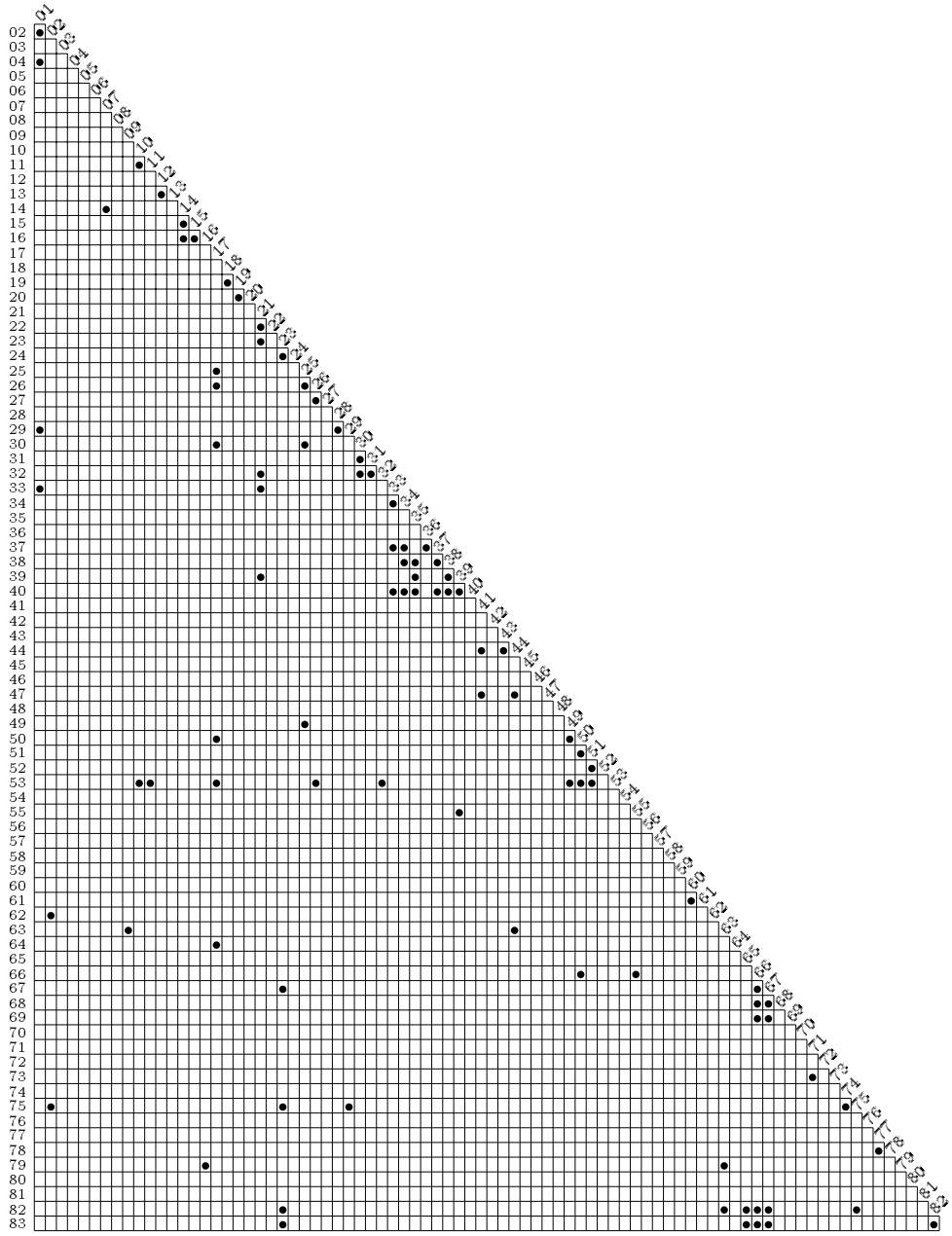
Matrices

This appendix contains the following matrices:

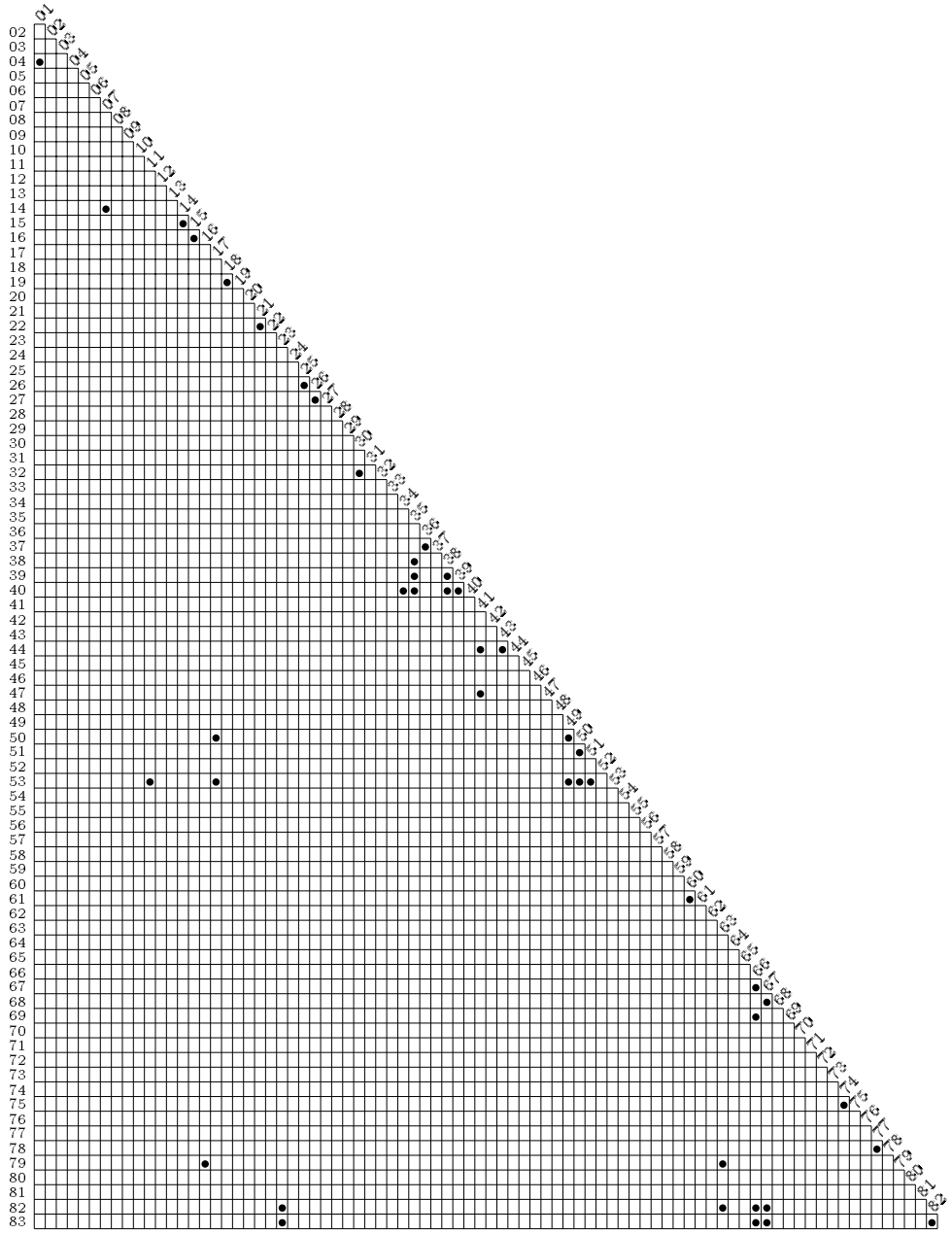
Matrix for 3-link threshold



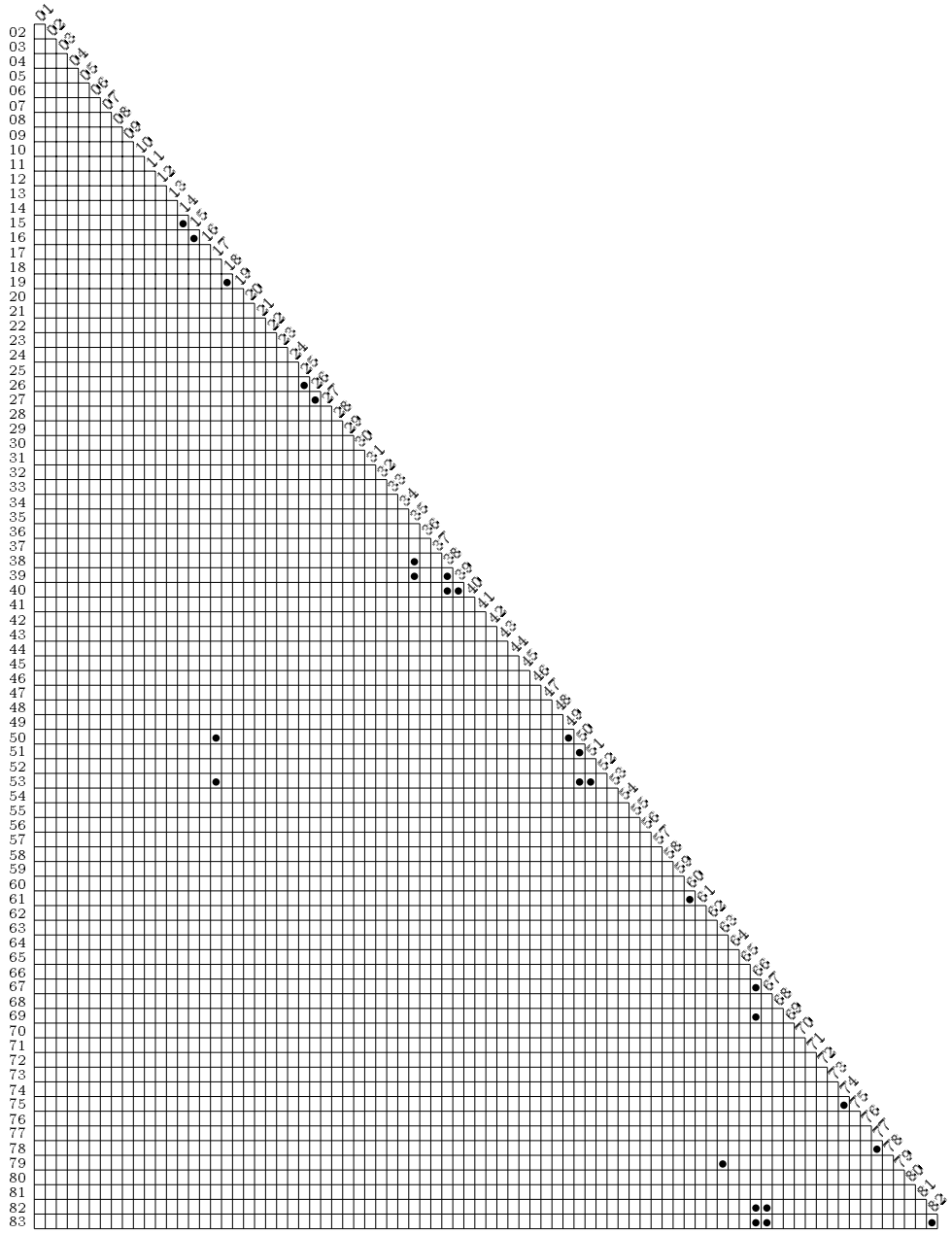
Matrix for 4-link threshold



Matrix for 5-link threshold



Matrix for 6-link threshold



Appendix 2

Stop list for `words`

a
a a consequence
a a result
a accordingly
accord
accord to
acros
ala
alway
apropo
as follow
at all event
doe
downward
der
dur
grant that
ha
hi
hundr
in ne of
in other word
inasmuch a
inde
insofar a
les
minu
mi
nevertheles
nonetheles
at the hand of
lot of
onward
ourselve
outward
plu
provid
regardles
sideway
sometime
themselve
thi
thu
tran
unles
upward
versu
wa
wherea
ye
yourselve
0
1
2
3
4
5
6
7
8
9
'd
'll
'm
'nt
's
't
've
a
a couple of
a kind of
a lot of
a matter of
able
aboard
about
above
above all
abroad
according to
accordingly
across
admitting that
after
afterward
afterwards
again
against
ago
ah
aha
ahah
ahead
alas
albeit
all
almost
along
alongside
already
also
alternatively
although
altogether
always
am
amid
amidst
among
amongst
an

and
another
anti
any
anybody
anyhow
anyone
anything
anyway
anywhere
apart
apropos
arch
are
aren
around
as
as a consequence
as a result
as far as
as follows
as if
as long as
as soon as
as though
aside
assuming that
at
at all events
at any rate
at the expense of
at the hands of
at variance with
atop
auto
away
b
back
backward
backwards
be
because
because of
been
before
beforehand
behind
being
below
beneath
beside
besides
between
beyond
billion
billionth
bn
but
but for
by
by comparison
by contrast
by dint of
by means of
by the same token
by the way
by virtue of
by way of
by way of comparison
by way of contrast
c
can
cannot
cent
co
consequently
considering
contrariwise
contrastingly
conversely
correspondingly
could
couldn
counter
d
de
despite
did
didn
directly that
ditto
does
doesn
doing
don
done
down
downwards
dr
due to
during
e
each
eh
eight
eighteen
eighteenth
eightieth
eighty
eigth
either
eleven
eleventh
else
enough
equally
even
ever
every
everybody
everyone
everything
everywhere
ex
except
except for
except that
f
few
fewer
fewest
fifteen
fifteenth
fifth
fiftieth
fifty
finally
first
five
for
for all that
for example
for the sake of
fore
forth
fortieth
forty
forward
four
fourteen
fourteenth
fourth
from
further
furthermore
g
given that
granted that
granting that
h
had
hadn

half
has
hasn
have
haven
having
he
hence
her
here
hers
herself
hey
him
himself
his
hither
hitherto
how
however
hundred
hundreth
hyper
I
if
immediately that
in
in accordance with
in addition
in addition to
in aid of
in any case
in any event
in case
in case of
in charge of
in common with
in comparison
in comparison with
in conclusion
in consequence
in contact with
in contrast
in exchange for
in face of
in favor of
in favour of
in front of
in lieu of
in line with
in need of
in order that
in other words
in place of
in quest of
in regard to
in relation to
in respect of
in return for
in search of
in spite of
in sum
in that
in that case
in the same way
in view of
inasmuch as
incidentally
indeed
inside
insofar as
instead
instead of
inter
into
is
it
its
itself
j
just
k
kind of
l
lastly
least
less
lest
like
likewise
little
lots of
m
mal
many
may
maybe
me
meantime
meanwhile
mid
might
million
millionth
mine
mini
minus
mis
miss
mm
mono
more
moreover
most
mr
mrs
ms
much
multi
must
mustn
my
myself
n
namely
near
nearby
nearly
neither
neo
never
nevertheless
next
nil
nine
nineteen
nineteenth
ninetieth
ninety
ninth
no
no sooner
no-one
nobody
non
none
none the less
nonetheless
noone
nor
not
nothing
notwithstanding
now that
nowhere
o
of
off
often
oh
oho
on

on account of
on behalf of
on pain on
on the contrary
on the one hand
on the other hand
on the strength of
on top of
once
one
only
onto
onwards
ooh
opposite
oppositely
or
other
others
otherwise
ouch
ought
oughtn
our
ours
ourselves
out
outside
outward
outwards
outwith
over
overall
ow
owing to
p
pan
partly
past
per
plus
poly
post
pre
presuming that
pro
proto
provided
providing
pseudo
q
quite
r
rather
re
regardless
round
s
save that
scarcely
second
secondly
seeing that
seldom
semi
seven
seventeen
seventeenth
seventh
seventy
several
shall
shan
she
should
shouldn
sideways
similarly
since
six
sixteen
sixteenth
sixth
sixty
so
so as
so as to
so far as
so long as
so that
some
somebody
somehow
someone
something
sometimes
somewhere
sort of
still
sub
such
such that
super
supposing
sur
t
ten
tenth
th
than
that
that is
the
the former
the latter
their
theirs
them
themself
themselves
then
thence
there
thereby
therefore
therein
these
they
thing
third
thirteen
thirtieth
thirty
this
thither
those
though
thousand
thousandth
three
thrice
through
throughout
thus
till
to
today
together with
tomorrow
too
toward
towards
trans
tri
twelfth
twelve
twentieth
twenty
twice
two
u
ugh

uh-huh
uhuh
ultra
unable
under
underneath
uni
unless
unlike
unlikely
until
up
upon
upwards
us
utmost
v
versus
very
via
vice
w
was
wasn
we
were
what
whatever
whatsoever
when
whence
whenever
where
whereas
whereby
whereof
whereupon
wherever
whether
which
whichever
while
whilst
whither
who
whoever
whom
whose
why
will
with
with regard to
within
without
worth
would
wow
x
y
yeah
yes
yesterday
yet
yippee
you
your
yours
yourself
yourselves
z
zero

Appendix 3

Lemmatisation file for `words`

good>better
good>best
bad>worse
bad>worst
arise>arose
arise>arisen
awake>awoke
awake>awoken
bear>bore
bear>borne
beat>beaten
become>became
begin>began
begin>begun
bind>bound
bite>bit
bite>bitten
bleed>bled
blow>blew
blow>blown
break>broke
break>broken
breed>bred
bring>brought
build>built
buy>bought
catch>caught
choose>chose
choose>chosen
cling>clung
creep>crept
deal>dealt
dig>dug
draw>drew
draw>drawn
drink>drank
drink>drunk
drive>drove
drive>driven
eat>ate
eat>eaten
fall>fell
fall>fallen
feed>fed
feel>felt
fight>fought
find>found
flee>fled
fly>flew
fly>flown
forbear>forbore
forbeat>forborn
forbid>forbidden
forget>forgot
forget>forgotten
forgive>forgave
forgive>forgiven
forsake>forsook
forsake>forsaken
forswear>forsworn
freeze>froze
freeze>frozen
get>got
get>gotten
give>gave
give>given
go>went
go>gone
grind>ground
grow>grew
grow>grown
hear>heard
hang>hung
hide>hid
hide>hidden
hold>held
keep>kept
know>knew
know>known
lay>laid
lead>led
leave>left
lend>lent
lose>lost
make>made
mean>meant
meet>met
pay>paid
ride>rode
ride>ridden
ring>rang
ring>rung
rise>rose
rise>risen
run>ran
saw>sawed
saw>sawn
say>said
see>saw
see>seen
seek>sought
sell>sold
send>sent
sew>sewn
shake>shook
shake>shaken
shine>shone
shoe>shod
shoot>shot
show>showed
show>shown
shrink>shrank
shrink>shrunk
sing>sang

sing>sung
sink>sank
sink>sunk
sit>sat
slay>slew
slay>slain
sleep>slept
slide>slid
sling>slung
slink>slunk
sow>sowed
sow>sown
speak>spoke
speak>spoken
spend>spent
spin>spun
spring>sprung
stand>stood
steal>stolen
stick>stuck
sting>stank
strew>strewn
stride>strode
stride>stridden
strike>struck
string>strung
strive>strove
strive>striven
swear>swore
swear>sworn
sweep>swept
swim>swam
swin>swum
swing>swung
take>took
take>taken
teach>taught
tear>tore
tear>torn
tell>told
think>thought
throw>threw
throw>thrown
tread>trod
tread>trodden
understand>understood
wear>wore
wear>worn
weep>wept
win>won
wind>wound
wring>wrung
write>wrote
write>written

burn>burnt
dream>dreamt
dwell>dwelt
fit>fitted
kneel>knelt
lean>leant
leap>leapt
light>lit
relay>relaid
smell>smelt
speed>sped
spell>spelt
spill>spilt
spoil>spoilt
wet>wetted
bid>bade
wake>woke
weave>wove
bid>bidden
lie>lain
mow>mown
prove>proven
swell>swollen
wake>woken
weave>woven
cancel>cancelled
dial>dialled
duel>duelled
enamel>enamelled
enrol>enrolled
enthral>enthralled
equal>equalled
fuel>fuelled
funnel>funnelled
hiccup>hiccupped
initial>initialled
kidnap>kidnapped
label>labelled
level>levelled
libel>libelled
marvel>marvelled
model>modelled
panel>panelled
pedal>pedalled
pencil>pencilled
program>programmed
pummel>pummelled
quarrel>quarrelled
refuel>refuelled
revel>revelled
rival>rivalled
shovel>shovelled
shrivel>shrivelled
snivel>snivelled

spiral>spiralled
stencil>stencilled
swivel>swivelled
total>totalled
travel>travelled
tunnel>tunnelled
unravel>unravelled
worship>worshipped
cancel>cancell
dial>diall
duel>duell
enamel>enamell
enrol>enroll
enthral>enthrall
equal>equall
fuel>fuell
funnel>funnell
hiccup>hiccupp
initial>initiall
kidnap>kidnapp
label>labell
level>levell
libel>libell
marvel>marvell
model>modell
panel>panell
pedal>pedall
pencil>pencill
program>programm
pummel>pummell
quarrel>quarrell
refuel>refuell
revel>revell
rival>rivall
shovel>shovell
shrivel>shrivell
snivel>snivell
spiral>spirall
stencil>stencill
swivel>swivell
total>totall
travel>travell
tunnel>tunnell
unravel>unravell
worship>worshipp
overcome>overcame
outdo>outdid
outdo>outdone
overdo>overdid
overdo>overdone
undo>undid
undid>undone
withdraw>withdrawn
overeat>overate

overeat>overeaten
befall>befell
befall>befallen
forego>forewent
forego>forgone
undergo>underwent
undergo>undergone
outgrow>outgrew
outgrow>outgrown
mishear>misheard
behold>beheld
uphold>upheld
withhold>withheld
mislead>misled
remake>remade
repay>repaid
override>overridden
outrun>outran
overrun>overran
re-run>re-ran
foresee>forsaw
foresee>forseen
oversee>oversaw
oversee>overseen
outsell>outsold
resell>resold
outshine>outshone
outshine>outshone
overshoot>overshot
oversleep>overslept
withstand>withstood
hamstring>hamstrung
mistake>mistook
overtake>overtook
overtake>overtaken
retake>retook
retake>retaken
undertake>undertook
undertake>undertaken
foretell>foretold
retell>retold
rethink>rethought
overthrow>overthrew
overthrow>overthrown
misunderstand>misunderstood
rewind>rewound
rewrite>rewrote
rewrite>rewritten
underwrite>underwrote
underwrite>underwritten
bottle-feed>bottle-fed
breast-feed>breast-fed
force-feed>force-fed
spoon-feed>spoon-fed

baby-sit>baby-sat
ghost-write>ghost-wrote
ghost-write>ghost-written
ban>bann
bar>barr
bat>batt
beg>begg
blot>blott
blur>blurr
bob>bobb
brag>bragg
brim>brimm
bug>bugg
cap>capp
chat>chatt
chip>chipp
chop>chopp
clap>clapp
clog>clogg
clot>clott
cram>cramm
crib>cribb
crop>cropp
cup>cupp
dab>dabb
dam>damm
dim>dimm
din>dinn
dip>dipp
dot>dott
drag>dragg
drop>dropp
drug>drugg
drum>drumm
dub>dubb
fan>fann
fit>fitt
flag>flagg
flap>flapp
flip>flipp
flop>flopp
fog>fogg
fret>frett
gas>gass
gel>gell
glut>glutt
grab>grabb
grin>grinn
grip>gripp
grit>gritt
grub>grubb
gun>gunn
gut>gutt
hem>hemm

hop>hopp
hug>hugg
hum>humm
jam>jamm
jet>jett
jig>jigg
jog>jogg
jot>jott
knit>knitt
knot>knott
lag>lagg
lap>lapp
log>logg
lop>lopp
mar>marr
mob>mobb
mug>mugg
nag>nagg
net>nett
nip>nipp
nod>nodd
pad>padd
pat>patt
peg>pegg
pen>penn
pet>pett
pin>pinn
plan>plann
plod>plodd
plug>plugg
pop>popp
prod>prodd
prop>propp
rib>ribb
rig>rigg
rob>robb
rot>rott
rub>rubb
sag>sagg
scan>scann
scar>scarr
scrap>scrapp
scrub>scrubb
ship>shipp
shop>shopp
shred>shredd
shrug>shrugg
shun>shunn
sin>sinn
sip>sipp
skid>skidd
skim>skimm
skin>skinn
skip>skipp

slam>slamm
slap>slapp
slim>slimm
slip>slipp
slop>slopp
slot>slott
slum>slumm
slur>slurr
snag>snagg
snap>snapp
snip>snipp
snub>snubb
sob>sobb
spot>spott
squat>squatt
stab>stabb
star>starr
stem>stemm
step>stepp
sitr>sitrr
stop>stopp
strap>strapp
strip>stripp
strut>strutt
stun>stunn
stunt>stuntt
sun>sunn
swab>swabb
swap>swapp
swat>swatt
swig>swigg
swot>swott
tag>tagg
tan>tann
tap>tapp
thin>thinn
thorb>thorbb
tip>tipp
top>topp
trap>trapp
trek>trekk
trim>trimm
trip>tripp
trot>trott
vet>vett
wag>wagg
wrap>wrapp
abet>abett
abhor>abhorr
acquit>acquitt
admit>admitt
allot>allott
commit>committ
compel>compell
confer>conferr
control>controll
defer>deferr
deter>deterr
distil>distill
embed>embedd
emit>emitt
enrol>enroll
enthral>enthrall
emit>emitt
enrol>enroll
equip>equipp
excel>excell
expel>expell
incur>incurr
instil>instill
occur>occurr
omit>omitt
outwit>outwitt
patrol>patroll
propel>propell
rebel>rebell
rebut>rebutt
recap>recapp
recur>recurr
refer>referr
regret>regrett
remit>remitt
repel>repell
submit>submitt
transfer>transferr
transmit>transmitt
handicap>handicapp

Appendix 4

Computer and manual analysis of the same text

Masters of Political Thought (Hoey, 1991b, pp.249–252)

[1] What is attempted in the following volume is to present to the reader a series of actual excerpts from the writings of the greatest political theorists of the past; selected and arranged so as to show the mutual coherence of various parts of an author's thought and his historical relation to his predecessors or successors; and accompanied by introductory notes and intervening comments designed to assist the understanding of the meaning and importance of the doctrine quoted. [2] The book does not purport to be a history of political theory, with quotations interspersed to illustrate the history. [3] It is rather a collection of texts, to which I have endeavoured to supply a commentary. [4] I have tried rather to render the work of Aristotle, Augustine, and the rest accessible to the students, than to write a book about them; and the main object of this work will have been achieved if it serves not as a substitute for a further study of the actual works of these authors, but as an incentive to undertake it.

[5] Nor does the commentary make any pretension of being exhaustive. [6] Very often after a long passage has been quoted a single point has been selected for comment; and sometimes this point has been selected not because it was the most important, but because it was one which I had something to say. [7] I have not tried to cover all ground, and shall have done my part if the reader is stimulated, by the samples which I have offered, to complete a commentary of his own.

[8] The selection has been confined to a few authors, for reasons not only of space, or of limitations of my own knowledge (though either of these reasons would have been sufficient), but because it is part of the plan of the book to concentrate attention upon the most important works. [9] A knowledge of Plato's Republic, of Aristotle's Politics, of parts of Augustine's City of God, belongs to a general education. [10] The works of lesser writers, or the lesser works of these writers, are doubtless worth reading; but a man who is not a specialist may ignore them without reproach.

[11] If the commentary is secondary to the text, still more so must be any introductory remarks which I make here. [12] In commending the writings which follow to the reader's attention, I will indeed stake my credit on the assertion that the study of them will correct the judgment and enlighten the understanding upon matters in which it is important to be enlightened and correct. [13] But if a proof of this assertion is demanded, there is no proof except that of asking the inquirer to make an experiment. [14] The introducer may suggest lines of reasoning, he may try to convey certain lights which he has himself derived from the study, but in doing this he must be tentative and not dogmatic, and in the last resort he must say to the reader, 'Go and read for yourself, and try whether this is confirmed by your experience". [15] In this respect his position is like that of the critic

of a work of art. [16] However useful the critic's remarks may be in preparing an approach to the work, they can never dispense the reader from the necessity of studying the work itself, nor deprive him of the right, on the basis of this study, of turning critic himself and standing in judgment on the reasonings by which he was led to it in the first place.

Key

***** Link has been picked out by manual analysis as well

+ Link has not been picked out by manual analysis

Link has been disregarded for comparative purposes because it was considered arguable by Hoey (1991)

$ Link has been disregarded for comparative purposes because it is beyond capability of computer program (substitution, deixis, ellipsis)

Links detected by computer analysis

Sentences		Links
1	2	political→political*
1	4	actual→actual*, author→authors*
1	6	selected→selected+, comments→comment*, quoted→quoted*
1	7	reader→reader*, parts→part+
1	8	parts→part+, author→authors*
1	9	parts→parts+
1	11	introductory→introductory#
1	12	following→follow*, reader→reader*, writings→writings*, understanding→understanding+
1	14	reader→reader*
1	16	reader→reader*
2	4	book→book*
2	8	book→book*
3	5	commentary→commentary*
3	7	commentary→commentary*
3	11	texts→text*, commentary→commentary*
4	7	tried→tried*
4	8	work→works*, book→book*, authors→authors*
4	9	Aristotle→Aristotle*, Augustine→Augustine*
4	10	work→works*
4	12	study→study*
4	14	study→study*
4	15	work→work+
4	16	work→work*, study→study+
5	7	commentary→commentary*
5	11	commentary→commentary*, make→make+
5	13	make→make+
6	8	important→important+
6	12	important→important+
6	14	say→say+
7	8	part→part+, own→own+
7	9	part→parts+
7	11	commentary→commentary*
7	12	reader→reader*
7	14	reader→reader*
7	16	reader→reader*
8	9	knowledge→knowledge+, part→parts+
8	10	works→works*
8	12	attention→attention+, important→important#

Continued on next page

Continued from previous page

Sentences		Links
8	14	reasons→reasoning+
8	15	works→work+
8	16	reasons→reasonings+, works→work*
10	14	reading→read*
10	15	works→work+
10	16	works→work*
11	13	make→make+
11	16	remarks→remarks*
12	13	assertion→assertion*
12	14	reader→reader*, study→study*
12	16	reader→reader*, study→study*, judgment→judgment+
14	16	reasoning→reasonings+, study→study*, reader→reader*
15	16	critic→critic*, work→work*

Key

***** Link has been picked out by computer analysis as well

+ Link has not been picked out by computer analysis

Link has been disregarded for comparative purposes because it was considered arguable by Hoey (1991)

$ Link has been disregarded for comparative purposes because it is beyond capability of computer program (substitution, deixis, ellipsis)

Links detected by manual analysis

Sentences		Links
1	2	political→political*, theorists→theory+, historical→history+, quoted→quotations+, intervening→interspersed+, volume→book+
1	3	comments→commentary+, attempted→endeavoured+
1	4	actual→actual* author→authors*, writings→works+, attempted→tried+, volume→book+
1	5	comments→commentary+
1	6	quoted→quoted*, comments→comment*, excerpt→passage+
1	7	reader→reader*, comments→commentary+, present→offered+, attempted→tried+
1	8	author→authors*, selected→selection+, importance→important+, writings→works+, volume→book+
1	9	political→politics+
1	10	reader→reading+, writings→writers+
1	11	introductory→introductory#, comments→commentary+, notes→remarks#
1	12	writings→writings*, reader→reader*, following→follow*, importance→important+
1	14	reader→read*, introductory→introducer#, attempted→try+
1	16	reader→reader*
2	3	the book→it+
2	4	book→book*
2	6	quotations→quoted+
2	8	book→book*
2	9	political→politics+
3	4	I→I#, endeavoured→tried +
3	5	commentary→commentary*
3	6	I→I#, commentary→comment+
3	7	commentary→commentary*, I→I#, endeavoured→tried+
3	8	I→my#, of texts→0$
3	11	commentary→commentary*, texts→text*, I→I#
3	12	I→I#
3	14	endeavoured→try+, I→Introducer#

Continued on next page

Continued from previous page

Sentences		Links
4	6	I→I#
4	7	tried→tried*, I→I#, object...achieved→done my part+
4	8	authors→authors*, book→book*, works→works*, I→my#
4	9	Aristotle→Aristotle*, Augustine→Augustine*
4	10	works→works*, authors→writers+
4	11	I→I#
4	12	study→study*, I→I#, works→writings+
4	14	tried→try+, study→study*, I→Introducer#
4	16	work→work*
5	6	commentary→comment+
5	7	commentary→commentary*, being exhaustive→cover all the ground+
5	11	commentary→commentary*
6	7	I→I#, comment→commentary+
6	8	I→my#
6	11	I→I#, comment→commentary*
6	12	I→I#
6	14	I→Introducer#
7	8	I→my#
7	10	reader→reading+
7	11	commentary→commentary+, I→I#
7	12	reader→reader*, I→I#
7	14	tried→try+, reader→read*, I→Introducer+
7	16	reader→reader*
8	10	works→works*, authors→writers+
8	11	my→I#
8	12	my→I#, important→important#
8	14	my→introducer#
8	16	works→work*
9	10	Plato etc→these$
10	12	reading→reader+, writer→writings
10	14	reading→read*
10	16	works→work*
11	12	I→I#
11	14	introductory→introducer#
11	16	remarks→remarks*
12	13	assertion→assertion*
12	14	study→study*, reader→reader*, enlighten→lights+
12	16	study→study*, reader→reader*
14	15	introducer→his$, sentence 14→this$
14	16	study→study*, reader→reader*, reasoning→reasoning+
15	16	work→work*, critic→critic*, of art→0$

Appendix 5

San Marino text

[0001] (**Introduction**) San Marino, republic in southern Europe, an enclave in northern Italy, south of the city of Rimini. *[0002]* With a total area of only 61 sq km (24 sq mi), San Marino is one of the smallest republics in the world. *[0003]* **Land and Population** Located in the central Apennines, east of Florence, Italy, San Marino has a terrain dominated by the three-peaked Mount Titano (739 m/2424 ft). *[0004]* The country is watered by several streams, including the Ausa, Marano, and San Marino. *[0005]* The climate is mild with an average annual precipitation of 686 mm (27 in). *[0006]* The population of San Marino (1989 estimate) was 22,900. *[0007]* The people speak Italian, use Italian currency, and are mostly Roman Catholic. *[0008]* The capital is San Marino (population, 1990 estimate, 4185), which is located on the slopes of Mount Titano. *[0009]* Other population centers include Borgo Maggiore, on the mountain's lower slope, and Serravalle. *[0010]* **Economy and Government** The economy of San Marino is based on agriculture, but light industry is of growing importance. *[0011]* In the late 1980s annual government revenue and expenditure were balanced at about $183 million. *[0012]* Wheat, barley, maize, olives, wine, and livestock and dairy products dominate agricultural output. *[0013]* Some building stone is quarried. *[0014]* Manufactures include textiles, cement, leather goods, synthetic rubber products, and ceramics. *[0015]* Other important sources of income are tourism and the sale of postage stamps. *[0016]* San Marino is governed by the Great and General Council, a legislative body of 60 members, elected by universal suffrage for a term of five years. *[0017]* Two members of the council, called captains-regent, are elected for six months to preside over the country's executive body, the Congress of State. *[0018]* **History** According to tradition, San Marino was founded in AD 301 by a Christian stonecutter, Marinus, who sought refuge on Mount Titano from religious persecution. *[0019]* The commune that developed maintained its sovereignty, despite repeated incursions by neighboring rulers of Rimini, and in 1291 Pope Nicholas IV recognized San Marino's independence. *[0020]* The governing laws of the republic were promulgated during the Middle Ages. *[0021]* San Marino has had a treaty of friendship (revised several times) with Italy since 1862. *[0022]* From 1945 to 1957 the republic was ruled by a coalition of Communists and Socialists. *[0023]* In 1957 the Christian Democratic party, aided by Communist dissidents, took control of the government. *[0024]* In 1978 a coalition led by Communists again came to power. *[0025]* The 1983 election left control in leftist hands, but in July 1986 a new Christian Democrat-Communist coalition was formed. *[0026]* In March

1992 the Christian Democrats formed a coalition government with the Socialists, a status which continued after the May 1993 general election. *[0027]* San Marino became a member of the United Nations in 1992. *[0028]* "San Marino," Microsoft (R) Encarta. *[0029]*

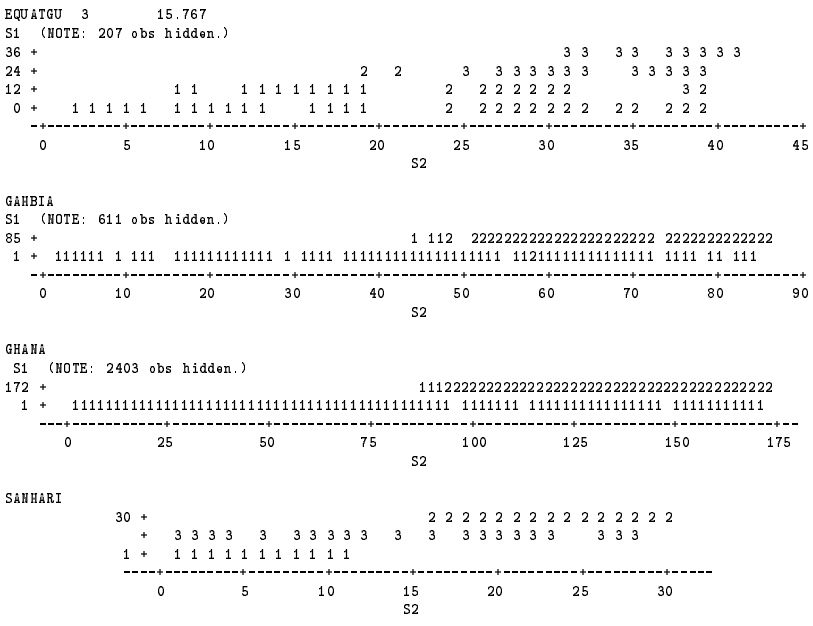
Appendix 6

Chosen CCC values

Text File	Clusters	CCC peak
botswa	2	9.584
burundi	2	15.459
cameroo	2	23.846
centafr	2	17.276
comoros	3	11.965
congo	2	16.376
cotediv	2	18.926
djibou	3	4.454
equatgu	3	15.767
gabon	2	16.943
gambia	2	12.798
ghana	2	39.733
lesotho	2	17.248
liberia	2	28.861
malawi	2	18.616
mozamb	2	17.993
namibia	2	19.112
niger	2	21.770
sanmari	3	13.254
senegal	2	13.742
sierral	2	21.451
somalia	2	25.171
swazil	3	12.402
togo	2	17.859
uganda	2	34.128

Appendix 7

Sample plots of clusters



Appendix 8

LSM segmentation performance by text

Text	Sections	Provis. boundaries (Peaks)	Final Boundaries	Matching boundaries	Recall		Precision	
					%	Rank	%	Rank
1	9	29	12	3	33.33	10	25.00	19
2	14	29	15	3	21.43	22	20.00	22
3	20	51	25	7	35.00	8	28.00	16
4	15	33	15	5	33.33	10	33.33	8
5	5	9	4	1	20.00	23	25.00	19
6	15	35	17	5	33.33	10	29.41	13
7	19	34	17	6	31.58	15	35.29	6
8	5	11	7	0	0.00	25	0.00	25
9	5	14	8	4	80.00	1	50.00	1
10	17	38	18	7	41.18	6	38.89	5
11	15	37	16	2	13.33	24	12.50	24
12	26	64	35	9	34.62	9	25.71	18
13	12	23	14	4	33.33	10	28.57	14
14	27	56	27	7	25.93	19	25.93	17
15	19	41	21	6	31.58	15	28.57	14
16	23	40	25	10	43.48	5	40.00	3
17	8	19	10	4	50.00	4	40.00	3
18	23	43	20	6	26.09	18	30.00	12
19	20	44	23	8	40.00	7	34.78	7
20	4	8	6	3	75.00	2	50.00	1
21	29	45	24	8	27.59	17	33.33	8
22	23	38	24	5	21.74	21	20.83	21
23	26	41	18	6	23.08	20	33.33	8
24	3	18	10	2	66.67	3	20.00	22
25	18	41	19	6	33.33	10	31.58	11

Appendix 9

Random segmentation performance by text

Text	Sections	Provis. bound-aries	Final Bound-aries	Match-ing bound-aries	Recall		Precision	
					%	Rank	%	Rank
1	9	23	17	3	33.33	5	17.65	11
2	14	34	26	2	14.29	21	7.69	24
3	20	48	30	5	25.00	13	16.67	12
4	15	37	23	5	33.33	5	21.74	7
5	5	14	9	2	40.00	1	22.22	6
6	15	32	22	2	13.33	22	9.09	23
7	19	33	19	5	26.32	10	26.32	3
8	5	14	11	2	40.00	1	18.18	10
9	5	13	8	1	20.00	16	12.50	18
10	17	33	26	2	11.76	23	7.69	24
11	15	33	18	3	20.00	16	16.67	12
12	26	65	47	10	38.46	3	21.28	8
13	12	34	17	4	33.33	5	23.53	5
14	27	59	33	7	25.93	12	21.21	9
15	19	36	24	7	36.84	4	29.17	1
16	23	38	24	7	30.43	9	29.17	1
17	8	28	15	2	25.00	13	13.33	17
18	23	38	26	4	17.39	18	15.38	15
19	20	38	27	3	15.00	20	11.11	20
20	4	10	9	1	25.00	13	11.11	20
21	29	41	26	3	10.34	25	11.54	19
22	23	39	23	6	26.09	11	26.09	4
23	26	51	32	3	11.54	24	9.38	22
24	3	12	7	1	33.33	5	14.29	16
25	18	38	18	3	16.67	19	16.67	12

Appendix 10

Expert segmentation performance by text

Text	Sections	Inserted Boundaries	Matching Boundaries	Recall		Precision	
				%	Rank	%	Rank
1	9	6	6	66.67	3	100	1
2	14	3	3	21.43	18	100	1
3	19	13	13	68.42	2	100	1
4	16	4	4	25.00	17	100	1
5	5	2	2	40.00	12	100	1
6	15	2	2	13.33	22	100	1
7	19	8	8	42.11	11	100	1
8	5	1	1	20.00	19	100	1
9	5	2	2	40.00	12	100	1
10	17	11	11	64.71	4	100	1
11	15	2	2	13.33	22	100	1
12	26	5	5	19.23	20	100	1
13	12	2	2	16.67	21	100	1
14	27	3	3	11.11	24	100	1
15	19	5	5	26.32	16	100	1
16	23	12	12	52.17	5	100	1
17	8	6	6	75.00	1	100	1
18	23	11	11	47.83	7	100	1
19	20	10	10	50.00	6	100	1
20	4	0	0	0.00	25	.	.
21	29	13	13	44.83	9	100	1
22	23	10	10	43.48	10	100	1
23	26	12	12	46.15	8	100	1
24	3	1	1	33.33	14	100	1
25	18	6	6	33.33	14	100	1

Appendix 11

Recall by LSM and TextTile

Text	Both LSM and TextTile	LSM only	TextTile only	Total
1	2 28.57%	1 14.29%	4 57.14%	7 100%
2	1 20.00%	2 40.00%	2 40.00%	5 100%
3	5 31.25%	2 12.50%	9 56.25%	16 100%
4	0 0.00%	5 62.50%	3 37.50%	8 100%
5	0 0.00%	1 33.33%	2 66.67%	3 100%
6	0 0.00%	5 71.43%	2 28.57%	7 100%
7	2 16.67%	4 33.33%	6 50.00%	12 100%
8	0 0.00%	0 0.00%	1 100.00%	1 100%
9	1 20.00%	3 60.00%	1 20.00%	5 100%
10	7 63.64%	0 0.00%	4 36.36%	11 100%
11	0 0.00%	2 50.00%	2 50.00%	4 100%
12	1 7.69%	8 61.54%	4 30.77%	13 100%
13	0 0.00%	4 66.67%	2 33.33%	6 100%
14	0 0.00%	7 70.00%	3 30.00%	10 100%

Continued on next page

Continued from previous page

Text	Both LSM and TextTile	LSM only	TextTile only	Total
15	2 22.22%	4 44.44%	3 33.33%	9 100%
16	4 22.22%	6 33.33%	8 44.44%	18 100%
17	3 42.86%	1 14.29%	3 42.86%	7 100%
18	1 6.25%	5 31.25%	10 62.50%	16 100%
19	4 28.57%	4 28.57%	6 42.86%	14 100%
20	0 0.00%	3 100.00%	0 0.00%	3 100%
21	5 31.25%	3 18.75%	8 50.00%	16 100%
22	1 7.14%	4 28.57%	9 64.29%	14 100%
23	4 28.57%	2 14.29%	8 57.14%	14 100%
24	1 50.00%	1 50.00%	0 0.00%	2 100%
25	3 33.33%	3 33.33%	3 33.33%	9 100%
Total	47	80	103	230

Appendix 12

Text 9

[0001] Equatorial Guinea, independent republic in western Africa, consisting of a mainland section (Río Muni) on the western coast and the coastal islets of Corisco, Elobey Grande, and Elobey Chico as well as the islands of Bioko (formerly Macías Nguema Biyogo and previously Fernando Po), and Annobón (Pagalu) in the Gulf of Guinea; total area, 28,051 sq km (10,831 sq mi). **[0002] Land and Resources** Mainland Equatorial Guinea is bounded on the north by Cameroon, on the east and south by Gabon, and on the west by the Gulf of Guinea. *[0003]* The terrain is gently rolling and heavily forested; about 60 percent of the area is drained by the Mbini (formerly Benito) River. *[0004]* With Corisco and the Elobeys islands it comprises the continental (formerly Río Muni) region, an area of 26,017 sq km (10,045 sq mi). *[0005]* The main island of Equatorial Guinea is Bioko (2017 sq km/779 sq mi), which is located off the western coast of Africa in the Bight of Bonny (Biafra). *[0006]* The island, primarily of volcanic origin, is mountainous and thickly wooded, with a steep, rocky coast. *[0007]* Its highest peak is Pico de Santa Isabel (3008 m/9868 ft). *[0008]* The island has fertile volcanic soils and is watered by several streams, and lakes are found in the mountains. *[0009]* Together with the small island of Annobón, lying about 640 km (about 400 mi) to the southwest, it comprises the insular (formerly Bioko) region. *[0010]* The climate is tropical; the average annual temperature is about 25 C (about 77 F) and the annual rainfall is more than 2005 mm (more than 79 in) in most areas. **[0011] Population** The population of Equatorial Guinea (1990 estimate) was 348,000. *[0012]* The overall population density was about 12 persons per sq km (about 32 per sq mi). *[0013]* The population is composed almost entirely of black Africans: the Bantu-speaking Bubis, most of whom live on Bioko; the Bengas on Elobey and Corisco; and the Fang (Spanish Pamúes) on the mainland. *[0014]* Persons of European descent and of mixed black and European descent make up the remainder. *[0015]* Spanish is the official language, and Roman Catholicism is the predominant religion. *[0016]* The capital of the continental region is Bata (1983 census, 24,100), on the mainland, and the largest city, chief port, and capital of the republic is Malabo, formerly Santa Isabel (15,253), on the northern coast of Bioko. **[0017] Economy and Government** Agriculture is the main source of livelihood in Equatorial Guinea. *[0018]* The principal export is cacao, which is grown almost entirely on Bioko. *[0019]* Coffee is grown on the mainland, which also produces tropical hardwood timber. *[0020]* Rice, bananas, yams, and millet are the staple foods. *[0021]* Local manufacturing industries include the processing of oil and soap, cacao,

yucca, coffee, and seafood. *[0022]* The monetary system is based on the franc system (2864 CFA francs equal U S $1; 1990). *[0023]* Under the 1982 constitution, the president was elected by universal suffrage to a seven-year term, and members of the legislature were elected to five-year terms. *[0024]* The Democratic Party of Equatorial Guinea was the sole legal political party. *[0025]* A new multiparty constitution was approved in 1991. **[0026] History** The island of Fernando Po was sighted in 1471 by Fernão do Po, a Portuguese navigator. *[0027]* Portugal ceded the island to Spain in 1778. *[0028]* From 1827 to 1844, with the permission of the Spanish government, Great Britain maintained a naval station at Fernando Po and also administered the island. *[0029]* In 1844 the Spanish settled in the area that became the province of Río Muni. *[0030]* In 1904 Fernando Po and Río Muni were organized into the Western African Territories, later known as Spanish Guinea. *[0031]* On October 12, 1968, the territory became the independent republic of Equatorial Guinea, with Francisco Macias Nguema as president. *[0032]* In 1972 Nguema appointed himself president for life. *[0033]* Extreme dictatorial and repressive policies led to the flight of an estimated 100,000 refugees to neighboring countries; at least 50,000 of those who remained were killed, and another 40,000 were sent into forced labor. *[0034]* In 1979 Nguema was overthrown in a military coup, tried for treason, and executed. *[0035]* Lieutenant Colonel Teodoro Obiang Nguema Mbasogo, who led the coup, then became president. *[0036]* Parliamentary elections, based on a single slate of candidates, were held in 1983 and 1988. *[0037]* Although the first multiparty elections took place in November 1993, they were internationally condemned and boycotted by approximately 80 percent of the eligible voters. *[0038]* Opposition forces called for the boycott after the Obiang Nguema government refused to prepare an accurate electoral roll and guarantee the right to campaign without harassment. **[0039] Further Reading** "Equatorial Guinea," Microsoft (R) Encarta. *[0040]*

Appendix 13

Links in text 9

1 link or more

Sentences	Total Links	Links
0001 0002	4	equatorial guinea mainland gulf
0001 0003	2	formerly area
0001 0004	8	río muni corisco elobey island formerly area sq
0001 0005	8	equatorial guinea western africa coast island bioko sq
0001 0006	2	coast island
0001 0008	1	island
0001 0009	4	island bioko formerly annobón
0001 0010	1	area
0001 0011	2	equatorial guinea
0001 0012	1	sq
0001 0013	4	mainland corisco elobey bioko
0001 0016	5	republic mainland coast bioko formerly
0001 0017	2	equatorial guinea
0001 0018	1	bioko
0001 0019	1	mainland
0001 0024	2	equatorial guinea
0001 0026	3	island fernando po
0001 0027	1	island
0001 0028	3	island fernando po
0001 0029	3	río muni area
0001 0030	6	guinea western río muni fernando po
0001 0031	4	equatorial guinea independent republic
0001 0032	1	nguema
0001 0034	1	nguema
0001 0035	1	nguema
0001 0038	1	nguema
0001 0039	2	equatorial guinea

Continued on next page

Continued from previous page

Sentences	Total Links	Links
0002 0005	2	equatorial guinea
0002 0011	2	equatorial guinea
0002 0013	1	mainland
0002 0016	1	mainland
0002 0017	2	equatorial guinea
0002 0019	1	mainland
0002 0024	2	equatorial guinea
0002 0030	1	guinea
0002 0031	2	equatorial guinea
0002 0039	2	equatorial guinea
0003 0004	2	area formerly
0003 0009	1	formerly
0003 0010	1	area
0003 0016	1	formerly
0003 0029	1	area
0003 0037	1	percent
0003 0038	1	roll
0004 0005	3	island sq km
0004 0006	1	island
0004 0008	1	island
0004 0009	5	island comprise formerly region km
0004 0010	1	area
0004 0012	2	sq km
0004 0013	2	corisco elobey
0004 0016	3	continental formerly region
0004 0026	1	island
0004 0027	1	island
0004 0028	1	island
0004 0029	3	río muni area
0004 0030	2	río muni
0005 0006	2	island coast
0005 0008	1	island
0005 0009	3	island bioko km
0005 0011	2	equatorial guinea
0005 0012	2	sq km
0005 0013	1	bioko
0005 0016	2	bioko coast
0005 0017	3	main equatorial guinea
0005 0018	1	bioko
0005 0024	2	equatorial guinea
0005 0026	1	island
0005 0027	1	island

Continued on next page

Continued from previous page

Sentences	Total Links	Links
0005 0028	1	island
0005 0030	2	guinea western
0005 0031	2	equatorial guinea
0005 0039	2	equatorial guinea
0006 0008	2	island volcanic
0006 0009	1	island
0006 0016	1	coast
0006 0026	1	island
0006 0027	1	island
0006 0028	1	island
0007 0016	2	santa isabel
0008 0009	1	island
0008 0026	1	island
0008 0027	1	island
0008 0028	1	island
0009 0012	1	km
0009 0013	1	bioko
0009 0016	3	formerly bioko region
0009 0018	1	bioko
0009 0026	1	island
0009 0027	1	island
0009 0028	1	island
0010 0019	1	tropical
0010 0029	1	area
0011 0012	1	population
0011 0013	1	population
0011 0017	2	equatorial guinea
0011 0024	2	equatorial guinea
0011 0030	1	guinea
0011 0031	2	equatorial guinea
0011 0039	2	equatorial guinea
0012 0013	1	population
0012 0014	1	person
0013 0014	1	black
0013 0015	1	spanish
0013 0016	2	bioko mainland
0013 0018	2	entirely bioko
0013 0019	1	mainland
0013 0028	1	spanish
0013 0029	1	spanish
0013 0030	2	african spanish
0015 0028	1	spanish

Continued on next page

Continued from previous page

Sentences	Total Links	Links
0015 0029	1	spanish
0015 0030	1	spanish
0016 0018	1	bioko
0016 0019	1	mainland
0016 0031	1	republic
0017 0024	2	equatorial guinea
0017 0028	1	government
0017 0030	1	guinea
0017 0031	2	equatorial guinea
0017 0038	1	government
0017 0039	2	equatorial guinea
0018 0019	1	grown
0018 0021	1	cacao
0019 0021	1	coffee
0022 0036	1	ba
0023 0025	1	constitution
0023 0031	1	president
0023 0032	1	president
0023 0035	1	president
0024 0030	1	guinea
0024 0031	2	equatorial guinea
0024 0039	2	equatorial guinea
0025 0037	1	multiparty
0026 0027	1	island
0026 0028	3	island fernando po
0026 0030	2	fernando po
0027 0028	1	island
0028 0029	1	spanish
0028 0030	3	spanish fernando po
0028 0038	1	government
0029 0030	3	spanish río muni
0029 0031	1	became
0029 0035	1	became
0030 0031	1	guinea
0030 0039	1	guinea
0031 0032	1	president
0031 0035	2	became president
0031 0039	2	equatorial guinea
0032 0034	1	nguema
0032 0035	2	nguema president
0032 0038	1	nguema
0033 0035	1	led

Continued on next page

Continued from previous page

Sentences	Total Links	Links
0034 0035	2	nguema coup
0034 0038	1	nguema
0035 0038	2	obiang nguema
0036 0037	1	election
0037 0038	1	boycott
0039 0040	1	microsoft
0040 0041	2	copyright corporation

Appendix 14

Link sets in text 9

1 link or more

Sentence	Median	Link set
1	13	2 2 2 2 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 6 6 8 9 9 9 9 9 10 11 11 12 12 13 13 13 13 16 16 16 16 16 17 17 18 19 24 24 26 26 26 27 28 28 28 29 29 29 30 30 30 30 30 30 31 31 31 31 31 32 34 35 38 39 39
2	16.5	1 1 1 1 5 5 11 11 13 16 17 17 19 24 24 30 31 31 39 39
3	9.5	1 1 4 4 9 10 16 29 37 38
4	9	1 1 1 1 1 1 1 1 1 3 3 5 5 5 6 8 9 9 9 9 9 10 12 12 13 13 16 16 16 26 27 28 29 29 29 30 30
5	11	1 1 1 1 1 1 1 1 1 2 2 4 4 4 6 6 8 9 9 9 11 11 12 12 13 16 16 17 17 17 18 24 24 26 27 28 30 30 31 31 39 39
6	8	1 1 4 5 5 8 8 9 16 26 27 28
7	16	16 16
8	6	1 4 5 6 6 9 26 27 28
9	5	1 1 1 1 1 3 4 4 4 4 4 5 5 5 6 8 12 13 16 16 16 18 26 27 28
10	4	1 3 4 19 29
11	17	1 1 2 2 5 5 12 13 17 17 24 24 30 31 31 39 39
12	5	1 1 4 4 5 5 9 11 13 14
13	13	1 1 1 1 2 4 4 5 9 11 12 14 15 16 16 18 18 19 28 29 30 30
14	12.5	12 13
15	28.5	13 28 29 30
16	5	1 1 1 1 1 2 3 4 4 4 5 5 6 7 7 9 9 9 13 13 18 19 31
17	17.5	1 1 2 2 5 5 5 11 11 24 24 28 30 31 31 38 39 39

Continued on next page

Continued from previous page

Sentence	Median	Link set
18	13	1 5 9 13 13 16 19 21
19	13	1 2 10 13 16 18 21
21	18.5	18 19
22	36	36
23	31.5	25 31 32 35
24	11	1 1 2 2 5 5 11 11 17 17 30 31 31 39 39
25	30	23 37
26	8.5	1 1 1 4 5 6 8 9 27 28 28 28 30 30
27	7	1 4 5 6 8 9 26 28
28	16	1 1 1 4 5 6 8 9 13 15 17 26 26 26 27 29 30 30 30 38
29	11.5	1 1 1 3 4 4 4 10 13 15 28 30 30 30 31 35
30	13	1 1 1 1 1 1 2 4 4 5 5 11 13 13 15 17 24 26 26 28 28 28 29 29 29 31 39
31	20	1 1 1 1 1 2 2 5 5 11 11 16 17 17 23 24 24 29 30 32 32 34 35 35 35 38 39 39
32	32.5	1 23 31 31 34 35 35 38
33	35	35
34	33.5	1 31 32 35 35 38
35	32	1 23 29 31 31 31 32 32 33 34 34 38 38
36	29.5	22 37
37	30.5	3 25 36 38
38	31.5	1 3 17 28 31 32 34 35 35 37
39	14	1 1 2 2 5 5 11 11 17 17 24 24 30 31 31 40
40	41	39 41 41
41	40	40 40

2 links or more

Sentence	Median	Link set
1	12.5	2 2 2 2 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 6 6 9 9 9 9 9 11 11 12 12 13 13 13 13 16 16 16 16 16 17 17 24 24 26 26 26 28 28 28 29 29 29 30 30 30 30 30 30 31 31 31 31 31 39 39
2	14	1 1 1 1 5 5 11 11 17 17 24 24 31 31 39 39
3	2.5	1 1 4 4
4	9	1 1 1 1 1 1 1 1 1 3 3 5 5 5 9 9 9 9 9 12 12 13 13 16 16 16 29 29 29 30 30
5	9	1 1 1 1 1 1 1 1 1 2 2 4 4 4 6 6 9 9 9 11 11 12 12 16 16 17 17 17 24 24 30 30 31 31 39 39

Continued on next page

Continued from previous page

Sentence	Median	Link set
6	5	1 1 5 5 8 8
7	16	16 16
8	6	6 6
9	4	1 1 1 1 1 4 4 4 4 4 5 5 5 16 16 16
11	17	1 1 2 2 5 5 17 17 24 24 31 31 39 39
12	4	1 1 4 4 5 5
13	10	1 1 1 1 4 4 16 16 18 18 30 30
16	5	1 1 1 1 1 4 4 4 5 5 7 7 9 9 9 13 13
17	11	1 1 2 2 5 5 5 11 11 24 24 31 31 39 39
18	13	13 13
24	11	1 1 2 2 5 5 11 11 17 17 31 31 39 39
26	28	1 1 1 28 28 28 30 30
28	26	1 1 1 26 26 26 30 30 30
29	4	1 1 1 4 4 4 30 30 30
30	9	1 1 1 1 1 1 4 4 5 5 13 13 26 26 28 28 28 29 29 29
31	14	1 1 1 1 1 2 2 5 5 11 11 17 17 24 24 32 32 35 35 35 39 39
32	33	31 31 35 35
34	35	35 35
35	32	31 31 31 32 32 34 34 38 38
38	35	35 35
39	11	1 1 2 2 5 5 11 11 17 17 24 24 31 31
40	41	41 41
41	40	40 40

3 links or more

Sentence	Median	Link set
2	1	1 1 1 1
4	5	1 1 1 1 1 1 1 1 1 5 5 5 9 9 9 9 9 16 16 16 29 29 29
5	2.5	1 1 1 1 1 1 1 1 1 4 4 4 9 9 9 17 17 17
9	4	1 1 1 1 1 4 4 4 4 4 5 5 5 16 16 16
13	1	1 1 1 1
16	4	1 1 1 1 1 4 4 4 9 9 9
17	5	5 5 5
26	14.5	1 1 1 28 28 28
28	26	1 1 1 26 26 26 30 30 30
29	4	1 1 1 4 4 4 30 30 30
30	14.5	1 1 1 1 1 1 28 28 28 29 29 29

Continued on next page

Continued from previous page

Sentence	Median	Link set
31	1	1 1 1 1 1 35 35 35
35	31	31 31 31

Appendix 15

LSM performance (2 links or more)

Text	Sections	Provis. bound-aries (Peaks)	Final Bound-aries	Match-ing bound-aries	Recall		Precision	
					%	Rank	%	Rank
1	9	10	7	2	22.22	13	28.57	19
2	14	13	9	1	7.14	23	11.11	23
3	20	34	19	6	30.00	7	31.58	17
4	15	21	11	4	26.67	10	36.36	9
5	5	6	3	1	20.00	17	33.33	12
6	15	22	15	7	46.67	2	46.67	3
7	19	24	12	4	21.05	16	33.33	12
8	5	4	2	0	0.00	24	0.00	24
9	5	8	6	1	20.00	17	16.67	21
10	17	20	14	5	29.41	8	35.71	10
11	15	9	5	2	13.33	21	40.00	6
12	26	38	21	8	30.77	6	38.10	7
13	12	10	9	3	25.00	12	33.33	12
14	27	35	21	7	25.93	11	33.33	12
15	19	32	18	8	42.11	3	44.44	4
16	23	24	14	4	17.39	19	28.57	19
17	8	11	8	3	37.50	4	37.50	8
18	23	17	14	5	21.74	15	35.71	10
19	20	28	17	7	35.00	5	41.18	5
20	4	5	4	2	50.00	1	50.00	1
21	29	27	17	5	17.24	20	29.41	18
22	23	21	13	2	8.70	22	15.38	22
23	26	27	14	7	26.92	9	50.00	1
24	3	6	6	0	0.00	24	0.00	24
25	18	25	12	4	22.22	13	33.33	12

Appendix 16

LSM performance (3 links or more)

Text	Sections	Provis. bound-aries (Peaks)	Final Bound-aries	Match-ing bound-aries	Recall		Precision	
					%	Rank	%	Rank
1	9	2	2	0	0.00	15	0.00	15
2	14	3	2	0	0.00	15	0.00	15
3	20	4	4	0	0.00	15	0.00	15
4	15	4	4	1	6.67	9	25.00	9
5	5	2	2	0	0.00	15	0.00	15
6	15	3	3	1	6.67	9	33.33	6
7	19	7	4	1	5.26	12	25.00	9
8	5	0	0	0	0.00	15	.	.
9	5	4	2	1	20.00	3	50.00	2
10	17	5	4	0	0.00	15	0.00	15
11	15	2	2	0	0.00	15	0.00	15
12	26	10	9	2	7.69	6	22.22	12
13	12	1	1	0	0.00	15	0.00	15
14	27	9	7	2	7.41	8	28.57	8
15	19	7	5	2	10.53	4	40.00	4
16	23	7	5	1	4.35	13	20.00	13
17	8	8	6	2	25.00	1	33.33	6
18	23	3	2	0	0.00	15	0.00	15
19	20	4	4	0	0.00	15	0.00	15
20	4	1	1	1	25.00	1	100.0	1
21	29	6	4	1	3.45	14	25.00	9
22	23	7	5	2	8.70	5	40.00	4
23	26	6	4	2	7.69	6	50.00	2
24	3	1	1	0	0.00	15	0.00	15
25	18	6	5	1	5.56	11	20.00	13

Appendix 17

Research article corpus

Text No	Title	Source
001	Inferior Parietal Perfusion, Lateralization, and Neuropsychological Dysfunction in Alzheimer's Disease	Brain and Cognition 32, 365-383 (1996)
002	Modularity of Language Reconsidered	Brain and Language 55, 240-263 (1996)
003	Tip-of-the-Tongue States and Lexical Access in Dementia	Brain and Language 54, 196-215 (1996)
004	Sentence Context Influences the Interpretation of Word Meaning by Alzheimer Patients	Brain and Language 54, 233 - 245 (1996)
005	Acoustic and perceptual evidence for complete neutralization of manner of articulation in Korean	Journal of Phonetics (1996) 24 , 295 - 312
006	Phrase Repetition in Alzheimer's Disease: Effect of Meaning and Length	Brain and Language 54, 246 - 261 (1996)
007	Connectionist Modeling of the Recovery of Language Functions Following Brain Damage	Brain and Language 52, 7 - 24 (1996)
008	Intelligence and the Frontal Lobe: The Organization of Goal- Directed Behavior	Cognitive Psychology 30, 257 - 303 (1996)
009	The Slow Time-Course of Visual Attention	Cognitive Psychology 30, 79 - 109 (1996)
010	Constructing and Validating Motive Bridging Inferences	Cognitive Psychology 30, 1 - 38 (1996)
011	Evidence for Conjoint Retention of Information Encoded from Spatial Adjunct Displays	Contemporary Educational Psychology 21, 221 - 239 (1996)
012	Effects of Headings on Text Recall and Summarization	Contemporary Educational Psychology 21, 261 - 278 (1996)
013	The Effects of Explanations and Pictures on Learning, Retention, and Transfer of a Procedural Assembly Task	Contemporary Educational Psychology 21, 129 - 148 (1996)
014	Problems in Academic Motivation Research and Advantages and Disadvantages of Their Solutions	Contemporary Educational Psychology 21, 149 - 165 (1996)

Continued on next page

Continued from previous page

Text No	Title	Source
015	Topic Interest, Text Representation, and Quality of Experience	Contemporary Educational Psychology 21, 3 - 18 (1996)
016	A Computer Method to Model the Dose Distribution of High Energy Photon Grid Therapy in Three Dimensions	Computers and Biomedical Research 29, 247 - 258 (1996)
017	Geometric Properties of the Fractured Tibia Stabilized by Unreamed Interlocking Nail: Development of a Three-Dimensional Finite Element Model	Computers and Biomedical Research 29, 259 - 270 (1996)
018	Simulation Calculations of Cardiac Virtual Cathode Effects	Computers and Biomedical Research 29, 77 - 84 (1996)
019	Storing Sparse and Repeated Data in Multivariate Markovian Models of Tuberculosis Spread	Computers and Biomedical Research 29, 85 - 92 (1996)
020	Empire, emigration and school geography: changing discourses of Imperial citizenship, 1880 - 1925	Journal of Historical Geography, 22, 4 (1996) 373 - 387
021	Geocentric education and antiimperialism: theosophy, geography and citizenship in the writings of J. H. Cousins	Journal of Historical Geography, 22, 4 (1996) 399 - 411
022	Visual culture and geographical citizenship: England in the 1940s	Journal of Historical Geography, 22, 4 (1996) 424 - 439
023	The spatial organization of a regional economy: central places in Northwest England in the early-eighteenth century	Journal of Historical Geography, 22, 2 (1996) 147 - 159
024	Geographical practice and its significance in Peter the Great' s Russia	Journal of Historical Geography, 22, 2 (1996) 160 - 176
025	Dental microwear of European Miocene catarrhines: evidence for diets and tooth use	Journal of Human Evolution (1996) 31, 335 - 366
026	Exploitation of large bovids and seals at Middle and Later Stone Age sites in South Africa	Journal of Human Evolution (1996) 31, 315 - 334
027	Brothers in Arms: Sport, the Law and the Construction of Gender Identity	International Journal of the Sociology of Law 1996, 24, 145 - 162
028	Complete Control? Judicial and Practical Approaches to the Negotiation of Commercial Music Contracts	International Journal of the Sociology of Law 1996, 24, 89 - 115
029	Football and the Civilizing Process: Penal Discourse and the Ethic of Collective Responsibility in Sports Law	International Journal of the Sociology of Law 1996, 24, 163 - 188
030	Towards a better measure of readability: Explanation of empirical performance results	Word 40, 223-234 (1989)
031	No Soul to be Damned, No Body to be Kicked: Responsibility, Blame and Corporate Punishment	International Journal of the Sociology of Law 24, 1 - 19 (1996)
032	Law Enforcement, Justice and Democracy in the Transnational Arena: Reflections on the War on Drugs	International Journal of the Sociology of Law 24, 61 - 75 (1996)

Continued on next page

Continued from previous page

Text No	Title	Source
033	School Quality and Real House Prices: Inter- and Intra-metropolitan Effects	Journal of Housing Economics 5, 351 - 368 (1996)
034	Depreciation, Maintenance, and Housing Prices	Journal of Housing Economics 5, 369 - 389 (1996)
035	Housing Supply under Rapid Economic Growth and Varying Regulatory Stringency: An International Comparison	Journal of Housing Economics 5, 274 - 289 (1996)
036	Credit Rationing and Public Housing Loans in Japan	Journal of Housing Economics 5, 227 - 246 (1996)
037	Deposit Deregulation and the Sensitivity of Housing	Journal of Housing Economics 5, 207 - 226 (1996)
038	Detecting Discrimination: Analyzing Racial Disparities in Public Contracting	Social Science Research 25, 400 - 422 (1996)
039	What is the good of health care?	Bioethics 10, 269-291 (1996)
040	Wealth Accumulation across the Adult Life Course: Stability and Change in Sociodemographic Covariate Structures of Net Worth Data in the Survey of Income and Program Participation, 1984 - 1991	Social Science Research 25, 423 - 462 (1996)
041	Language Development in Williams Syndrome: A Case Study	Cognitive Neuropsychology 13, 1017-1040 (1996)
042	Evaluation of cognitive-behavioural counselling for the distress associated with an abnormal cervical smear result	British Journal of Health Psychology 1, 327-338 (1996)
043	Interaction in Public Reports	English for Specific Purposes 14, 189-200 (1995)
044	Theory and Practice in Content-Based ESL Reading Instruction	English for Specific Purposes 14, 223-230 (1995)
045	The Effect of Genre Awareness on Linguistic Transfer	English for Specific Purposes 14, 247-256 (1995)
046	The effects of enriched prenatal care services on Medicaid birth outcomes in New Jersey	Journal of Health Economics 15, 455-476
047	Vital exhaustion, neuroticism and symptom reporting in patients with cardiac and noncardiac chest pain	British Journal of Health Psychology 1, 4, 301-315 (1996)
048	Dieting in adolescence: An application of the theory of planned behaviour	British Journal of Health Psychology 1, 4, 315-326 (1996)
049	Influence of Salient Stimuli on Rats' Performance in an Eight-Arm Radial Maze	Learning and Motivation 27, 294-306 (1996)
050	Positive but not negative life-events predict vulnerability to upper respiratory illness	British Journal of Health Psychology 1, 4, 339-348 (1996)
051	Behavioural and mental health profiles in childhood hay fever	British Journal of Health Psychology 1, 4, 349-357 (1996)
052	Pakistani women and maternity care: raising muted voices	Sociology of Health of Illness — A journal of medical sociology 18, 1, 45-63 (1996)
053	On the status of equality	Political Theory 24, 3, 394- 400 (1996)

Continued on next page

Continued from previous page

Text No	Title	Source
054	Nothing Human is Alien to Me	Religion (1996) 26, 297 - 309
055	An ethnography of risk management amongst illicit drug injectors and its implications for the development of community-based interventions	Sociology of Health of Illness — A journal of medical sociology 18, 1, 86-106 (1996)
056	Immigration and Internal Security: Political Deportations During the McCarthy Era	Science and Society 60, 4, 393-426 (1996-1997)
057	Western Buddhism: Tradition and Modernity	Religion (1996) 26, 311 - 321
058	Two Sociological Approaches to Religion in Modern Britain	Religion (1996) 26, 331 - 342
059	Mathematical Horizons: Light and Darkness in Portugal in the 18th Century	Historia Mathematica 23 (1996), 239 - 245
060	An Example of the Secant Method of Iterative Approximation in a Fifteenth-Century Sanskrit Text	Historia Mathematica 23 (1996), 246 - 256
061	Morphological Processing and Visual Word-Recognition: Evidence from Acquired Dyslexia	Cognitive Neuropsychology 13, 1041-1058 (1996)
062	Experimenting with Embryos: Can Philosophy Help?	Bioethics 10, 292-309 (1996)
063	Albert Harry Wheeler (1873 - 1950): A Case Study in the Stratification of American Mathematical Activity	Historia Mathematica 23 (1996), 269 - 287
064	Interlabial pressure during production of bilabial phones	Journal of Phonetics (1996) 24 , 337 - 349
065	Early bilingual acquisition of the voicing contrast in English and Spanish	Journal of Phonetics (1996) 24 , 351 - 365
066	On explaining certain male-female differences in the phonetic realization of vowel categories	Journal of Phonetics (1996) 24 , 187 - 208
067	Hardy-Ramanujan's Asymptotic Formula for Partitions and the Central Limit Theorem	Advances in Mathematics 125, 114 - 120 (1997)
068	Invariants of Finite Groups over Fields of Characteristic p	Advances in Mathematics 124, 25 - 48 (1996)
069	Theory of the Anderson Impurity Model: The Schrieffer-Wolff Transformation Reexamined	Annals of Physics 252, 1 - 32 (1996)
070	Dissipation and Topologically Massive Gauge Theories in the Pseudo-Euclidean Plane	Annals of Physics 252, 115 - 132 (1996)
071	Gravitational Wave Interaction with Normal and Superconducting Circuits	Annals of Physics 248, 34 - 59 (1996)
072	Are Anomalously Short Tunnelling Times Measurable?	Annals of Physics 248, 122 - 133 (1996)
073	Classical and Quantum Transitions to Chaos for a Family of Periodically Driven Hamiltonians	Annals of Physics 246, 369 - 380 (1996)

Continued on next page

Continued from previous page

Text No	Title	Source
074	Turbulent Two-Dimensional Magnetohydrodynamics and Conformal Field Theory	Annals of Physics 246, 446 - 458 (1996)
075	The Nonconfining Schwinger Model	Annals of Physics 249, 34 - 43 (1996)
076	Cytokines in the serum and brain in mice infected with distinct species of Lyme disease Borrelia	Microbial Pathogenesis 21, 413 - 419 (1996)
077	A possible mechanism for host-specific pathogenesis of Salmonellaserovars	Microbial Pathogenesis 21, 435 - 446 (1996)
078	Natural Abundance Isotopic Fractionation in the Fermentation Reaction: Influence of the Nature of the Yeast	Bioorganic Chemistry 24, 319 - 330 (1996)
079	Comparison of Resorufin Acetate and pNitrophenyl Acetate as Substrates for Chymotrypsin	Bioorganic Chemistry 24, 331 - 339 (1996)
080	Platelet-Activating Factor and Nitric Oxide Mediate Microvascular Permeability in Ischemia-Reperfusion Injury	Microvascular Research 52, 210 - 220 (1996)
081	Functional Microcirculatory Impairment: A Possible Source of Reduced Skin Oxygen Tension in Human Diabetes Mellitus	Microvascular Research 52, 115 - 126 (1996)
082	Simultaneous Analysis of Peripheral Blood Granulocytes, Lymphocytes, and Monocytes Adhering to Human Microvascular Endothelial Cells	Microvascular Research 52, 101 - 114 (1996)
083	Lotka's Game in Predator-Prey Theory: Linking Populations to Individuals	Theoretical Population Biology 50, 368-393 (1996)
084	The Determinants of Young Women's Wages: Comparing the Effects of Individual and Occupational Labor Market Characteristics	Science Research 25, 240-259 (1996)
085	The Stability and Persistence of Mutualisms Embedded in Community Interactions	Theoretical Population Biology 50, 281 - 297 (1996)
086	Lightning injury: A review and case presentations	The Canadian Journal of Plastic Surgery 2, 4 (1994)
087	Interdisciplinary Collaboration in Teacher Education: A Constructivist Approach	TESOL Quarterly 30, 231 - 252 (1996)
088	ESDA and the analysis of contested contemporaneous notes of police interviews	Forensic Linguistics 1, 71 - 90 (1994)
089	Buttocks lift for tight thighs	The Canadian Journal of Plastic Surgery 4, 1, (1996)
090	Corpus Work at HCRC	International Journal of Corpus Linguistics 1, 121 - 130 (1996)
091	The Empty Lexicon	International Journal of Corpus Linguistics 1, 99 - 120 (1996)
092	Contextual Dependency and Lexical Sets	International Journal of Corpus Linguistics 1, 75 - 98 (1996)

Continued on next page

Continued from previous page

Text No	Title	Source
093	Analysis of Temporal Changes in Corpora	International Journal of Corpus Linguistics 1, 61 - 74 (1996)
094	The Role of Corpora in Compiling the Cambridge International Dictionary of English	International Journal of Corpus Linguistics 1, 39 - 60 (1996)
095	Evolution of Structure, Phase Composition, and X-Ray Reflectivity of Multilayer Mirrors Mo - (B / C) after Annealing at 250 - 11007C	Journal of X-Ray Science and Techonology 6, 141 - 149 (1996)
096	Connecting Current Research on Authentic and Performance Assessment Through Portfolios	Assessing Writing 1, 247 - 266 (1994)
097	Time Reversal Focusing Applied to Lithotripsy	Ultrasonic Imaging 18, 106 - 121 (1996)
098	The evaluation of waste management options	Waste Management & Research (1996) 14, 515 - 526
099	Phase Insensitive Homomorphic Image Processing for Speckle Reduction	Ultrasonic Imaging 18, 122 - 139 (1996)
100	Bioaerosol exposure during collection of mixed domestic waste - An intervention study on compactor truck design	Waste Management & Research (1996) 14, 527 - 536

Appendix 18

Business report corpus

Text No	Company
001	IDS/Balcor Income Partners
002	IDS Certificate Company
003	Hardin Bancorp, Inc
004	Harcourt General, Inc.
005	Harken Energy Corporation
006	Harrow Corporation
007	Harry's Farmers Market, Inc.
008	Hartford Life Insurance Company
009	Nantucket Industries, Inc.
010	Harsco Corporation
011	Harte-Hanks Corporation
012	Hartford Steam Boiler Inspection and Insurance Company
013	Hartmarx Corporation
014	Harvard Industries, Inc.
015	ICG Communications, Inc.
016	ICN Merger Corp.
017	ICO Corporation
018	Idaho Power Company
019	Identix Incorporated
020	Four M Corporation
021	IDEX Corporation
022	IDS Life Account
023	IDS Managed Futures
024	IDS/Shurgard Income Growth Partners
025	IEC Electronics Corp.
026	IES Industries Inc.
027	IES Utilities Inc.
028	IFR Systems, Inc.
029	IGEN Corporation
030	IGI, Inc.
031	LabOne, Inc.
032	Leggett & Platt Inc.
033	Legg Mason, Inc.
034	Lance, Inc.
035	National Housing Partnership Realty Fund
036	National Computer Systems, Inc.

Continued on next page

Continued from previous page

Text No	Company
037	National Data Corporation
038	National Diversified Services, Inc.
039	National Fuel Gas Company
040	National Commerce Bancorporation
041	National Gas & Oil Company
042	National Home Health Care Corp.
043	National Income Realty Trust
044	National Micronetics, Inc.
045	NAC Re Corp.
046	National Mortgage Acceptance Corporation
047	National Properties Corp.
048	Badger Meter, Inc.
049	N.U. Pizza Holding Corporation
050	Goodyear Tire & Rubber Company
051	Halifax Corporation
052	Ideon Group, Inc.
053	Lands' End, Inc.
054	Lahaina Acquisitions, Inc.
055	Lakeland Industries, Inc.
056	Lamcor Incorporated
057	Lamson & Sessions Co.
058	Lancit Media Productions, Ltd.
059	Lancaster Colony Corporation
060	Landauer, Inc.
061	National Auto Credit, Inc.
062	Lane Plywood, Inc.
063	Larcan-TTC Inc.
064	Larizza Industries, Inc.
065	Larson Davis Inc.
066	LaserMaster Technologies, Inc.
067	Laser Photonics, Inc.
068	LBO Capital Corp.
069	Landmark Graphics Corporation
070	Nalco Chemical Company
071	Pacific Bell
072	Pacific Real Estate Investment Trust
073	PaineWebber R&D Partners
074	Paris Business Forms, Inc.
075	Parlex Corporation
076	PC Quote, Inc.
077	Penn Engineering & Manufacturing Corp
078	P & F Industries, Inc.
079	Radiant Technology Corp.
080	RADVA Corporation
081	RAL Income + Equity Growth V Limited Partnership
082	Lee Enterprises Inc.
083	R.F. Management Corp.
084	Saddlebrook Resorts, Inc.
085	Sage Laboratories, Inc.
086	Sanchez-O'Brien Drilling Company
087	SB Partners

Continued on next page

Continued from previous page

Text No	Company
088	SBM Certificate Company
089	Scan-Optics, Inc.
090	S/M Real Estate Fund VII, Ltd.
091	S&T Bancorp, Inc.
092	Tab Products Co.
093	Tandy Brands Accessories, Inc.
094	Taurus Petroleum, Inc.
095	TCI International, Inc.
096	LecTec Corporation
097	Tech/Ops Sevcon, Inc.
098	LeCroy Corporation
099	Vacu-dry Company
100	Laserscope-Registered Trademark

Appendix 19

Encyclopedia article corpus

Text No	Title
001	Academy
002	Act of Union
003	Acupuncture
004	Adaptation
005	Adoption
006	Adult Education
007	Adventists
008	Aesthetics
009	Aegean Civilization
010	Aggression
011	Air Pollution
012	Air Warfare
013	Alcohol
014	Algae
015	Analytic and Linguistic Philosophy
016	Anti-Semitism
017	Arabs
018	Archaeology
019	Audiovisual Education
020	Babylon
021	Bacteria
022	Biblical Archaeology
023	African immigration to the Americas
024	Boer War
025	Buddhism
026	Calendar
027	Caliphate
028	Church of England
029	City Planning
030	Code
031	Confucianism
032	Contract
033	Cooperatives
034	Cost of Living
035	Crime Detection
036	Descartes

Continued on next page

Continued from previous page

Text No	Title
037	Dinosaur
038	Election
039	Energy Supply
040	Evolution
041	Festivals and Feasts
042	Flood Control
043	Forum
044	God
045	Gothic Art and Architecture
046	Hammurabi
047	Hanseatic League
048	Health Insurance
049	Heat
050	Heating
051	House
052	Housing
053	Ice Skating
054	Indian Music
055	Machine Tools
056	Madrigal
057	Mental Disorders
058	Mongol Empire
059	Mythology
060	Native American Languages
061	Naval Vessels
062	Novel
063	Orthodox Church
064	Pest Control
065	Picasso
066	Plato
067	Positivism
068	Preschool Education
069	Protestantism
070	Psychical Research
071	Friends, Society of
072	Radio
073	Railroads
074	Rhetoric
075	Rome, History of
076	Rose
077	Shiites
078	Short Story
079	Soccer
080	Space Exploration
081	Temple (building)
082	Thanatology
083	Tiberius
084	Unitarianism
085	Zen or Chan
086	Moldova
087	Monaco

Continued on next page

Continued from previous page

Text No	Title
088	French Guiana
089	Northern Ireland
090	Suriname
091	Poland
092	Portugal
093	Romania
094	Djibouti
095	San Marino
096	Scotland
097	Serbia
098	Slovakia
099	Slovenia
100	Ethiopia

Appendix 20

Predicted recall and precision

The following tables provide the predicted recall and precision values for each corpus.

- In tables referring to predicted **recall** values, *residual* refers to the absolute value of the difference between recall and predicted recall;
- In tables referring to predicted **precision** values, *residual* refers to the absolute value of the difference between precision and predicted precision;
- Texts are sorted according to values of residual;
- Predicted values are not available for 'outliers'.

Predicted recall values for research article corpus

Text	Recall	Predicted Recall	Residual	Sections	Avg. Median Diff.	Tokens/ Sent.	Tokens/ Section
28	33.33	33.4856	0.15562	6	26.3	17.8089	932.00
94	35.29	34.8809	0.40906	17	20.4	14.6541	229.29
25	28.13	27.5006	0.62936	32	25.5	16.8676	179.22
34	38.46	37.7305	0.72952	13	19.9	15.0120	289.85
53	33.33	34.3355	1.00546	3	6.1	16.3896	420.67
60	40.00	38.3472	1.65278	5	9.8	15.1290	281.40
77	45.45	47.2096	1.75960	11	7.9	22.5111	184.18
67	28.57	30.9589	2.38894	7	6.7	10.1509	76.86
86	41.67	44.6001	2.93014	12	27.2	11.4118	145.50
21	25.00	28.1156	3.11561	4	10.7	18.7321	786.75
29	25.00	21.8378	3.16215	4	38.5	17.8920	1614.75
10	29.27	32.4412	3.17116	41	51.5	12.5952	203.37
31	33.33	36.5638	3.23384	6	15.9	18.7044	632.83
13	25.93	29.4466	3.51665	27	23.5	15.4083	164.93
46	33.33	36.8641	3.53411	9	8.1	18.7296	330.89
55	40.00	36.4291	3.57090	10	21.7	15.1058	471.30
20	25.00	28.8727	3.87270	4	13.6	20.8516	948.75
32	25.00	28.9257	3.92571	4	15.3	17.0749	798.25
2	40.00	35.9655	4.03448	5	34.3	15.9905	1010.60
85	44.44	40.3421	4.09788	9	15.0	14.7939	215.33
45	40.00	35.7848	4.21518	5	4.1	17.7684	337.60
30	33.33	37.5733	4.24328	3	12.0	16.9121	513.00
76	44.44	39.2986	5.14140	9	8.0	16.2597	139.11
100	38.89	33.6293	5.26075	18	12.3	16.7500	130.28
49	28.57	33.8853	5.31528	14	13.3	15.0113	189.79
1	37.50	42.8565	5.35654	8	16.4	18.7598	419.75

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sections	Avg. Median Diff.	Tokens/ Sent.	Tokens/ Section
41	42.86	37.0662	5.79378	7	14.4	16.2732	425.43
15	42.86	36.9892	5.87082	14	20.4	13.4612	210.57
70	40.00	33.7038	6.29617	5	14.6	16.4778	593.20
24	0.00	6.3127	6.3127	3	23.2	15.6263	1547.00
88	25.00	31.4977	6.4977	4	36.4	13.9434	1108.50
38	43.48	36.8449	6.6351	23	21.7	18.0042	187.87
44	42.86	35.9837	6.8763	7	4.4	17.1800	245.43
43	46.15	37.8315	8.3185	13	17.2	15.0681	221.38
37	50.00	41.4818	8.5182	10	10.6	18.3852	224.30
52	50.00	41.0122	8.9878	8	26.8	15.5939	571.13
58	25.00	34.9710	9.9710	4	15.9	20.4815	829.50
91	33.33	43.7980	10.4680	9	26.7	15.1489	441.00
68	41.67	31.1728	10.4972	24	31.7	9.4401	121.54
78	25.00	35.5636	10.5636	12	7.8	16.5780	150.58
19	50.00	39.3559	10.6441	6	6.6	17.0357	238.50
40	45.45	34.6298	10.8202	11	12.2	21.2188	555.55
5	23.53	35.0920	11.5620	17	18.8	14.1684	158.35
81	50.00	38.2828	11.7172	14	9.8	18.3136	154.36
6	50.00	37.9385	12.0615	10	15.3	14.6982	248.40
17	25.00	37.3108	12.3108	4	8.5	17.8500	446.25
99	55.56	42.7451	12.8149	9	25.5	13.4426	351.00
36	25.00	38.0195	13.0195	8	21.7	14.6148	469.50
23	0.00	13.4964	13.4964	2	7.1	19.5124	1180.50
93	55.56	41.8499	13.7101	9	13.0	16.0472	189.00
7	50.00	36.2804	13.7196	8	23.2	14.1171	527.63
80	23.08	36.9752	13.8952	13	15.8	14.0621	156.85
26	57.14	42.7447	14.3953	7	24.7	18.3252	644.00

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sections	Avg. Median Diff.	Tokens/ Sent.	Tokens/ Section
98	55.56	40.8126	14.7474	9	15.3	15.3429	238.67
3	22.22	37.6117	15.3917	9	14.7	15.1808	298.56
8	24.44	40.4897	16.0497	45	64.1	13.4746	229.67
39	62.50	46.2733	16.2267	8	31.3	15.9225	565.25
54	0.00	16.9379	16.9379	3	25.3	18.3640	1463.00
79	25.00	42.9587	17.9587	4	4.4	20.1613	312.50
27	60.00	41.9418	18.0582	5	16.3	23.5422	781.60
90	22.22	40.6191	18.3991	9	9.3	16.9121	171.00
89	16.67	35.6128	18.9428	6	3.3	13.5000	72.00
42	58.33	39.2790	19.0510	12	15.7	16.1839	234.67
57	57.14	37.6731	19.4669	7	17.5	14.8418	415.57
50	60.00	40.2723	19.7277	10	13.8	16.7219	252.50
75	60.00	40.2518	19.7482	5	22.7	10.7834	338.60
14	20.00	39.7939	19.7939	15	21.9	15.7952	262.20
83	15.38	35.2065	19.8265	13	16.1	15.7213	295.08
73	57.14	37.1805	19.9595	7	14.0	13.1630	253.86
11	45.45	24.7745	20.6755	22	11.6	15.8809	169.64
84	55.56	34.6217	20.9383	9	10.6	18.3756	434.89
22	50.00	28.1477	21.8523	4	23.5	17.8577	1067.00
82	14.29	36.2149	21.9249	14	7.5	17.7368	120.36
61	14.29	37.3176	23.0276	7	15.6	15.3060	400.14
4	50.00	26.9353	23.0647	4	15.8	15.7970	797.75
18	50.00	26.8542	23.1458	2	10.9	14.6022	679.00
63	50.00	26.4405	23.5595	4	8.3	19.5843	812.75
96	14.29	37.8754	23.5854	7	19.4	17.7071	604.57
97	12.50	36.1409	23.6409	8	20.0	12.7824	381.88
62	50.00	25.9566	24.0434	4	17.6	16.8858	924.50

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sections	Avg. Median Diff.	Tokens/ Sent.	Tokens/ Section
33	14.29	38.7477	24.4577	7	10.4	18.2534	380.71
71	60.00	35.2069	24.7931	5	21.8	13.1991	570.20
64	7.14	32.5627	25.4227	14	9.0	16.5062	191.00
87	66.67	41.1470	25.5230	9	22.5	16.6255	478.44
69	18.18	43.9278	25.7478	11	38.5	11.1317	468.55
95	0.00	25.8095	25.8095	2	3.7	15.8378	586.00
56	0.00	25.8118	25.8118	5	27.0	16.1147	1095.80
12	52.94	26.9521	25.9879	17	11.3	15.5147	248.24
92	20.00	46.7012	26.7012	10	31.1	13.8205	377.30
66	12.50	39.3743	26.8743	8	20.4	16.9020	517.63
35	60.00	32.4750	27.5250	5	13.1	14.8941	506.40
9	66.67	37.3565	29.3135	21	32.7	14.5202	341.57
65	0.00	31.5991	31.5991	6	9.0	17.3464	517.50
74	0.00	OUTLIER	.	3	14.7	11.8411	422.33
72	0.00	34.4631	34.4631	5	11.1	15.0803	413.20
51	0.00	OUTLIER	.	4	8.9	15.0519	289.75
48	80.00	OUTLIER	.	5	15.2	15.2670	583.20
47	100.00	OUTLIER	.	3	8.1	14.9844	319.67
16	100.00	OUTLIER	.	3	7.9	15.3483	455.33
59	100.00	OUTLIER	.	1	12.8	14.5243	1496.00

Predicted precision values for research article corpus

Text	Precision	Predicted Precision	Residual	Sections	Sents.	Avg. Median Diff.	Links/ Sent.	Tokens/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
82	15.38	15.3746	0.0054	14	95	7.5	43.105	17.7368	6.92632	292.50	120.36	47.000
52	7.55	7.5261	0.0239	8	293	26.8	67.625	15.5939	4.25597	2476.75	571.13	155.875
70	6.06	6.0897	0.0297	5	180	14.6	53.833	16.4778	3.82222	1938.00	593.20	137.600
53	10.00	9.9131	0.0869	3	77	6.1	30.130	16.3896	5.88312	773.33	420.67	151.000
89	14.29	14.1207	0.1693	6	32	3.3	15.750	13.5000	5.53125	84.00	72.00	29.500
18	5.00	4.7371	0.2629	2	93	10.9	26.108	14.6022	4.62366	1214.00	679.00	215.000
6	13.16	13.4623	0.3023	10	169	15.3	53.018	14.6982	3.86982	896.00	248.40	65.400
56	0.00	0.3290	0.3290	5	340	27.0	63.385	16.1147	4.77059	4310.20	1095.80	324.400
69	2.33	1.9320	0.3980	11	463	38.5	66.652	11.1317	1.62203	2805.45	468.55	68.273
26	9.09	8.6914	0.3986	7	246	24.7	58.809	18.3252	5.41463	2066.71	644.00	190.286
90	14.29	13.8884	0.4016	9	91	9.3	18.253	16.9121	7.43956	184.56	171.00	75.222
94	12.24	12.6980	0.4580	17	266	20.4	64.718	14.6541	3.68045	1012.65	229.29	57.588
86	18.52	18.0473	0.4727	12	153	27.2	13.752	11.4118	4.52941	175.33	145.50	57.750
4	5.71	6.2439	0.5339	4	202	15.8	89.777	15.7970	3.43564	4533.75	797.75	173.500
46	12.00	11.1846	0.8154	9	159	8.1	87.969	18.7296	4.93082	1554.11	330.89	87.111
41	10.34	9.4571	0.8829	7	183	14.4	57.191	16.2732	4.78142	1495.14	425.43	125.000
10	11.65	10.7662	0.8838	41	662	51.5	111.343	12.5952	1.85045	1797.78	203.37	29.878
39	10.00	9.0911	0.9089	8	284	31.3	59.838	15.9225	3.95775	2124.25	565.25	140.500
59	6.25	5.3386	0.9114	1	103	12.8	13.573	14.5243	6.78641	1398.00	1496.00	699.000
67	15.38	14.2189	1.1611	7	53	6.7	7.811	10.1509	3.94340	59.14	76.86	29.857
34	12.20	10.9072	1.2928	13	251	19.9	89.263	15.0120	3.39841	1723.46	289.85	65.615
28	3.85	5.4420	1.5920	6	314	26.3	86.462	17.8089	4.80892	4524.83	932.00	251.667
54	0.00	1.6332	1.6332	3	239	25.3	27.372	18.3640	7.60669	2180.67	1463.00	606.000
24	0.00	-1.6401	1.6401	3	297	23.2	54.710	15.6263	5.09428	5416.33	1547.00	504.333
15	16.22	14.4843	1.7357	14	219	20.4	57.457	13.4612	3.01370	898.79	210.57	47.143
13	13.46	15.2097	1.7497	27	289	23.5	128.415	15.4083	2.58478	1374.52	164.93	27.667

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sections	Sents.	Avg. Median Diff.	Links/ Sent.	Tokens/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
79	11.11	12.8836	1.7736	4	62	4.4	38.887	20.1613	7.74194	602.75	312.50	120.000
49	13.79	15.5668	1.7768	14	177	13.3	74.938	15.0113	2.89266	947.43	189.79	36.571
63	8.00	6.2224	1.7776	4	166	8.3	55.470	19.5843	7.51807	2302.00	812.75	312.000
57	10.81	9.0321	1.7779	7	196	17.5	41.388	14.8418	4.52041	1158.86	415.57	126.571
32	3.70	5.5089	1.8089	4	187	15.3	46.128	17.0749	6.02139	2156.50	798.25	281.500
27	12.00	10.1892	1.8108	5	166	16.3	46.675	23.5422	9.33133	1549.60	781.60	309.800
55	8.33	6.4763	1.8537	10	312	21.7	87.654	15.1058	3.72115	2734.80	471.30	116.100
73	14.81	12.9511	1.8589	7	135	14.0	29.822	13.1630	3.67407	575.14	253.86	70.857
85	13.33	15.2412	1.9112	9	131	15.0	37.962	14.7939	4.61069	552.56	215.33	67.111
7	6.35	4.4246	1.9254	8	299	23.2	57.214	14.1171	3.40134	2138.38	527.63	127.125
8	9.17	11.1173	1.9473	45	767	64.1	133.751	13.4746	1.98044	2279.71	229.67	33.756
68	18.18	16.1800	2.0000	24	309	31.7	70.272	9.4401	1.17152	904.75	121.54	15.083
99	9.43	11.5383	2.1083	9	235	25.5	54.013	13.4426	2.91915	1410.33	351.00	76.222
35	9.09	6.9530	2.1370	5	170	13.1	51.065	14.8941	4.00000	1736.20	506.40	136.000
23	0.00	2.1844	2.1844	2	121	7.1	56.058	19.5124	6.39669	3391.50	1180.50	387.000
91	7.69	9.8810	2.1910	9	262	26.7	49.126	15.1489	4.85115	1430.11	441.00	141.222
40	9.80	7.4679	2.3321	11	288	12.2	177.191	21.2188	3.97222	4639.18	555.55	104.000
78	18.75	16.3224	2.4276	12	109	7.8	37.771	16.5780	5.54128	343.08	150.58	50.333
31	5.13	7.7113	2.5813	6	203	15.9	57.906	18.7044	6.07389	1959.17	632.83	205.500
22	5.00	2.3674	2.6326	4	239	23.5	38.757	17.8577	6.37657	2315.75	1067.00	381.000
38	17.54	20.2331	2.6931	23	240	21.7	93.450	18.0042	3.74167	975.13	187.87	39.043
20	2.94	5.7901	2.8501	4	182	13.6	58.022	20.8516	7.26923	2640.00	948.75	330.750
36	5.88	8.7807	2.9007	8	257	21.7	94.973	14.6148	2.70039	3051.00	469.50	86.750
21	4.35	7.2602	2.9102	4	168	10.7	66.726	18.7321	6.50000	2802.50	786.75	273.000
75	13.64	10.6243	3.0157	5	157	22.7	24.293	10.7834	2.84713	762.80	338.60	89.400
2	3.08	0.0214	3.0586	5	316	34.3	50.323	15.9905	4.35127	3180.40	1010.60	275.000
60	10.00	13.0994	3.0994	5	93	9.8	29.301	15.1290	4.88172	545.00	281.40	90.800

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sections	Sents.	Avg. Median Diff.	Links/ Sent.	Tokens/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
71	8.33	5.1790	3.1510	5	216	21.8	40.394	13.1991	3.52315	1745.00	570.20	152.200
45	14.29	11.1260	3.1640	5	95	4.1	73.516	17.7684	4.91579	1396.80	337.60	93.400
62	7.14	3.7262	3.4138	4	219	17.6	63.215	16.8858	4.96804	3461.00	924.50	272.000
1	9.38	12.9564	3.5764	8	179	16.4	69.028	18.7598	4.86034	1544.50	419.75	108.750
58	2.94	6.8822	3.9422	4	162	15.9	37.253	20.4815	7.41975	1508.75	829.50	300.500
3	7.69	11.6697	3.9797	9	177	14.7	74.559	15.1808	3.67797	1466.33	298.56	72.333
43	18.18	14.1062	4.0738	13	191	17.2	38.304	15.0681	4.87958	562.77	221.38	71.692
25	14.75	18.9544	4.2044	32	340	25.5	82.029	16.8676	3.80588	871.56	179.22	40.438
80	13.04	17.5567	4.5167	13	145	15.8	38.993	14.0621	3.97241	434.92	156.85	44.308
30	5.88	10.4552	4.5752	3	91	12.0	24.407	16.9121	6.08791	740.33	513.00	184.667
48	10.81	5.9162	4.8938	5	191	15.2	62.545	15.2670	3.45550	2389.20	583.20	132.000
19	21.43	16.4527	4.9773	6	84	6.6	39.155	17.0357	4.52381	548.17	238.50	63.333
5	13.33	18.3461	5.0161	17	190	18.8	57.568	14.1684	3.15263	643.41	158.35	35.235
93	22.73	17.6484	5.0816	9	106	13.0	31.113	16.0472	5.13208	366.44	189.00	60.444
88	1.64	-3.4689	5.1089	4	318	36.4	46.730	13.9434	3.52201	3715.00	1108.50	280.000
97	2.70	7.8372	5.1372	8	239	20.0	58.845	12.7824	2.79498	1758.00	381.88	83.500
29	1.52	-3.6485	5.1685	4	361	38.5	45.133	17.8920	6.10526	4073.25	1614.75	551.000
84	13.89	8.6318	5.2582	9	213	10.6	106.606	18.3756	4.21127	2523.00	434.89	99.667
87	15.00	9.4924	5.5076	9	259	22.5	80.757	16.6255	4.84556	2324.00	478.44	139.444
37	22.73	17.2079	5.5221	10	122	10.6	60.852	18.3852	4.96721	742.40	224.30	60.600
17	5.26	10.9179	5.6579	4	100	8.5	41.030	17.8500	5.77000	1025.75	446.25	144.250
64	5.56	11.2197	5.6597	14	162	9.0	113.975	16.5062	3.77778	1318.86	191.00	43.714
83	4.55	10.7532	6.2032	13	244	16.1	70.795	15.7213	3.75000	1328.77	295.08	70.385
95	0.00	6.3197	6.3197	2	74	3.7	42.811	15.8378	4.60811	1584.00	586.00	170.500
96	2.33	8.7702	6.4402	7	239	19.4	86.146	17.7071	5.10460	2941.29	604.57	174.286
81	26.92	20.4426	6.4774	14	118	9.8	53.178	18.3136	5.00847	448.21	154.36	42.214
50	21.43	14.9083	6.5217	10	151	13.8	49.139	16.7219	4.92053	742.00	252.50	74.300

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sections	Sents.	Avg. Median Diff.	Links/ Sent.	Tokens/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
16	15.79	9.1830	6.6070	3	89	7.9	36.292	15.3483	5.32584	1076.67	455.33	158.000
61	3.23	9.8488	6.6188	7	183	15.6	64.989	15.3060	4.25683	1699.00	400.14	111.286
66	2.27	9.2020	6.9320	8	245	20.4	90.620	16.9020	3.70204	2775.25	517.63	113.375
100	28.00	21.0592	6.9408	18	140	12.3	54.950	16.7500	4.35000	427.39	130.28	33.833
14	7.14	14.5488	7.4088	15	249	21.9	50.839	15.7952	4.17269	843.93	262.20	69.267
92	3.57	11.2010	7.6310	10	273	31.1	32.392	13.8205	4.19780	884.30	377.30	114.600
74	0.00	8.2191	8.2191	3	107	14.7	18.963	11.8411	3.53271	676.33	422.33	126.000
33	3.85	12.1390	8.2890	7	146	10.4	75.712	18.2534	4.54110	1579.14	380.71	94.714
9	14.14	5.7643	8.3757	21	494	32.7	128.536	14.5202	2.06073	3023.67	341.57	48.476
65	0.00	8.4837	8.4837	6	179	9.0	116.844	17.3464	3.26816	3485.83	517.50	97.500
42	24.14	15.2615	8.8785	12	174	15.7	63.443	16.1839	4.12644	919.92	234.67	59.833
98	23.81	14.7694	9.0406	9	140	15.3	49.293	15.3429	4.50000	766.78	238.67	70.000
72	0.00	9.2773	9.2773	5	137	11.1	47.934	15.0803	4.46715	1313.40	413.20	122.400
51	0.00	11.2980	11.2980	4	77	8.9	21.675	15.0519	6.27273	417.25	289.75	120.750
47	25.00	11.9229	13.0771	3	64	8.1	22.250	14.9844	5.48438	474.67	319.67	117.000
44	27.27	10.8183	16.4517	7	100	4.4	97.290	17.1800	4.48000	1389.86	245.43	64.000
12	23.68	OUTLIER	.	17	272	11.3	141.960	15.5147	2.71324	2271.35	248.24	43.412
11	30.30	OUTLIER	.	22	235	11.6	129.685	15.8809	2.58298	1385.27	169.64	27.591
77	50.00	OUTLIER	.	11	90	7.9	56.211	22.5111	6.97778	459.91	184.18	57.091
76	57.14	OUTLIER	.	9	77	8.0	31.234	16.2597	5.83117	267.22	139.11	49.889

Predicted recall values for business report corpus

Text	Recall	Predicted Recall	Residual	Sents.	Bound-aries	Tokens/ Sent.	Types/ Sent.
11	30.00	29.9347	0.0653	240	81	17.0500	3.82083
29	33.33	33.2016	0.1284	352	124	18.2642	3.57386
40	34.38	34.1911	0.1889	271	97	16.2288	3.31734
95	33.33	33.9408	0.6108	152	59	17.7763	5.73684
43	18.75	17.6641	1.0859	115	21	19.4348	5.22609
65	29.63	28.4766	1.1534	150	54	16.7200	5.69333
16	28.21	26.8898	1.3202	346	118	17.0665	3.79769
97	27.27	25.8906	1.3794	36	14	14.0000	5.94444
73	28.57	29.9859	1.4159	48	18	18.0208	6.56250
30	28.00	29.4361	1.4361	223	75	18.0897	4.51570
3	44.19	45.6648	1.4748	562	215	17.0391	2.29359
14	34.21	35.7103	1.5003	441	159	20.2857	4.04989
21	25.00	26.5198	1.5198	104	31	18.0962	5.58654
27	30.77	32.4893	1.7193	222	80	16.2793	4.15315
52	34.48	36.4236	1.9436	278	101	17.3957	3.56835
47	27.27	24.9947	2.2753	46	14	14.7609	5.65217
79	19.05	21.4090	2.3590	102	28	15.5098	5.22549
2	33.33	30.9657	2.3643	83	22	18.2410	4.32530
9	18.18	15.6572	2.5228	203	54	16.3892	4.23645
12	23.81	26.4982	2.6882	214	69	16.9065	4.15888
19	28.85	26.1107	2.7393	421	140	19.4181	3.53919
82	22.22	24.9852	2.7652	53	17	18.2075	7.24528
26	29.41	32.3411	2.9311	267	96	16.3258	3.94007
33	31.25	28.2906	2.9594	272	88	18.8382	3.85294
6	22.22	18.8185	3.4015	46	9	21.8261	9.19565
74	27.27	31.1140	3.8440	121	44	15.9008	4.91736

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sents.	Bound-aries	Tokens/ Sent.	Types/ Sent.
55	20.00	23.8541	3.8541	222	70	16.5586	4.20721
25	26.32	30.2159	3.8959	147	51	17.1293	5.02721
5	47.37	43.3168	4.0532	163	65	21.1104	5.26380
50	28.21	32.3250	4.1150	313	105	20.7796	3.99361
57	36.84	32.4703	4.3697	74	34	18.1622	7.35135
44	30.00	25.5168	4.4832	69	24	15.1739	5.97101
39	35.00	40.2762	5.2762	167	63	20.5988	5.01198
83	28.57	23.1057	5.4643	175	52	16.0629	4.17143
15	36.36	30.8560	5.5040	393	132	20.1603	3.28499
59	38.46	32.9292	5.5308	73	31	16.7397	6.02740
63	38.89	33.1639	5.7261	144	54	15.2569	4.38889
61	32.26	26.2513	6.0087	171	56	15.5263	4.38012
22	16.67	22.7046	6.0346	30	4	20.9000	8.06667
67	25.00	18.5830	6.4170	256	79	16.1094	4.43359
100	37.50	30.1876	7.3124	305	107	16.7082	3.85902
64	16.67	24.1085	7.4385	119	35	17.0588	5.34454
66	39.29	31.6486	7.6414	264	92	18.0682	4.24621
53	18.75	26.6215	7.8715	127	44	15.4882	5.34646
94	35.29	27.3153	7.9747	58	21	15.4138	5.89655
78	33.33	25.0989	8.2311	68	21	14.4559	5.17647
87	22.22	30.4591	8.2391	55	21	20.1091	7.43636
13	23.08	31.5660	8.4860	116	42	17.6207	5.54310
96	25.00	16.4637	8.5363	156	39	15.9679	4.51282
80	22.22	30.7792	8.5592	74	26	16.7973	5.39189
86	35.71	26.8370	8.8730	56	18	14.1964	4.98214
49	32.00	40.8813	8.8813	240	96	15.0167	3.52917
60	18.18	27.0829	8.9029	70	26	14.5000	5.64286

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sents.	Bound-aries	Tokens/ Sent.	Types/ Sent.
68	33.33	24.3250	9.0050	35	13	15.1714	6.71429
45	43.48	34.1789	9.3011	222	78	17.6126	3.87387
77	14.29	23.8517	9.5617	46	13	16.7391	6.58696
72	25.71	35.2738	9.5638	316	117	16.2532	3.67722
1	33.33	23.6990	9.6310	25	11	15.3600	7.36000
8	45.00	35.3504	9.6496	172	64	15.9070	3.97093
56	16.67	26.5206	9.8506	38	12	12.2368	4.39474
28	16.67	26.5755	9.9055	66	21	17.9091	6.50000
88	33.33	22.8595	10.4705	41	10	16.3902	6.43902
17	25.00	35.6220	10.6220	140	54	17.5714	5.13571
54	33.33	22.2098	11.1202	118	30	18.1864	5.28814
42	41.67	30.4324	11.2376	104	32	18.9135	5.19231
98	25.00	13.3001	11.6999	250	64	18.5560	4.00400
69	20.99	32.8700	11.8800	283	100	18.1131	4.12367
38	37.50	24.6581	12.8419	76	20	16.9605	5.48684
58	20.00	33.1438	13.1438	139	50	19.1223	5.63309
10	44.44	31.2845	13.1555	141	51	18.4610	5.85816
18	14.81	28.0295	13.2195	366	125	18.1530	3.84699
4	15.79	29.5789	13.7889	86	28	16.7209	5.08140
85	38.46	24.1868	14.2732	52	18	15.8462	6.71154
70	40.00	25.2946	14.7054	93	31	15.8495	5.83871
75	7.69	22.4469	14.7569	85	26	13.9529	5.28235
41	14.29	29.5161	15.2261	109	36	16.0092	4.60550
89	22.22	37.6147	15.3947	152	62	14.6711	4.11842
37	16.00	31.7423	15.7423	137	43	19.2117	4.63504
20	36.84	20.6132	16.2268	178	45	19.3315	4.42135
36	45.00	28.5094	16.4906	112	35	17.5804	5.07143

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Sents.	Bound-aries	Tokens/ Sent.	Types/ Sent.
48	14.29	30.8460	16.5560	86	33	13.6744	4.53488
93	16.67	34.8372	18.1672	90	37	15.4667	4.87778
99	50.00	31.4354	18.5646	66	28	13.1061	4.72727
31	0.00	18.8956	18.8956	148	40	15.9257	4.79054
51	47.06	28.1616	18.8984	127	44	16.7717	5.52756
24	0.00	19.7570	19.7570	55	13	15.5273	6.30909
92	14.29	34.3764	20.0864	87	37	13.3908	4.33333
34	50.00	29.7437	20.2563	59	22	17.1356	6.18644
23	0.00	20.8102	20.8102	107	20	19.3458	4.85981
84	0.00	21.0344	21.0344	38	10	15.6316	6.81579
32	0.00	21.7669	21.7669	90	23	16.8444	5.63333
46	0.00	22.3762	22.3762	22	1	24.8636	9.90909
91	62.50	37.6823	24.8177	115	47	18.5565	5.62609
71	53.85	29.0137	24.8363	95	31	17.3474	5.40000
90	60.00	34.0946	25.9054	45	23	16.4222	6.20000
35	0.00	OUTLIER	.	66	21	16.4091	5.65152
7	58.82	29.7492	29.0708	147	52	17.4558	5.51020
76	56.25	OUTLIER	.	86	27	16.9186	5.90698
81	60.00	OUTLIER	.	24	8	16.7083	7.00000
62	83.33	OUTLIER	.	82	24	14.5000	5.69512

Predicted precision values for business report corpus

Text	Recall	Predicted Recall	Residual	Links/ Sent.	Tokens/ Sent.	Types/ Section
45	24.39	24.4093	0.0193	79.761	17.6126	37.391
66	21.57	21.5458	0.0242	105.038	18.0682	40.036
65	28.57	28.5426	0.0274	44.440	16.7200	31.630
27	11.76	11.9253	0.1653	66.721	16.2793	70.923
90	23.08	22.7005	0.3795	20.400	16.4222	55.800
8	25.00	24.4111	0.5889	57.500	15.9070	34.150
79	28.57	27.9495	0.6205	44.853	15.5098	25.381
64	28.57	29.3027	0.7327	58.697	17.0588	26.500
23	0.00	-0.8556	0.8556	65.065	19.3458	130.000
96	20.00	19.0252	0.9748	76.622	15.9679	44.000
98	18.75	17.6131	1.1369	142.748	18.5560	41.708
63	24.14	22.8233	1.3167	56.326	15.2569	35.111
80	33.33	31.8395	1.4905	44.541	16.7973	22.167
35	0.00	-1.5905	1.5905	33.258	16.4091	124.333
43	30.00	28.3199	1.6801	79.557	19.4348	37.563
59	27.78	29.7902	2.0102	27.932	16.7397	33.846
5	31.03	29.0192	2.0108	83.006	21.1104	45.158
19	23.44	21.0041	2.4359	166.637	19.4181	28.654
11	27.27	24.8082	2.4618	85.283	17.0500	30.567
95	21.43	23.9742	2.5442	55.263	17.7763	48.444
29	19.70	17.0767	2.6233	168.631	18.2642	32.256
22	33.33	36.1695	2.8395	31.933	20.9000	40.333
1	16.67	19.5429	2.8729	12.080	15.3600	61.333
10	15.38	12.3213	3.0587	44.624	18.4610	91.778
33	22.73	25.9155	3.1855	102.618	18.8382	32.750
44	21.43	24.7701	3.3401	21.043	15.1739	41.200

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Links/ Sent.	Tokens/ Sent.	Types/ Section
26	11.11	14.7607	3.6507	68.989	16.3258	61.882
16	17.74	21.4268	3.6868	105.523	17.0665	33.692
40	21.57	25.3020	3.7320	72.948	16.2288	28.094
68	22.22	26.0360	3.8160	15.971	15.1714	39.167
54	26.67	22.8237	3.8463	62.559	18.1864	52.000
72	16.07	20.2708	4.2008	101.775	16.2532	33.200
12	16.13	20.5700	4.4400	85.332	16.9065	42.381
50	20.75	25.2224	4.4724	146.224	20.7796	32.051
9	16.67	21.2742	4.6042	79.108	16.3892	39.091
78	22.22	17.6062	4.6138	19.441	14.4559	58.667
52	18.52	23.2664	4.7464	94.482	17.3957	34.207
61	33.33	28.3116	5.0184	45.520	15.5263	24.161
74	14.29	19.4477	5.1577	43.298	15.9008	54.091
41	7.14	12.3816	5.2416	55.633	16.0092	71.714
21	38.46	33.2133	5.2467	37.327	18.0962	29.050
89	19.35	24.6131	5.2631	63.993	14.6711	23.185
49	15.69	20.9801	5.2901	71.058	15.0167	33.880
14	16.25	21.7435	5.4935	124.512	20.2857	47.000
69	36.17	30.4531	5.7169	102.466	18.1131	14.407
82	25.00	31.1795	6.1795	18.226	18.2075	42.667
99	23.08	16.8841	6.1959	19.561	13.1061	52.000
4	27.27	33.6018	6.3318	25.791	16.7209	23.000
73	22.22	28.9106	6.6906	27.500	18.0208	45.000
100	12.77	6.0469	6.7231	117.167	16.7082	73.563
83	28.57	21.7686	6.8014	81.120	16.0629	34.762
57	41.18	34.2833	6.8967	30.622	18.1622	28.632
67	32.43	25.4831	6.9469	81.770	16.1094	23.646

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Links/ Sent.	Tokens/ Sent.	Types/ Section
30	17.07	24.0883	7.0183	83.117	18.0897	40.280
3	19.00	11.7521	7.2479	197.810	17.0391	29.977
39	22.58	29.9105	7.3305	75.347	20.5988	41.850
55	14.29	21.8341	7.5441	82.351	16.5586	37.360
25	17.86	25.6719	7.8119	55.878	17.1293	38.895
93	5.56	13.6473	8.0873	30.778	15.4667	73.167
38	30.00	21.8996	8.1004	47.500	16.9605	52.125
28	11.11	19.2589	8.1489	32.712	17.9091	71.500
15	24.62	16.4000	8.2200	217.524	20.1603	29.341
51	33.33	24.7980	8.5320	49.953	16.7717	41.294
2	9.09	0.4634	8.6266	62.687	18.2410	119.667
18	6.25	14.9181	8.6681	128.773	18.1530	52.148
17	10.00	18.8263	8.8263	62.886	17.5714	59.917
20	33.33	24.3137	9.0163	100.837	19.3315	41.421
70	12.50	3.4642	9.0358	24.387	15.8495	108.600
53	12.50	21.5890	9.0890	50.394	15.4882	42.438
60	15.38	24.6601	9.2801	24.514	14.5000	35.909
56	14.29	23.7814	9.4914	13.132	12.2368	27.833
36	40.91	30.2961	10.6139	54.482	17.5804	28.400
37	17.39	28.2452	10.8552	110.431	19.2117	25.400
42	38.46	27.0737	11.3863	59.558	18.9135	45.000
13	13.64	25.1083	11.4683	39.879	17.6207	49.462
84	0.00	11.5190	11.5190	14.737	15.6316	86.333
91	41.67	30.1314	11.5386	39.861	18.5565	40.438
7	37.04	24.6910	12.3490	45.517	17.4558	47.647
97	42.86	30.2928	12.5672	13.861	14.0000	19.455
77	14.29	26.9334	12.6434	25.565	16.7391	43.286

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Links/ Sent.	Tokens/ Sent.	Types/ Section
87	20.00	32.7554	12.7554	31.982	20.1091	45.444
92	11.76	24.6839	12.9239	29.253	13.3908	26.929
58	16.67	29.7210	13.0510	57.403	19.1223	39.150
48	11.11	24.8646	13.7546	30.314	13.6744	27.857
86	45.45	29.6667	15.7833	21.464	14.1964	19.929
75	7.69	23.8326	16.1426	25.353	13.9529	34.538
31	0.00	17.6300	17.6300	57.953	15.9257	54.538
85	50.00	30.7125	19.2875	23.404	15.8462	26.846
71	50.00	28.3359	21.6641	35.632	17.3474	39.462
62	33.33	10.7786	22.5514	24.207	14.5000	77.833
24	0.00	OUTLIER	.	22.164	15.5273	49.571
94	54.55	31.6303	22.9197	26.638	15.4138	20.118
32	0.00	OUTLIER	.	45.633	16.8444	46.091
88	50.00	25.9762	24.0238	25.268	16.3902	44.000
34	9.09	OUTLIER	.	25.017	17.1356	182.500
6	66.67	OUTLIER	.	17.174	21.8261	47.000
47	60.00	OUTLIER	.	21.587	14.7609	23.636
76	64.29	OUTLIER	.	29.640	16.9186	31.750
46	0.00	OUTLIER	.	26.045	24.8636	72.667
81	75.00	OUTLIER	.	13.208	16.7083	33.600

Predicted recall values for encyclopedia article corpus

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
58	33.33	33.5466	0.2166	8.2	20.120	7.2400	251.50	192.167	90.500
99	40.00	39.0426	0.9574	11.6	15.759	6.1724	274.20	237.200	107.400
26	25.00	26.0802	1.0802	6.4	35.423	4.7042	628.75	248.750	83.500
86	40.00	41.4769	1.4769	12.4	13.598	5.0805	236.60	218.200	88.400
35	25.00	26.7146	1.7146	9.8	14.883	6.9870	95.50	96.417	44.833
20	33.33	31.5704	1.7596	2.1	14.357	8.3929	134.00	154.333	78.333
66	27.27	29.3225	2.0525	15.5	22.017	5.2650	234.18	140.636	56.000
92	39.47	36.9839	2.4861	34.6	29.771	4.8229	212.32	96.500	34.395
1	50.00	47.1951	2.8049	2.7	14.714	9.3571	206.00	256.000	131.000
93	37.21	40.0373	2.8273	35.0	37.368	4.8576	250.28	96.605	32.535
69	37.50	34.4274	3.0726	17.9	27.568	6.3333	151.63	88.417	34.833
4	33.33	30.0807	3.2493	3.1	9.032	7.6774	93.33	155.000	79.333
18	35.48	39.0120	3.5320	43.1	60.822	4.5772	928.03	251.290	69.839
57	20.00	23.8975	3.8975	10.8	18.658	6.1139	147.40	106.500	48.300
65	36.36	31.8669	4.4931	4.9	25.032	8.9524	143.36	102.182	51.273
76	33.33	37.9968	4.6668	4.3	15.450	6.4500	206.00	199.000	86.000
15	50.00	45.3301	4.6699	7.3	21.070	6.0986	374.00	267.750	108.250
80	21.05	25.7431	4.6931	31.9	77.366	3.9639	789.95	152.816	40.474
52	18.18	13.3971	4.7829	12.0	49.139	5.0927	674.55	193.636	69.909
25	35.00	39.8356	4.8356	29.0	23.639	4.6466	294.30	163.300	57.850
60	26.67	31.9143	5.2443	9.7	67.355	5.4380	543.33	138.867	43.867
7	50.00	44.1051	5.8949	5.6	8.243	7.2973	152.50	263.000	135.000
22	100.00	94.0334	5.9666	8.6	12.864	8.4848	424.50	573.500	280.000
100	41.38	34.9355	6.4445	31.3	19.806	5.2938	144.10	97.172	38.517
78	27.27	33.7436	6.4736	12.5	21.965	8.3953	171.73	130.364	65.636
89	20.83	27.3450	6.5150	21.2	35.796	5.2857	219.25	80.042	32.375

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
91	36.21	42.7458	6.5358	50.0	57.197	4.1242	444.76	106.552	32.069
41	25.00	31.6706	6.6706	5.4	23.296	7.5211	413.50	274.750	133.500
96	43.59	36.9050	6.6850	36.1	38.938	4.5033	303.51	105.077	35.103
83	33.33	26.6326	6.6974	3.1	10.130	8.7826	77.67	122.333	67.333
68	25.00	31.7927	6.7927	4.2	16.000	6.7692	156.00	149.750	66.000
37	50.00	43.0286	6.9714	8.1	13.000	7.5556	175.50	214.250	102.000
33	42.86	35.5629	7.2971	7.4	33.292	6.6517	423.29	211.714	84.571
47	50.00	42.5568	7.4432	2.8	9.545	11.0000	105.00	210.500	121.000
74	33.33	40.9274	7.5974	2.5	15.310	11.4483	148.00	193.667	110.667
39	44.83	37.2270	7.6030	23.6	45.319	4.7685	337.55	114.483	35.517
90	20.00	27.7008	7.7008	9.7	5.964	6.4364	65.60	140.000	70.800
72	40.00	32.1414	7.8586	15.0	53.728	4.7333	698.47	213.467	61.533
55	33.33	25.3434	7.9866	11.5	24.315	5.5326	124.28	73.889	28.278
98	33.33	41.4751	8.1451	13.2	20.607	6.1589	367.50	256.667	109.833
61	44.44	35.6019	8.8381	10.3	35.685	4.9130	364.78	160.556	50.222
85	25.00	34.2490	9.2490	4.6	15.556	9.4444	140.00	158.250	85.000
50	7.69	17.0727	9.3827	9.0	62.730	4.1006	767.23	185.308	50.154
97	33.33	23.8833	9.4467	8.7	10.800	5.9167	108.00	125.500	59.167
40	60.00	49.9763	10.0237	16.2	32.915	5.7805	539.80	272.400	94.800
63	25.00	35.4071	10.4071	12.1	24.541	7.0450	170.25	116.875	48.875
23	30.36	19.9136	10.4464	49.6	144.274	4.6266	1373.18	162.054	44.036
73	44.44	33.9474	10.4926	26.9	55.514	4.9804	786.44	232.278	70.556
44	43.75	30.8938	12.8562	12.5	33.950	6.3762	214.31	98.688	40.250
17	33.33	20.2519	13.0781	4.3	7.615	8.9231	33.00	67.167	38.667
42	50.00	36.7758	13.2242	8.0	26.378	6.6081	488.00	284.500	122.250
62	25.00	38.6582	13.6582	13.9	26.144	9.1667	287.58	196.500	100.833
46	50.00	36.2057	13.7943	3.1	5.733	13.2000	43.00	150.000	99.000

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
48	14.29	28.2341	13.9441	5.5	22.463	4.8060	215.00	133.286	46.000
38	25.00	39.3550	14.3550	6.5	28.343	5.7286	496.00	267.000	100.250
5	50.00	35.5055	14.4945	2.7	18.412	7.0294	313.00	254.000	119.500
94	0.00	14.6078	14.6078	7.1	4.176	5.8529	35.50	93.000	49.750
34	50.00	35.0915	14.9085	3.6	10.000	7.2000	125.00	185.000	90.000
12	57.14	41.4777	15.6623	6.0	27.228	7.3291	307.29	199.714	82.714
87	33.33	17.3754	15.9546	3.8	5.577	7.0769	48.33	108.667	61.333
59	35.71	18.9940	16.7160	11.7	53.955	5.5414	605.07	169.214	62.143
14	50.00	32.8326	17.1674	8.7	18.296	6.2535	216.50	170.667	74.000
95	33.33	15.2102	18.1198	4.8	4.667	6.5185	42.00	104.000	58.667
3	50.00	31.6808	18.3192	2.2	10.783	8.0870	124.00	177.500	93.000
67	50.00	31.3038	18.6962	1.6	5.917	11.6667	35.50	114.000	70.000
11	25.00	43.8068	18.8068	7.1	12.340	7.9434	163.50	216.500	105.250
31	100.00	80.7962	19.2038	8.0	15.806	7.4167	569.00	569.000	267.000
54	66.67	46.9840	19.6860	6.0	10.913	7.9783	167.33	247.000	122.333
16	20.00	40.7702	20.7702	5.6	23.395	7.5116	402.40	282.200	129.200
45	18.75	39.5510	20.8010	16.8	37.748	6.2147	384.56	171.375	63.313
9	58.33	37.2799	21.0501	14.9	22.810	5.9683	239.50	158.917	62.667
27	16.67	37.7441	21.0741	9.4	22.132	6.6044	335.67	229.833	100.167
8	12.50	33.5836	21.0836	9.3	22.190	6.3048	291.25	193.000	82.750
70	66.67	45.4919	21.1781	5.1	13.028	9.6667	156.33	221.667	116.000
81	57.14	35.8829	21.2571	8.6	17.259	6.7176	209.57	182.571	81.571
29	50.00	28.7054	21.2946	13.1	43.895	5.9211	667.20	244.600	90.000
28	66.67	45.2324	21.4376	4.0	28.667	8.8039	487.33	320.333	149.667
77	0.00	23.0592	23.0592	3.0	9.050	8.2000	60.33	101.667	54.667
32	0.00	25.5275	25.5275	2.4	15.806	5.7742	163.33	140.667	59.667
84	50.00	23.1889	26.8111	4.4	5.778	6.5926	78.00	162.000	89.000

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
75	16.67	43.6175	26.9475	19.5	44.335	6.3107	761.08	304.750	108.333
82	0.00	28.0834	28.0834	2.8	7.895	8.2105	75.00	144.500	78.000
2	50.00	20.8695	29.1305	3.1	10.083	6.7500	60.50	87.500	40.500
36	0.00	29.9958	29.9958	6.0	10.026	8.2308	130.33	190.333	107.000
88	0.00	30.6765	30.6765	9.7	6.933	7.1556	104.00	193.667	107.333
53	0.00	30.9775	30.9775	3.0	10.917	6.4167	131.00	167.000	77.000
51	0.00	31.3832	31.3832	7.3	16.899	6.5316	190.71	163.714	73.714
30	0.00	32.5195	32.5195	3.4	20.889	8.1111	150.40	126.000	58.400
13	66.67	33.6024	33.0676	6.3	11.068	5.6364	162.33	187.000	82.667
24	0.00	33.8183	33.8183	3.1	18.316	8.4737	174.00	162.750	80.500
43	0.00	34.6529	34.6529	3.5	10.172	8.3448	147.50	221.000	121.000
21	0.00	36.5711	36.5711	6.4	19.600	8.1333	235.20	199.600	97.600
6	0.00	37.4550	37.4550	5.1	26.344	6.6230	401.75	242.750	101.000
56	66.67	22.4322	44.2378	2.7	10.826	7.3478	83.00	110.667	56.333
49	83.33	OUTLIER	.	5.5	30.603	5.7931	295.83	164.500	56.000
19	75.00	28.2418	46.7582	3.5	13.132	6.7105	124.75	139.000	63.750
64	100.00	OUTLIER	.	4.4	12.300	6.5000	246.00	282.000	130.000
79	0.00	OUTLIER	.	5.1	18.170	7.1915	427.00	378.000	169.000
10	100.00	OUTLIER	.	4.5	9.281	7.8438	148.50	234.500	125.500
71	0.00	OUTLIER	.	7.8	15.857	6.8413	499.50	470.500	215.500

Predicted precision values for encyclopedia article corpus

Text	Precision	Predicted Precision	Residual	Sections	Bound-aries	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
78	23.08	23.1076	0.0276	11	27	8.3953	171.73	130.364	65.636
4	16.67	16.2952	0.3748	3	10	7.6774	93.33	155.000	79.333
72	13.64	12.9621	0.6779	15	79	4.7333	698.47	213.467	61.533
86	13.33	12.3901	0.9399	5	28	5.0805	236.60	218.200	88.400
41	11.11	10.0873	1.0227	4	17	7.5211	413.50	274.750	133.500
92	27.78	28.9284	1.1484	38	108	4.8229	212.32	96.500	34.395
99	11.76	10.4981	1.2619	5	31	6.1724	274.20	237.200	107.400
76	16.67	15.3834	1.2866	3	14	6.4500	206.00	199.000	86.000
20	20.00	18.5392	1.4608	3	9	8.3929	134.00	154.333	78.333
52	9.52	7.9449	1.5751	11	46	5.0927	674.55	193.636	69.909
18	12.50	10.9075	1.5925	31	177	4.5772	928.03	251.290	69.839
25	15.91	14.2520	1.6580	20	97	4.6466	294.30	163.300	57.850
34	16.67	15.0068	1.6632	2	8	7.2000	125.00	185.000	90.000
15	16.67	14.9952	1.6748	4	23	6.0986	374.00	267.750	108.250
66	16.67	14.9512	1.7188	11	37	5.2650	234.18	140.636	56.000
96	29.82	28.0508	1.7692	39	109	4.5033	303.51	105.077	35.103
57	14.29	16.1210	1.8310	10	30	6.1139	147.40	106.500	48.300
35	18.75	20.9716	2.2216	12	36	6.9870	95.50	96.417	44.833
69	31.03	28.4750	2.5550	24	58	6.3333	151.63	88.417	34.833
98	7.41	10.3557	2.9457	6	44	6.1589	367.50	256.667	109.833
5	14.29	11.2712	3.0188	2	12	7.0294	313.00	254.000	119.500
28	18.18	14.9973	3.1827	3	18	8.8039	487.33	320.333	149.667
87	16.67	12.9627	3.7073	3	9	7.0769	48.33	108.667	61.333
23	18.48	14.7402	3.7398	56	192	4.6266	1373.18	162.054	44.036
68	12.50	16.2637	3.7637	4	17	6.7692	156.00	149.750	66.000
40	21.43	17.3084	4.1216	10	58	5.7805	539.80	272.400	94.800

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
26	6.25	10.3938	4.1438	4	26	4.7042	628.75	248.750	83.500
100	30.77	26.6018	4.1682	29	78	5.2938	144.10	97.172	38.517
42	14.29	10.0266	4.2634	4	31	6.6081	488.00	284.500	122.250
47	16.67	20.9546	4.2846	2	9	11.0000	105.00	210.500	121.000
80	11.27	15.7629	4.4929	38	144	3.9639	789.95	152.816	40.474
31	11.76	7.0544	4.7056	2	30	7.4167	569.00	569.000	267.000
89	17.86	22.5917	4.7317	24	55	5.2857	219.25	80.042	32.375
90	9.09	13.9570	4.8670	5	15	6.4364	65.60	140.000	70.800
48	9.09	14.1988	5.1088	7	28	4.8060	215.00	133.286	46.000
60	23.53	18.0254	5.5046	15	35	5.4380	543.33	138.867	43.867
83	25.00	19.4546	5.5454	3	7	8.7826	77.67	122.333	67.333
67	33.33	27.7132	5.6168	2	4	11.6667	35.50	114.000	70.000
16	6.67	12.3323	5.6623	5	33	7.5116	402.40	282.200	129.200
38	9.09	14.8331	5.7431	4	20	5.7286	496.00	267.000	100.250
33	21.43	15.5334	5.8966	7	32	6.6517	423.29	211.714	84.571
61	22.22	16.2274	5.9926	9	35	4.9130	364.78	160.556	50.222
22	16.67	10.6741	5.9959	2	23	8.4848	424.50	573.500	280.000
85	14.29	20.4522	6.1622	4	13	9.4444	140.00	158.250	85.000
93	25.81	32.0792	6.2692	43	120	4.8576	250.28	96.605	32.535
65	33.33	26.8501	6.4799	11	23	8.9524	143.36	102.182	51.273
73	20.51	13.9240	6.5860	18	85	4.9804	786.44	232.278	70.556
7	16.67	9.9723	6.6977	2	17	7.2973	152.50	263.000	135.000
1	11.11	17.8533	6.7433	2	11	9.3571	206.00	256.000	131.000
59	17.86	10.9548	6.9052	14	54	5.5414	605.07	169.214	62.143
50	5.00	12.0140	7.0140	13	43	4.1006	767.23	185.308	50.154
27	6.25	13.4172	7.1672	6	31	6.6044	335.67	229.833	100.167
55	30.00	22.8131	7.1869	18	40	5.5326	124.28	73.889	28.278

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
91	27.27	34.4589	7.1889	58	166	4.1242	444.76	106.552	32.069
14	21.43	14.2335	7.1965	6	27	6.2535	216.50	170.667	74.000
63	17.39	24.8011	7.4111	16	45	7.0450	170.25	116.875	48.875
8	5.88	13.4641	7.5841	8	40	6.3048	291.25	193.000	82.750
81	23.53	15.8470	7.6830	7	30	6.7176	209.57	182.571	81.571
62	10.00	18.0797	8.0797	12	55	9.1667	287.58	196.500	100.833
84	16.67	8.5294	8.1406	2	10	6.5926	78.00	162.000	89.000
95	20.00	11.6952	8.3048	3	6	6.5185	42.00	104.000	58.667
11	9.09	17.5015	8.4115	4	17	7.9434	163.50	216.500	105.250
29	18.52	9.9938	8.5262	10	61	5.9211	667.20	244.600	90.000
75	5.00	13.6720	8.6720	12	78	6.3107	761.08	304.750	108.333
12	28.57	19.8478	8.7222	7	27	7.3291	307.29	199.714	82.714
70	28.57	19.8112	8.7588	3	14	9.6667	156.33	221.667	116.000
88	0.00	9.0573	9.0573	3	14	7.1556	104.00	193.667	107.333
97	22.22	12.8988	9.3212	6	22	5.9167	108.00	125.500	59.167
74	33.33	23.9859	9.3441	3	7	11.4483	148.00	193.667	110.667
3	25.00	15.3816	9.6184	2	7	8.0870	124.00	177.500	93.000
71	0.00	9.7350	9.7350	2	20	6.8413	499.50	470.500	215.500
45	8.57	18.5312	9.9612	16	67	6.2147	384.56	171.375	63.313
13	22.22	11.9263	10.2937	3	15	5.6364	162.33	187.000	82.667
39	36.11	25.4628	10.6472	29	78	4.7685	337.55	114.483	35.517
94	0.00	11.1361	11.1361	4	11	5.8529	35.50	93.000	49.750
36	0.00	11.1851	11.1851	3	18	8.2308	130.33	190.333	107.000
2	28.57	17.2578	11.3122	4	10	6.7500	60.50	87.500	40.500
37	28.57	16.4077	12.1623	4	19	7.5556	175.50	214.250	102.000
9	29.17	16.9961	12.1739	12	52	5.9683	239.50	158.917	62.667
43	0.00	12.4732	12.4732	2	8	8.3448	147.50	221.000	121.000

Continued on next page

Continued from previous page

Text	Recall	Predicted Recall	Residual	Avg. Median Diff.	Links/ Sent.	Types/ Sent.	Links/ Section	Tokens/ Section	Types/ Section
54	28.57	16.0571	12.5129	3	14	7.9783	167.33	247.000	122.333
58	28.57	15.2017	13.3683	6	22	7.2400	251.50	192.167	90.500
79	0.00	13.4432	13.4432	2	19	7.1915	427.00	378.000	169.000
32	0.00	13.5748	13.5748	3	9	5.7742	163.33	140.667	59.667
53	0.00	14.3892	14.3892	2	5	6.4167	131.00	167.000	77.000
6	0.00	14.4320	14.4320	4	20	6.6230	401.75	242.750	101.000
51	0.00	15.0852	15.0852	7	29	6.5316	190.71	163.714	73.714
17	40.00	24.8642	15.1358	6	7	8.9231	33.00	67.167	38.667
64	28.57	12.2645	16.3055	2	13	6.5000	246.00	282.000	130.000
82	0.00	16.4323	16.4323	2	7	8.2105	75.00	144.500	78.000
21	0.00	16.4986	16.4986	5	21	8.1333	235.20	199.600	97.600
10	28.57	10.5090	18.0610	2	14	7.8438	148.50	234.500	125.500
77	0.00	19.3746	19.3746	3	4	8.2000	60.33	101.667	54.667
24	0.00	19.6821	19.6821	4	10	8.4737	174.00	162.750	80.500
30	0.00	OUTLIER	.	5	7	8.1111	150.40	126.000	58.400
19	50.00	OUTLIER	.	4	14	6.7105	124.75	139.000	63.750
44	58.33	OUTLIER	.	16	39	6.3762	214.31	98.688	40.250
49	62.50	OUTLIER	.	6	23	5.7931	295.83	164.500	56.000
46	100.00	OUTLIER	.	2	4	13.2000	43.00	150.000	99.000
56	100.00	OUTLIER	.	3	6	7.3478	83.00	110.667	56.333

Author Index

Darnton (1987), 77, 78, 80, 187, 427
Davies (1985), 220, 427
Davies (1994), 19, 68–70, 72, 73, 81, 427
de Beaugrande and Dressler (1981), 153, 427
de Beaugrande (1997), 181, 427
Di Pietro (1983), 12, 427
Donaldson et al. (1996), 5, 427

Eggins (1994), 134, 135, 175, 185, 404, 427
Ehrich and Koster (1983), 7, 427
Everitt (1974), 220, 427

Firth (1957), 4, 427
Fisher (1994), 5, 427
Frawley (1987), 181, 428
Fries (1990), 7, 10, 11, 428
Fries (1995), 17, 428

Garcia-Berrio and Mayordomo (1987), 8, 428
Georgakopoulou and Goutsos (1997), 2, 87, 409, 414, 428
Giora (1983), 7, 19, 74–76, 81, 409, 428
Girden (1992), 338, 428
Glass (1983), 21, 428
Gledhill (1995), 86, 428
Good (1977), 221, 428
Goutsos (1996a), 7, 19, 63–67, 81, 192, 279, 396–398, 406, 408, 411, 412, 428
Goutsos (1996b), 85, 184, 428
Graustein and Thiele (1983), 396, 397, 428
Grefenstette and Tapainen (1994), 335, 429
Gregory (1985a), 12, 429
Gregory (1985b), 10, 429
Grimes (1975), 8–10, 163, 429
Grimshaw (1991), 11, 429
Grosz and Sidner (1986), 7, 20, 21, 399, 429
Grosz et al. (1989), 187, 429
Grosz (1995), 5, 8, 429
Guide to Corporate filings (1997), 332, 429

Haas and Losee (1994), 90, 429
Hahn and Strube (1997), 124, 420, 429
Hak and Helsloot (1995), 172, 429
Halliday and Hasan (1976), 129–132, 136, 140–142, 145, 155, 175, 403, 404, 406, 430
Halliday and Hasan (1989), 7, 19, 31–36, 81, 129, 136–140, 142–146, 149, 150, 153, 175, 186, 260, 261, 399, 403, 404, 412, 430
Halliday (1978), 10, 429
Halliday (1985), 68, 132, 133, 175, 404, 430
Halliday (1994), 133, 185, 430
Harris (1951), 9, 430
Harris (1989), 4, 430
Hasan (1977), 6, 10, 430
Hasan (1984), 104, 185, 186, 403, 411, 430
Hasan (1996a), 37, 410, 430
Hasan (1996b), 25, 430
Hearst and Plaunt (1993), 109, 111, 113, 117, 122, 125, 215, 296, 403, 404, 430
Hearst (1993), 21, 109, 111, 117, 215, 296, 401, 430
Hearst (1994a), 6, 111, 112, 215, 216, 281, 296, 402, 406, 430
Hearst (1994b), 7, 109, 188, 215, 296, 430
Hinds (1979), 7, 431
Hirschberg and Litman (1993), 80, 93, 431
Hockey and Ide (1994a), 1, 4, 410, 431
Hockey and Ide (1994b), 1, 4, 410, 431
Hoey and Winter (1986), 411, 432

Hoey and Wools (1995), 257, 432
Hoey (1983), 7, 10, 19, 52–58, 81, 399, 404, 411, 413, 431
Hoey (1985), 8, 132, 184, 431
Hoey (1986), 322, 431
Hoey (1988), 149, 403, 413, 431
Hoey (1991a), 403, 404, 431
Hoey (1991b), 12, 104, 122, 126, 129, 149–151, 153, 155–165, 169, 171, 173–176, 179, 180, 185, 186, 188–192, 195, 205, 206, 249, 254, 256–261, 281, 289, 396, 397, 404, 405, 409, 410, 413, 415, 431
Hoey (1993), 2, 431
Hoey (1994), 421, 431
Hoey (1995a), 13, 431
Hoey (1995b), 173, 431
Hoey (1996), 107, 109, 132, 184, 432
Hopkins and Dudley-Evans (1988), 27, 30, 399, 432
Hrebicek and Altmann (1993), 15, 432
Hughes and Atwell (1994), 230, 432
Hughes and Atwell (nd), 231, 432
Humphrey (1996), 116, 117, 123, 432
Hwang (1989), 184, 432
Hyland (1990), 27, 30, 399, 432

Jordan (1984), 7, 10, 432

Karlgren et al. (1995), 230, 432
Kirk (1994), 3, 432
Knott and Dale (1993), 5, 433
Kozima and Furugori (1993), 7, 87, 92, 122, 188, 403, 413, 433
Kozima (1993a), 92–96, 281, 403, 413, 433
Kozima (1993b), 92, 185, 402, 403, 413, 433
Kukharenko (1979), 7, 16, 17, 433
Kyto et al. (1988), 1, 433

Labov and Waletzky (1967), 6, 433
Labov (1972), 9, 433
Lamprecht (1988), 7, 433
Lancashire (1991), 1, 410, 433
Landow and Delany (1993), 1, 410, 433
Langleben (1979), 7, 433
Ledger (1989), 230, 433
Leech and Fligelstone (1992), 3, 434
Lewis-Beck (1980), 356, 358, 361, 434
Lewis (1996), 6, 434
Lohmann (1988), 186, 434
Longacre and Levinsohn (1978), 278, 434
Longacre (1976), 77, 434
Longacre (1979), 184, 434
Longacre (1983), 19, 76, 77, 79, 80, 182, 187, 396, 410, 434
Lorch and Lorch (1996), 85, 408, 434

Mann and Thompson (1986a), 10, 434
Mann and Thompson (1986b), 19, 21, 58, 59, 81, 399, 411, 434
Mann and Thompson (1987a), 15, 21, 58–62, 113, 182, 399, 434
Mann and Thompson (1987b), 182, 434
Mann and Thompson (1988), 182, 435
Mann and Thompson (1992), 11, 435
Mann et al. (1989), 58–61, 87, 88, 181, 399, 434
Marcu (1997), 62, 435
Markels (1983), 106, 435
Marshall (1991), 27, 30, 399, 435
Martin (1989), 63, 435
Martin (1992), 10, 133, 140, 435
Matthiessen (1988), 10, 435

McEnery and Wilson (1996), 1, 12, 435
Meyer and Rice (1984), 21, 435
Miall (1992), 4, 435
Miller et al. (1990), 112, 435
Milligan and Cooper (1985), 242, 243, 435
Mitchell (1957/1975), 10, 435
Morris and Hirst (1991), 105–107, 109, 113, 122, 216, 311, 402–404, 413, 436
Morris (1988), 21, 105, 107–109, 113, 114, 122, 123, 188, 216, 311, 402–404, 413, 436

Norusis (1990), 383, 436
Nwogu (1991), 27, 30, 399, 436

Okumura and Honda (1994), 108, 109, 122, 216, 310, 402, 413, 436
Ostler (1987), 21, 47, 436

Pêcheux (1969/1995), 176, 405, 437
Paduceva (1974), 184, 436
Paltridge (1994), 20, 36–38, 399, 405, 436
Parsons (1990), 19, 39, 144–146, 149, 153, 172, 175, 186, 369, 404, 436
Parsons (1996), 145, 153, 186, 436
Passonneau and Litman (1993), 82, 107, 120, 121, 393, 436
Passonneau and Litman (1995), 80, 100, 102, 436
Petöfi and Rieser (1973), 397, 437
Petöfi and Sözer (1987), 406, 437
Petöfi (1979), 397, 437
Petöfi (1982), 397, 437
Phillips (1985), 4, 13, 14, 83, 85–87, 119, 153, 183, 185, 186, 224, 225, 394, 395, 398, 408, 413, 437
Phillips (1989), 12, 181, 183, 398, 408, 413, 437
Pike and Pike (1977), 10, 437
Pike (1972), 10, 437
Pike (1982), 10, 437
Pitkin (1969), 19, 21, 47–52, 81, 397, 437
Pollard-Gott et al. (1979), 226, 437

Quirk et al. (1985), 228, 437

Raben (1991), 4, 437
Renouf and Collier (1995), 82, 168, 169, 437
Reynar (1994), 114–117, 123, 403, 438
Rietveld and Van Hout (1993), 267, 268, 438
Rodgers (1966), 184, 438
Rotondo (1984), 222, 226, 227, 438
Rumelhart (1975), 21, 438

Sacks et al. (1974), 41, 438
Salager-Meyer (1989), 27, 30, 399, 438
Salager-Meyer (1990), 27, 399, 438
Salton and Buckley (1991), 5, 401, 438
Salton et al. (1990), 5, 438
Salton et al. (1994), 5, 118, 119, 401, 438
Salton (1988), 5, 84, 438
Sarle (1983), 263, 438
SAS Institute Inc (1989a), 220, 223, 224, 338, 438
SAS Institute Inc (1989b), 351, 360, 361, 438
Schegloff and Sacks (1973), 63, 439
Schiffrin (1994), 9, 20, 225, 439
Schroeder et al. (1986), 359, 439
Scinto (1986), 7, 17, 439
Scott (1996), 331, 439
Scott (1997), 2–4, 16, 408, 413, 414, 439
Sibson (1972), 330, 439
Siegel (1975), 381, 439
Sinclair and Coulthard (1975), 63, 410, 439

Sinclair and Coulthard (1992), 19, 439
Sinclair (1966), 3, 4, 439
Sinclair (1991), 1, 4, 12, 16, 87, 395, 439
Sinclair (1994), 1, 15, 182, 439
Skorochod'ko (1972), 110, 439
Smadja (1992), 4, 439
Sparck Jones (1996), 9, 123, 124, 186, 439
St-Onge (1995), 122, 440
Stairmand and Black (1996), 123, 404, 440
Stairmand (1996a), 123, 404, 440
Stairmand (1996b), 123, 440
Stoddard (1991), 181, 440
Stubbs (1983), 13, 440
Stubbs (1996), 1, 12–15, 186, 410, 440
Svartvik (1990), 92, 440
Svartvik (1992), 1, 440
Svartvik (1996), 2, 440
Swales (1981), 27, 30, 399, 440
Swales (1990), 7, 27, 37, 63, 81, 86, 184, 399, 410, 440

Tabachnick and Fidell (1989), 338, 359, 360, 440
Thompson (1996), 65, 279, 441
Thorndyke (1977), 6, 441
Thury (1988), 4, 441
Tinberg (1988), 27, 30, 399, 441

van Dijk and Kintsch (1983), 410, 441
van Dijk and Petöfi (1977), 11, 441
van Dijk (1972), 397, 441
van Dijk (1980), 7, 19, 63, 396, 397, 399, 409–411, 441
van Dijk (1983), 410, 441
van Dijk (1985), 23, 441
van Rijsbergen (1979), 100, 441
Ventola (1979), 10, 441
Ventola (1986), 182, 441

Wessels (1993a), 154, 441
Wessels (1993b), 154, 171, 172, 441
Widdowson (1978), 153, 441
Wimmer and Dominick (1991), 282, 283, 441
Winburne (1962), 126–129, 176, 405, 442
Winter (1971), 52, 442
Winter (1974), 53, 170, 176, 404, 405, 413, 442
Winter (1977), 52, 53, 58, 150, 151, 442
Winter (1979), 53, 153, 154, 442
Woods et al. (1986), 221, 283, 442

Yarowsky (1992), 112, 442
Youmans (1991), 89–91, 93, 94, 185, 402, 403, 406, 442

Subject Index