Automatic Text Summarisation through Lexical Cohesion Analysis

Mohamed Benbrahim

Ph.D. Thesis

Submitted to the University of Surrey in partial fulfilment of requirements for the degree of Doctor of Philosophy

Artificial Intelligence Group
Department of Mathematical and Computing Sciences
University of Surrey
Guildford GU2 5XH
Surrey
United Kingdom

April 1996
Abstract

A methodology for automatically summarising scientific texts is presented using the patterns of lexical cohesion found in such texts. Lexical cohesion is a type of cohesion whereby certain lexical features of the text connect sentences with each other in the text. An analysis of lexical cohesion in text, primarily by counting repetitions, synonyms, and paraphrase, leads to the establishment of a network of sentences, some tightly bonded through lexical cohesion relations, some others having weak bonds or no bonds at all. The strength of connections in this cohesion network is used to identify key sentences in a text. Some sentences open key topics, some close topics, whilst others consolidate a given topic. Topic opening, closing, and consolidating, or central sentences, have different strengths and different connectivity patterns. A selection of these sentences can be construed as a summary of a given text. TELE-PATTAN (TExt and LExical cohesion PATTerns ANalysis), a system for summarising text automatically, extracts patterns of lexical cohesion in a text, categorises its sentences, and subsequently produces summaries of the text on the basis of these patterns. Experiments were conducted with human subjects to evaluate the summaries. The results of this preliminary evaluation are encouraging.
Acknowledgements

I would like to acknowledge Khurshid Ahmad for his continuous support and help throughout the period of my doctoral research. I thank him particularly for personally sharing some of the difficulties I encountered during the research period. I would like also to thank David Bowers for his support during the first year of this doctoral research.

My thanks are due to all former and current members of the Artificial Intelligence Group for their friendly support, collaboration, and suggestions, especially, Paul Holmes-Higgin for his early tips in mastering the 'Eleven Commandments of Prolog'. And, special thanks are due to Tracey Bale and Andrew Salway for their help in proof reading this thesis.

I would like to thank Professor Stephen A. Schuman and Professor Trevor J. Sweeting for providing a supportive and friendly research environment within the Department of Mathematical and Computing Sciences. I particularly extend my thanks to them and to Khurshid Ahmad for their sympathetic support after an unfortunate incident which lead to a delay in my studies. My thanks are also due to the Ministry of Industry and Mines of the Algerian Government for sponsoring my PhD.

Of course, I am indebted most to my parents who, since my birth and throughout the years of my education, have raised me, supported me, and provided me with the best advice. To them, I will always be grateful, and to them, I dedicate this thesis. Finally, I would like to thank all my friends, some in Britain, some in Algeria, and others thousands of miles away, ever so much for their sincere, enthusiastic, and affectionate support which, to me, was very precious indeed.
## Contents

### CHAPTER 1: INTRODUCTION

1.1 Motivation .......................................................... 1
1.2 Objectives .......................................................... 4
1.3 Achievements ....................................................... 6
1.4 Outline of the thesis ............................................. 9

### CHAPTER 2: PROGRAMS FOR UNDERSTANDING AND SUMMARISING TEXT - A REVIEW...

2.1 Conventional Artificial Intelligence approaches to text understanding ........................................ 10
   2.1.1 SAM .................................................................. 11
   2.1.2 PAM .................................................................. 13
2.2 The SMART text retrieval system ................................................................. 16
2.3 The TOPIC Text Condensation System .......................................................... 19
2.4 Lexical Chains ......................................................................................... 22
2.5 Cohesion index analysis ........................................................................... 25
2.6 SERAPHIN - a French text abstract generator ............................................. 28
2.7 The Text Tiling and discourse segmentation system ...................................... 30
2.8 Conclusions ............................................................................................. 32

### CHAPTER 3: LEXICAL COHESION AND ITS ROLE IN TEXT UNDERSTANDING

3.1 Introduction ................................................................................. 38
3.2 Lexical and syntactic cohesion .......................................................... 42
   3.2.1 Syntactic Cohesion ............................................... 43
   3.2.2 Lexical Cohesion .................................................. 44
3.3 ‘Patterns of lexis’ in text .................................................................. 49
   3.3.1 Simple Lexical Repetition ...................................... 49
   3.3.2 Complex Lexical Repetition ................................... 51
   3.3.3 Paraphrase .......................................................... 52
3.4 Automatic cohesion-based text analysis - Assumptions and hypotheses ............ 54
3.5 Conclusions ....................................................................................... 57

### CHAPTER 4: COHESION AND SUMMARISATION OF NON-NARRATIVE TEXTS

4.1 Introduction .................................................................................. 59
4.2 Sentence categorisation and lexical cohesion ........................................... 63
   4.2.1 Topic opening sentences ....................................... 63
   4.2.2 Topic closing sentences ........................................ 64
   4.2.3 Central sentences ............................................... 64
   4.2.4 Key Central sentences ......................................... 65
   4.2.5 Marginal sentences .............................................. 66
4.3 Text summarisation ........................................................................... 67
4.4 Computing lexical cohesion: The need for linguistic knowledge resources .... 69
4.5 A lexical cohesion-based text summarisation methodology ....................... 72
   4.5.1 Architecture ....................................................... 74
Chapter 1

Introduction

The flow of scientific and technical information is increasing dramatically everyday. Well-motivated scientists and engineers communicate with each other across languages, having no problems with each others’ terminology, and obtain the latest research paper, the most recent version of software documentation and so on. The question is, or perhaps should be, how others, that is those who have recently joined a scientific community, people across other disciplines, and users of scientific and technical artefacts and lay people, can retrieve, browse and understand the contents of this flexibly available public resource. It can be argued that the information sciences classification, used to organise texts in a paper library will not suffice due to the sheer volume, diversity, and interdisciplinarity of documents which blur into a digital Tower of Babel.

An effective way of disseminating (long) documents will be to make available a short summary of each of the documents. A potential user of a document will be able to retrieve, or indeed not to retrieve, a document depending on the information available in the summary. Of course, the most desirable thing would be to ask the authors to append a summary of each document they produce. The next best thing would be a computer-generated summary produced by a program which can mimic the summarising skills of a human being.

Summaries of long documents are produced in a variety of ways. It is customary now to attach an ‘executive summary’ of the document: the term ‘executive’ is deliberate as it refers generally to a busy executive who really cannot be expected to read the whole
document but is keen to get the gist. Executive summaries are, perhaps, a solution to
the problem of ‘having too much to read’. But, like abstracts, executive summaries are
also written generally by experts for experts. Most research papers, unlike research
reports, do not contain executive summaries.

Over the last two decades, computer scientists, particularly those working in artificial
intelligence (AI) and information retrieval (IR), have been trying to develop and test
various methods for studying, analysing, and summarising text. AI methods for
generating summaries consisted of subjecting a suitably formatted text to the ‘slot-
filling’ operations of ‘frames’ and ‘scripts’ which are known as knowledge
representation schemata. Some of the (AI) systems that were used to summarise short
stories based on these ‘rigid’ knowledge representation schemata include the systems
SAM and PAM (Schank, 1980; Cullingford, 1986).

IR methods for text summarisation are based on the use of document vectors, essentially
lists of pre-selected keywords which are organised and ordered for retrieving and
browsing documents (Salton, 1989). The reduction of text into in a list of pre-selected
keywords which are used as indicative elements to produce document abstracts has a
few problems. First, by its very nature, the list of keywords has to be compiled and
revised at fixed points in time, thus documents containing neologisms cannot be indexed
properly. Second, the selection of keywords is based only on the single pragmatic
feature of ‘frequency of use’ whilst ignoring the communicative intent of the author, the
theme of the document, genre and tenor. Third, keywords or document feature vectors
are known to be ‘reductive’ in that these words/terms tend to reflect the highest
common usage (of words) throughout a domain, neglecting rare, but sometimes crucial
words. Although IR methods have proved efficient in retrieving hundreds of documents at a touch of a button in response to a query or a request, information retrieval experts have come to agree that in order to achieve goals such as automatic text processing and summarisation, an in-depth study of the nature of cohesion in text is of prime importance (Paice, 1990; Hearst, 1993; Sparck Jones, 1992). Such a study may lead to an understanding of how sentences, paragraphs, and sections, which are structurally distinct units of a text, are connected to form a coherent and meaningful text.

The work we present in this thesis is about text analysis, understanding, and summarisation using the notion of cohesion in text: the understanding of elements of information described in some parts of the text depends on or refers to other elements of information in other parts. Text linguists argue that the way writers and composers of text link sentences together by using syntactic and lexical signals such as pronominal reference, conjunction, repetition, comparison, contrast, etc., constitutes the very notion of textual cohesion (or cohesion in text). This notion is regarded as essential for the existence of text, its meaningfulness, and the consistency and clarity of the message it conveys. Cohesion in texts manifests itself at many levels: authors of texts repeat keywords in texts more than other words with the exception of the so-called closed-class words (that include determiners, moral verbs, conjunctions, and pronouns), this repetition is essentially to 'hammer away at' the basic message of the text. This repetition may be just the repetition of the same word, its plurals, its synonym, or its paraphrase.
1.1 Motivation

The linguists, Halliday and Hasan, define cohesion as part of the system of a language that describes primarily a semantic role which is set-up to account for relations in discourse without the implication that there is a structural unit above the sentence itself. They specifically describe cohesion as '...the range of possibilities that exist for linking something with what has gone before'. They further characterise cohesion in text as '...a set of semantic resources for linking a sentence with what has gone before' (1976:10). The expression, 'what has gone before', means all previous sentences, paragraphs, sections, or chapters of a texts. Halliday and Hasan describe how sentences within a text which are in principle structurally independent of one another may be 'linked through particular features of their interpretation'. They argue that text has 'texture' by virtue of its property of 'being a text' and the fact that it 'functions as a unity with respect to its environment'. Cohesion in text is one of the resources that are used to create 'texture'. This texture is established partly because sentences, paragraphs, and sections which contain cohesively related items are 'tied' to one another. A text can therefore be characterised by the number and types of the cohesive ties it has.

Hoey (1991), in his work on 'patterns of lexis' in text, has shown that lexis and text are an important level of organisation and that they interact constructively to form a regular contiguous unit. He demonstrated by analysing small excerpts of non-narrative text, e.g. scientific text, that patterns of lexis exist in text and can be used as a metaphor for text abridgement. Hoey argues that text is patterned but does not have structure such that one can 'make predictive statements about text organisation', and that such patterning reveals 'complex ways in which topics may interrelate in their development'. And, these
lexical patterns 'may allow us to say interesting thing about it [i.e. text] and to elicit coherent subtexts from it'.

The notion of cohesion in text may be quite important for both AI and IR workers in that it may help to resolve the question of how the form and content of a text can be described and represented in a computer system. Some authors will insist that we must first find the structure of a given text and use the structural information to build a text processing system. Indeed, if one can find such a structure or structures, particularly the rules that govern these structures, and if such structures can be described with a minimal theoretical baggage and as unambiguously as possible, then operationalising such structures would not be much of a problem, at least in a formal sense.

We believe that until people can find these structures and the rules governing these structures, we will focus, like many others who investigate text, on how texts are organised. The word organisation is used here as a weaker term, and also a broader term, than the much used term structure. The search for organising principles that give a collection of words and the accompanying graphetics and pictures, etc., the status of a text, is not a search for mechanisms that can be used to predict either the form or content of a text, rather this is a search for principles that are used by a writer in writing a text and by a reader in reading the written text.

The AI and IR approaches for generating summaries are by their very nature reductive in that the original text is modified and reduced. We believe that reducing the original text by transforming it raises questions about the credibility of a summary. Reductive text summarisation also poses the question of the credibility of the agent responsible for
producing the summary—a question that is often ignored perhaps at the expense of a growing need to reduce large volumes of text in short periods of time. We argue therefore that cohesion based text analysis and summarisation is perhaps the least intrusive and that summaries based on the categorisation based on text cohesion, but not the reduction of sentences may have more credibility than other summaries.

1.2 Objectives

Our inquiry has two objectives. First, we attempt to test and verify the claims made by text linguistics experts such as Halliday, Hasan, and Hoey. Specifically, we wish to investigate the claims that cohesion in text is a major organisational feature of text and that it is established through ‘repetition patterns’ that range from the syntactic binding of references and pronouns to the semantic linking of collocations and paraphrases. Many of these claims were based on the analysis of short stretches of a text (c. 40 sentences). We intend to extend the analysis to larger stretches of texts including conference papers and a report of the size of a book comprising hundreds of sentences.

Our second objective is to develop a computational framework, that is methods, tools and techniques, to exploit the notion of cohesion, in particular lexical cohesion, for the purposes of text understanding and summarisation. The purpose of this framework is to extract patterns of lexical cohesion from text and study the distribution of these patterns in various text and its effects on text summarisation.

Previous AI work on text understanding, which focused on syntactic analysis of one sentence at a time, raised issues related to ambiguous parsing (i.e. multiple parse trees). Problems like these are the hallmark of early and current natural language processing
systems. We believe that, unless one is equipped with proper and powerful parsers, the study of syntactic cohesion may lead to the same problems encountered in AI. Moreover, recent research reveals that in non-narrative text, the prevalent type of cohesion is that of lexical cohesion and that syntactic cohesion has little effect in this respect. We therefore focus in our study on lexical cohesion in non-narrative text, particularly scientific text, and its role in the automatic summarisation of such texts.

1.3 Achievements

We have developed a system that can be used to examine Halliday, Hasan, and Hoey’s theories related to the existence and purpose of patterns of cohesion, in particular lexical cohesion. We have shown that their claims are valid for even larger texts than those experimented by Hoey (from 100 to 2600 sentences). We have studied the lexical cohesion patterns and confirmed that the distribution of these patterns follows a regular variation in different texts irrespective of the texts’ sizes. We have found out that simple lexical repetition in text, that is the literal repetition of words and terms, constitutes most of the lexical cohesion in text, although other forms of repetition do constitute and do indeed produce a multiplier effect.

We have designed and implemented TELE-PATTAN (TExt and LExical cohesion PATTerns ANalysis) which is a text summarisation system based on the notion of lexical cohesion. TELE-PATTAN is capable of analysing texts for extant patterns of lexical cohesion, using these patterns to automatically categorise the sentences of the text, and subsequently producing text summaries. These summaries contain selections of topic opening, consolidatory, and topic closing sentences, just consolidatory sentences or just the topic opening/closing sentences. The number of sentences in a given summary can
be controlled automatically by defining bond-strength thresholds and then including only those sentences that reach or cross the thresholds.

In a summary evaluation experiment that we have conducted, a number of specialised scientists were asked to read both the original texts and the summaries produced by TELE-PATTAN and then give comments regarding the quality and accuracy of the content of the summaries by answering a number of questions. The results revealed that the summaries were readable and that their contents in terms of subject matter were quite accurate. In some of the summaries, the readers identified more topics than those reported in the abstracts or lists of keywords attached to the original texts. Most of them mentioned that the absence of equations, formulae, tables, and figures from the summaries was not really harmful to the readability of these summaries. The general comment was that the summaries were quite accurate and had, on average, a logical presentation of arguments.

TELE-PATTAN’s relevance to workers in text linguistics, computer science, and information retrieval, we believe can be two fold. First, the tools of TELE-PATTAN allow text linguists to verify the theory of cohesion in text, particularly at the lexical level. Its flexibility in changing the different parameters of the analysis allows one to see the effects of the different types of lexical cohesion patterns on the overall cohesion of text. The tools include also the analysis of the distribution of lexical cohesion link throughout the sentences of a text and the subsequent categorisation of its sentences. Second, the theory behind the design and implementation of this system can be of interest to computer scientists, particularly those working in the area of information retrieval. The text summarisation component can be used to make relevance judgements
in a given context or domain of knowledge. That is to say, the lexical cohesion links, simple lexical repetition or more complex forms of repetition, established through the sentences of a text are used to categorise sentences, say, into those which are more cohesive (i.e. having more lexical cohesion links) and therefore more related to the topic(s) of the text, and those which are less, or perhaps not cohesive at all.

1.4 Outline of the thesis

This thesis is structured into two parts. In the first part that consists of chapters 2 and 3, we describe the theoretical frameworks both in computer science and in linguistics. In chapter 2, we present a review of the major text understanding and summarisation methods and techniques used in artificial intelligence and information retrieval developed and implemented to date. Chapter 3 is an introductory understanding of the notion of cohesion in text. We focus particularly on lexical cohesion and establish its the role in text understanding and, most importantly, in text summarisation.

This leads us to the second part consisting of chapters 4, 5, and 6 where we describe in detail the architecture of our methodology and the results that we obtained from the analysis of a number of texts. We illustrate in chapter 4 the different phases of the analysis and the different methods for producing text summaries based solely on the existence and distribution the lexical cohesion patterns in text and on sentence categorisation. We present in chapter 5 the analysis results of five texts of various sizes. We discuss in these case studies the distribution of lexical cohesion bonds and the production of text summaries. We also present the results of a summary evaluation experiment. Finally in the chapter 6, we conclude the results of our research work and our contributions to computer science and linguistics and discuss further research issues.
Chapter 2

Programs for Understanding and Summarising Text - A Review

One of the interests in artificial intelligence, for more than two decades, has been natural language processing which focuses on how to understand, process, analyse, and generate natural language using automated techniques. This research which is heavily influenced by linguistics has given birth to what is now known as computational linguistics. Natural language processing (NLP) stands for the study of the structures of language and the development of computer models that would emulate the human cognitive linguistic power: the ability to understand and generate natural language text and utterances based solely on the information already possessed. Over the last 25 years, natural language understanding (NLU) research works have largely been concerned with syntactic and semantic analysis of sentences. Some of these works have focused on developing natural language front-ends for answering queries to database systems and others on automatic translation of textbook documents (e.g. technical journals, legal briefs, financial statements, etc.).

Automatic text summarisation is an example of NLU in which computers are used to try to transform pages of large documents into paragraphs and paragraphs into sentences—a process commonly known as text summarisation. A text summary can be regarded as the result of taking the body of information and reducing its size to a set of coherent sentences. The purposes of reducing the size of information vary significantly depending on the objectives and perhaps the means of summarisation. Shapiro reports that, given the claim that thematic understanding (i.e. theme or topic identification) plays a
significant role in the global understanding, the methods and techniques of text summarisation provide a means to explore such a claim (1991). Summaries (of long documents) are available in one or all of the following types of text fragments: abstracts, keywords, epitomes, overviews, abridgements, digests, and recapitulations. The interest in thinking of automatic text understanding and summarisation is mainly due to the growing size of resources of machine-readable texts in many disciplines and domains of knowledge.

In this chapter, we present a survey of some of the automatic text understanding and summarisation systems reported in the literature. In particular, we will discuss the early SAM and PAM—the so-called story understanding, slot-filler systems and the SMART system which is the subject of a 30 years on-going project primarily meant for automatic text processing and transformation using the information retrieval statistical methods. We will then discuss some cohesion-based text understanding systems, like the TOPIC system, Morris and Hirst's discourse segmentation system that uses lexical chains, and Hearst's TextTiller. In addition to these, we will discuss briefly some other systems that have used other methods for generating text abstracts. We will conclude by presenting a brief comparison between our strategy for generating text summaries using lexical cohesion analysis and the others discussed in this chapter.

2.1 Conventional Artificial Intelligence approaches to text understanding
One of the important issues in artificial intelligence is knowledge representation. Frames (Minsky, 1975), scripts (Schank & Abelson, 1977), conceptual graphs (Sowa, 1984) are all examples of knowledge representation schemata. Among these, the script-based
knowledge representation schema was widely used in the early attempts to automate natural language understanding.

Scripts, it is claimed, represent ‘memory structures’ that organise knowledge about stereotypical situations such as catching a bus, dining in a restaurant, attending a lecture, etc. Script-based structures contain a stereotypic sequence of actions and culturally shared events and settings (or locations of events) that occur in a stereotypical situation (Schank, 1980).

Basically, a script is defined as a structure that describes appropriate sequences of events in a particular context that is made up of slots and requirements about what can fill those slots. The scripts of a given situation or context are interconnected in that the content of one slot can affect the content of the others. Scripts are used to ‘handle stylised everyday situations’ that usually do not change and that are characterised by a predetermined, and stereotyped, sequence of actions in a situation that is commonly well-known (Schank & Abelson, 1977:41). Examples of some situations that can be described using scripts include the following: going to a restaurant, riding a bus, attending a party, etc. In order to cover most of the events in a given situation, say in a restaurant for instance, scripts must contain a large amount of information and an enormous variability of what can happen while having a meal in a restaurant. Consider, for instance, a situation where a teacher hands over a paper to a student (e.g. a teacher has given a paper to a student). A script that corresponds to this situation is shown below:
The keyword ATRANS represents the name of the script and the keywords ACTOR, OBJECT, FROM, and TO are the slots needed to be filled to describe the act of ‘transferring’ the paper from the teacher to the student. The structure of scripts that allows the inclusion of events, locations, and scenes was the result of efforts to overcome the problems encountered in the early stages of machine translation (MT). These problems were noted in the fact that MT systems lacked knowledge about the world. The importance of the scripts method is emphasised by the fact that it can be used for performing complex tasks such as pronominal resolution, word-sense disambiguation, and in supplying inferences. A number of text understanding systems were built using scripts including SAM—the script applier mechanism, PAM—the plan applier mechanism, FRUMP—a fast reading, understanding, and memory program, and others. We review the systems SAM and PAM in the rest of this section.

2.1.1 SAM

SAM (Script Applier Mechanism) is a script-based system built for investigating the knowledge of context and how this knowledge can be used to understand stories (Cullingford, 1986). SAM processes stories in English, a sentence at a time, in one pass and makes a number of inferences as it scans the sentence. The depth of understanding that SAM has achieved is demonstrated by its ability to generate natural language summaries or paraphrases of a given story during the process of answering questions about the story. There are two main components of the scripts in SAM: episodes and turning points. Episodes are collections of events linked into a causal chain where each
event triggers one or more events and the turning points are places in the script where several actions might follow. SAM reads the story in a top-down process where each script provides prestored expectations about what will be read based upon what has already been read. The basic strategy that SAM uses in reading and understanding newspaper stories consists of locating the events mentioned in the story and their situation(s). The structure of the situation (i.e. the script) is used to indicate how the chain of episodes should causally be connected.

Although SAM has a limited capability for expressing the time, the location, and the number of events, it produces summaries that are cohesive enough for a human subject to understand. To overcome such limitations, an extended schema based on scripts and known as Plans was used (Wilensky, 1986). The basic idea behind plans is that it is important to know the motivations and intentions of the actors in an event. In other words, it is important to know the plan pursued by the agent performing an act. This extension to the original scripts added more flexibility in that one can make inferences such as the type of plan being followed in the execution of events in a given situation, the type of plans likely to follow in future events, and facts about the actor(s) involved in such events.

To illustrate the use of scripts, consider an example of a story about a person going to a city for a meal. The example is adopted from Cullingford (1986).

*The story (66 words)*

Friday evening a car swerved off Route 69. The vehicle struck a tree. The passenger, a New Jersey man, was killed. David Hall, 27, was pronounced dead at the scene by Dr. Dana Blauchard, a medical examiner. Frank Miller, 32, of Foxon Road, the driver, was taken to Milford Hospital Flanagan Ambulance. He was treated and released. No charges were made. Patrolman Robert Onofrio investigated the accident.
A SAM's summary of the story (51 words)
AN AUTOMOBILE HIT A TREE NEAR HIGHWAY 69 FOUR DAYS AGO. DAVID HALL, AGE 27, RESIDENCE IN NEW JERSEY, THE PASSENGER DIED. FRANK MILLER, AGE 32, RESIDENCE AT 593 FOXON ROAD IN NEW HAVEN, CONNECTICUT, THE DRIVER, WAS SLIGHTLY INJURED. THE POLICE DEPARTMENT DID NOT FILE CHARGES.

Note that the length of the summary is not significantly different from that of the original story. Perhaps, due to this slight physical difference between the length of the text of the story and that of its summary, extracts produced by using scripts are often referred to as 'paraphrases' rather than summaries. In fact, a script-based paraphrase of a story can be longer than the story itself. Note, for instance, the expansion of the details of the driver Frank Miller in the story above where the full address is added from, perhaps, a prestored knowledge-base, or from a previous story where the same driver was mentioned.

SAM works by analysing the sentences in the text and representing them using a conceptual dependency grammar which is based the conceptual dependency theory—a theory of the representation of the meaning of sentences using a number of rules (or axioms)¹. These representations are brought into memory if they fit into scripts. The succeeding input sentences are analysed and the results are inferred from the scripts in memory. Such inputs may invoke other scripts as the story continues. Typical scripts that may be invoked in the car accident story above are 'car crash', 'ambulance ride', 'hospital treatment', and 'police investigation'.

¹ The axioms are: a) for any two sentences that are identical in meaning, regardless of the language, there should only one representation; b) any information in a sentence that is implicit must be made explicit in the representation of the meaning of that sentence; c) the meaning propositions underlying language are called conceptualisations which can be active or stative; d) an active conceptualisation has the form Actor Action Object Direction (Instrument); d) a stative conceptualisation has the form: Object (is in) State (with Value) (Schank & Abelson, 1977:11-12).
SAM's most important task in text understanding is to invoke the right scripts and fill their slots with the right words and phrases that are either explicitly mentioned in the text or are expected to be. It attempts to locate causally connected chains of events mentioned in the story. For example, while processing the car accident story, SAM connects the input concerning the crash with the one about the person in the hospital. This type of connection allows SAM to infer the fact that the driver was injured—something that is not explicitly stated in the original story.

SAM's knowledge-base of scripts describes almost any situation, particularly those found in stories that are reported in newspaper articles. The understanding is, therefore, domain specific and subject to the existence of scripts that match the events described in a story.

2.1.2 PAM

PAM—the Plan Applier Mechanism—is a system that uses plans to understand short texts. PAM's knowledge is about the kinds of actions, the plans, and the goals that people have and the relations of these plans and goals to the events described in a text. PAM, which was a successor to SAM, included mainly the idea of goals and plans in text understanding. It can be used to paraphrase text according to the points of view of the different characters (i.e. actors) in the text. The use of plans, in addition to scripts, has improved the ability to understand the coherence of the story due to the fact that plans provide valuable links between the events of the story—something that scripts alone could not do. PAM, also known as the 'naive explanation algorithm' (Wilensky, 1986), offers more flexibility to text understanding than previous systems based on
scripts and frames in that it has less rigidity in representing events and actions of a story in a text.

Consider the following story paraphrasing example illustrating the understanding of plan and goal relationships from Cullingford (1986):

The event
John and Mary were married. Then one day, John was killed in a car accident. Marry had to get a job.

Question: Why did Marry need employment?
PAM's answer: John died and so she needed a source of money.

PAM's plans include knowledge about goals, themes, and plans in order to determine the roles and intentions of the characters. Its task is to identify the different characters in the story, their goals, and the actions taken to achieve such goals. The explanation algorithm of PAM is based on the ‘points theory’ in which a story is thought of as a set of points that comprise the important content of the text (Wilensky, 1986). Beneath the level of points, there is a level of events that are connected with the major actors in the story but they are in themselves points. Consider the example illustrating PAM's explanation of goals and paraphrase generation from Wilensky (1986).

The event
Willa was hungry. She picked up the Michelin guide and got into her car.

Question: Why did Willa pick up the Michelin guide?
PAM's answer: Because Willa wanted to know where a restaurant was.

Question: Why did Willa get into her car?
PAM's answer: What were the consequences of Willa picking up.

Question: The Michelin guide?
PAM's answer: This enabled Willa to read the Michelin guide.
PAM's paraphrase of the story from Willa's point of view:
I wanted to get something to eat, but I did not know where a restaurant was. So I picked up the Michelin guide and I got into my car.

The development of SAM and PAM, which were first tested as question answering systems, was a major contribution to language understanding, although limited by virtue of Schank's assumption that there are 14 types of acts that can be used to represent all possible events in any situation. Furthermore, SAM and PAM, as two of the first automatic text understanding and summarisation systems, brought to the surface the features of language understanding and communication like anaphora, metaphor, and quantification. Needless to say that these features so frequently used by humans become problems for computers.

There is a variety of other script-based systems that were built to understand stories and to answer questions in natural language such as FRUMP (DeJong, 1979) and QUALM (Lehnert, 1986). SAM, PAM, FRUMP, and QUALM are all based on the notion of scripts and plans.

The NLU systems that have been developed by AI researchers in an attempt to understand natural language varied in the way they dealt with problems in knowledge representation. However, they all tended to explore the many ways of linking sentences, and consequently, events occurring in a given chunk of text or a story. Natural language processing research unfortunately does not concern itself with the cohesion of text at both the understanding and generation level of natural language. However, this should not detract us from noting that knowledge representation formalisms such as scripts, plans, and frames have lead to bringing AI researchers, psychologists, and now linguists together in the quest for efficient and accurate models of human cognition.
2.2 The SMART text retrieval system

SMART is a text retrieval system developed and refined at Cornell University by a team lead by Gerald Salton over the last 30 years (Salton, 1989; Salton et al., 1994a and 1994b). SMART is based on the vector-space model for representing documents and queries. The document and query vectors are lists of weighted terms extracted from the document or the query itself. Terms can be single words, compound words, or phrases that are chosen primarily from a lexicon, a thesaurus, or ideally from a large terminology database. A method for weighting terms is to use the well-known equation (term weight $= \frac{f}{f_c}$) where $(f)$ is the frequency of a term in a given document and $(f_c)$ is the frequency of this term in a collection of documents, i.e. a corpus (Salton, 1989). To retrieve a document in response to a given query, the term vector of the query is compared with the term vectors of all documents in a corpus and, based on a similarity threshold, some documents are retrieved and others are ignored.

Using the notion of document similarity, SMART can be used to represent text as a graph: the so-called ‘text relation map’. In this map, the nodes represent different text structures, e.g. chapters, sections, or paragraphs and the links are relations that stand for existing similarities between these different text structures. The main purpose of this map, as Salton argues, is to identify the theme or themes of a text by merely looking at possible gaps in such a map. These gaps represent a change of theme, a citation, or a related theme. The number of gaps therefore is an estimate of the themes present in a text and therefore helps to retrieve the gist of the text that should cover most of its theme(s).
Salton (1989) has argued that the lack of 'coherence' in the extract or summary of a text that is based on the notions of similarity and relevance is partly due to the word-frequency oriented metric used in passage retrieval systems such as SMART. He suggests that three important heuristics should be used in a text abstraction program. First, those sentences that contain specific reference to the status of work reported in a text, like 'the present research' or 'our work', should be automatically included in the abstract. Second, sentences referred to in earlier passages should be included in the extract, like sentences containing 'as described above', etc. Salton refers to this kind of reference as 'syntactic-coherence consideration'. The third heuristic relates to the location of a sentence. The first sentence of a paragraph, for instance, should always be included in the abstract.

Consider an example on the use of conventional document similarity measures for text summarisation (Salton et al., 1994a). The example is about summarising an article numbered “16585” on “Horatio, Viscount Nelson”. The similarity-based paragraph map for this article is shown in Figure 2-1. In this figure, the letter ‘p’ indicates the paragraphs of the article. The dashed path traverses all 'bushy' nodes which are defined as nodes that have at least six incident links. In this example, a link between two paragraphs is valid if the similarity between them is more than 0.20. The lines shown in the figure connect paragraphs which have similarities above 0.20. Other nodes corresponding to paragraphs with lower similarities are consequently not shown.
The bushy nodes in the figure above which are connected through a dashed line are paragraphs p3, p6, p9, and p11. A summary of the article consists of all the paragraphs joined by the dashed line, namely, paragraphs p3, p6, p9, and p11. A more detailed summary would include other paragraphs with a lesser number of incident links like paragraphs 7, 8, and 12. Salton argues that when the text relation map is substantially disconnected, the text traversal does not produce comprehensive summaries. This means that gaps in the map are undesirable for the production of summaries based on paragraph similarity measurements.

Although judicious text extraction methods are used in large systems like SMART to identify the themes that cover almost the totality of a textual document, Salton and his colleagues admit that, in the absence of deeper linguistic analysis methods, i.e. methods that use knowledge about the syntax and semantics of the language in which the text is written and that are applicable to unrestricted subjects areas, it is not possible to produce 'intellectually' satisfactory summaries (Salton, 1994a:1425). An intellectually satisfactory summary, according to Salton, is a summary that is cohesive and readable.
One should also note that text decomposition in SMART which goes to the level of paragraphs is, perhaps, the most attractive aspect of SMART.

2.3 The TOPIC Text Condensation System

TOPIC is an example of text understanding systems that uses a large knowledge base (Hahn & Reimer, 1985; and Hahn, 1990). TOPIC adopts frames as a knowledge representation schema and is used to perform deep syntactic and semantic text analyses using 'macro textual coherence methods' and 'micro textual cohesion considerations'.

In addition to text analysis, TOPIC covers a variety of text 'condensation modes' and supports many options for text retrieval (Reimer & Hahn, 1990). It can be used, for the extraction of facts and the acquisition of new concepts—two important issues in knowledge acquisition. It can also be used to create associations between thematic descriptions and relevant text. Such associations allow the retrieval of text fragments from a full-text database using corresponding passage descriptions. These text passages, particularly those that share thematic descriptions, can also be linked creating a navigational hypertext representation of the text. An important aspect of TOPIC which is of interest to us is text summarisation (or text condensation). TOPIC helps to generate a representation of topics in a text which varies from the generic and thematic level to more specific ones. This system performs shallow text understanding that is based on partial parsing of the source text and a graphic representation of the text condensates. The partial parsing which consists of recognising the thematic foci of the text and the significant facts related to them is realised by restricted text analysis and taxonomic knowledge representation using frames. The recognition of thematic foci involves continuous activation of slot-filling operations for various frames that are
predefined in a knowledge base. The condensation process transforms the taxonomic knowledge representation into various levels of descriptions of thematic specialisation. The result is a hierarchical graph of nodes and links. Each leaf node of this graph represents a topic description of the themes in a cohesive text passage. The nodes at higher levels comprise more general descriptions.

The graphs of a full text are combined to reduce redundancy giving rise to a global semantic network that reflects a hierarchical topic description of the text in which a variable number of thematically coherent units appear. The root nodes in the semantic network represent the most abstract and general description of the content of the corresponding text.

The construction of 'condensates' of text in TOPIC is based on the distribution of activation weights that trigger slot-filling operations for different single frames in the knowledge base, on the particular connectivity patterns that hold within groups of frames, and on the identification of significantly dominant concepts in the text. Consider the following example from Reimer and Hahn (1990) showing a text that describes a specific type of personal computer: the Zenon-X machine. A TOPIC graph corresponding to the following text is shown in Figure 2-2.

"The text passage is about personal computers. The Zenon-X is discussed in more details with respect to its CPU and its peripheral devices. Besides disk drives in general, the floppy disk drive which is available for the Zenon-X is focused on..."
In the figure above, the label 'identity' represents a link between superordinate and subordinate nodes and the label 'instance-of' represents an 'is-a' relationship between nodes. The other links denoted by '—slot—', '—s—', '—sf—' stand for slot and slot-filler links. The nodes of the text graph are associated with relevant text fragments that comprise the slot fillers and represent a variable number of thematically coherent units. The root nodes cover the most abstract and concise characterisation of the text whilst the leaf nodes give more specific descriptions of the topics or themes that the text covers.

One notes, therefore, that the idea of text condensation in TOPIC has more to do with knowledge acquisition than with text summarisation in that the system can be used to
extract a collection of text fragments where certain themes are described. The authors of TOPIC consider such a collection as an abstract of the text. TOPIC's condensation which relies on the use of a rich knowledge base does not modify the original structure of the text fragments. It can be thought of as an indicative abstract generation method in that it can be used to indicate the relevant text where a given description of a theme or a topic is likely to be found.

2.4 Lexical Chains

Morris and Hirst (1991) have used cohesion patterns in a text to extract 'lexical chains'. The cohesion patterns are the effect of the repetitions of words that authors often use explicitly or implicitly throughout a text for the purposes of emphasis, elaboration, or explanation of ideas, thoughts, and arguments. The lexical chains are lists or successions of related words spanning a 'topical unit' of a text and can therefore be thought of as collocations. The relations between words in these chains are lexical semantic relations and comprise synonymy, antonymy, meronymy (part-of), and hyponymy (kind-of). Most of the semantic relations between words of a lexical chain can be identified using a dictionary and/or a thesaurus. Furthermore, words in lexical chains are related by both lexical cohesion relationships and 'distance' relationships which can be within a given sentence or across sentences within a text. The distance relationship constitutes the difference between lexical chains and the concept of collocations in which collocating words must be adjacent (i.e. no distance between them). Consider the following example from Morris and Hirst (1991):

In front of me lay a virgin crescent cut out of pine bush. A dozen houses were going up in various stages of construction, surrounded by hummocks of dry earth and stands
of precariously tall trees nude halfway up their trunks. They were the kind of trees you might see in the mountains.

The highlighted words in the text above constitute an example of a lexical chain. The list of words \{virgin, pine bush, trees, trunks, trees\} indicates that the author intends to bring into focus a given situation, an event, or a phenomenon that is perhaps directly or indirectly related to trees, forests, or wood industry. The decision of which of these themes the author has in mind would require looking at other and, perhaps, longer lexical chains. In the example above, the list of words \{house, pine bush, trees, trunks\} appears to be describing part of a topic on wooden houses construction.

Halliday and Hasan have argued that pairs of lexical items that co-occur in a text constitute 'lexical sets' and contribute to the cohesion of the text. These lexical sets can therefore be thought of as 'cohesion chains' (1976:286). Morris and Hirst argue that lexical chains, which actually constitute a form of cohesion chains, are important for two reasons. First, every lexical chain in a text represents a word interpretation in context by virtue of the fact that such a word belongs to such a lexical chain. Second, the extraction of lexical chains from a text will help in the identification of the discourse structure of the text.

Lexical chains can be extracted using thesaural relations, transitivity of word relations, and distance (in sentence units) allowable between words in a chain (Morris & Hirst, 1991). Morris and Hirst argue that the strength of a lexical chain and its use in the interpretation of discourse structures depends on three parameters: the number of reiterations of a word, the density, and the length of the chain. Their claim is that lexical
chains provide a good indication of the segment boundaries that are assumed in Grosz and Sidner's theory (1986).

Grosz and Sidner (1986) emphasise the idea that there are three interacting components for a common discourse structure: the linguistic structure—that by which text is recognised as an ordered sequence of sentences, paragraphs, sections, etc., the intentional structure—the intention and purpose in engaging in discourse, and the attentional state—basically a number of entities, concepts, or notions that the attention is focused on at a given point in discourse. Based on this idea, Morris and Hirst claim that, within a given text, there is a tendency to use related words and that these words, if grouped together, would constitute lexical chains that can be used to indicate the discourse structure. They attempt to prove that when a lexical chain starts, it indicates the beginning a new linguistic segment and the end of another, and that, as a consequence of such an indication, one is able to establish the topic(s) of the segments in the text that correspond to the chain.

Morris and Hirst's demonstration of the use of lexical chains was based on the use of a small number of texts and a prototype thesaurus similar in structure to that of the Roget's Thesaurus (Roget, 1972). Although most of the lexical chains extraction was performed by hand, due to a non-availability of a machine readable version of the Roget's Thesaurus, the results indicate that topics or themes of a text are identified by drawing potential word-to-word relations from a thesaurus. However, relying fully on the thesaurus and limiting the distance between words of a chain is not free of difficulties particularly in cases where lexical chains span across paragraphs and therefore creating overlaps with other chains. One can think of solutions to these problems by assuming
that a chain will only span one paragraph at a time, or alternatively, one can take into account the words that appear in headers of paragraphs or sections as pilots for buildings chains.

Our work is different from Morris and Hirst's work, which was carried out almost entirely manually, in two aspects. First, we emphasise in our work the study of lexical cohesion patterns in text which, we believe, embodies that of lexical chains. Second, we have developed a computational methodology for the analysis and summarisation of text based the notion of lexical cohesion.

2.5 Cohesion index analysis

Stoddard (1993) has attempted to establish and study the relationships of texture, repetition patterns, and cohesion in written text. This research is based on the assumption that readers derive more meaning from their reading than the sum of words printed in the text. This would mean, argues Stoddard, that readers establish a kind of synergy which is partly due to the presence of various kinds of patterns that in turn give a text its texture. Stoddard claims that such patterns suggest a holistic view of text.

Stoddard’s work is a comparative study on cohesion patterns found in different texts in a variety of disciplines including non-fiction, essays, short stories, and biographies. The author has developed for this purpose a system that performs a number of tasks. First, the system helps to define appropriate terms that identify each word within a text as a node, a cohesive element, or 'a general word', and count the total number of words. Second, the system transforms the text into a two dimensional matrix containing words and links between them. Stoddard refers to a pair of words that represent cohesive
elements as a 'link'. The rows of this matrix are the actual lines of the text and the columns are the words in the lines. Words here are assumed to be six letters long and the lines twelve words long. The resulting table or ‘map’ is a representation of a cohesion networks in each text segment. Third, the system performs a statistical analysis including a study of a number of parameters such as the total number of words, the total number of nodes, and the total number of occurrences of cohesive elements.

Some of the results that Stoddard obtained from the analysis of different texts are shown in Table 2-1 below. In this table, there are three types of cohesive elements: definite articles which signal noun phrases such as the, pronouns such as he, they, she, etc., and agent displacement elements which Stoddard defines as ‘ed’ and ‘ing’ morphological variants.

<table>
<thead>
<tr>
<th>Texts</th>
<th>Def CE/Words</th>
<th>Def FCE/Words</th>
<th>Pro CE/Words</th>
<th>Pro FCE/Words</th>
<th>AgD CE/Words</th>
<th>AgD FCE/Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-fiction</td>
<td>14.40</td>
<td>11.84</td>
<td>4.11</td>
<td>2.97</td>
<td>4.00</td>
<td>2.21</td>
</tr>
<tr>
<td>Essay</td>
<td>14.84</td>
<td>12.56</td>
<td>6.50</td>
<td>5.84</td>
<td>3.32</td>
<td>1.99</td>
</tr>
<tr>
<td>Biography</td>
<td>13.10</td>
<td>9.74</td>
<td>7.89</td>
<td>7.59</td>
<td>3.40</td>
<td>2.51</td>
</tr>
<tr>
<td>Novel</td>
<td>12.68</td>
<td>8.74</td>
<td>10.48</td>
<td>9.60</td>
<td>2.73</td>
<td>1.83</td>
</tr>
<tr>
<td>Short Story</td>
<td>11.22</td>
<td>8.54</td>
<td>10.95</td>
<td>10.50</td>
<td>2.63</td>
<td>2.08</td>
</tr>
</tbody>
</table>

Table 2-1: Number of cohesive elements relative to the length of the text in (%) (where Def=Definite Article Cohesion, Pro=Pronominal Cohesion, AgD=Agent Displacement Cohesion, CE=Cohesive Element, FCE= Fulfilled Cohesive Element), taken from Stoddard (1993).

Stoddard does not mention other types of cohesive elements such as synonyms and explicit repetitions in her comparative analysis. And, although her analysis says little about the effect of cohesion patterns on the overall topic(s) of the text, it reveals interesting results about the distribution of the different cohesive elements in texts of different genre. In particular, pronouns, according to Stoddard, have less effect on text cohesion as one moves from short story texts to non-fiction texts such as scientific text.
Furthermore, Stoddard’s analysis appears to focus on syntactic cohesion elements and not lexical cohesion which, as we will demonstrate in the rest of this thesis, has a significant effect on the overall cohesion of text.

2.6 SERAPHIN - a French text abstract generator

SERAPHIN is a text abstraction system built as part of a project aiming at extracting ‘outstanding sentences’ (Berri et al., 1995). Given the variety of domains that can be encountered in arbitrary text, Berri et al. do not make use of any conventional knowledge representation schemata. Instead, the authors use ‘phraseological markers’ and refer to these markers as ‘pertinent linguistic indicators’ (PLI), and ‘contextual linguistic indicators’ (CLI). These markers are used to identify ‘important’ sentences and to associate ‘importance scores’ with such sentences. The phraseological markers are used to explore the ‘linguistic knowledge’ rather than the ‘domain knowledge’. Consider for instance the sentence:

\textit{It is necessary to emphasise that our research focuses on text summarisation.}

Here, the verb \textit{to emphasise} plays the role of a PLI while the phrase: \textit{It is necessary}, is an example of a CLI. In SERAPHIN, this sentence is considered potentially important and an importance weight (or score) is associated with it. The scoring is used to extract all the important sentences in a text in a specific importance order. Furthermore, SERAPHIN uses graphetic information such as font types, highlighting, underlining, etc. for computing the importance of a text fragment.

The two indicators, CLI and PLI, that SERAPHIN uses to produce sentence scores can be single words, compound words, and phrases, particularly thematic noun phrases (i.e.
noun phrases that relate to the topic or theme of a text in their wording). In fact, the list of indicators is a semantic grouping expected to cover the potential themes or topics of the text. It is, therefore, necessary to use either a dictionary or a thesaurus to recognise such groupings. Having said so, it is not clear to us how such groupings are achieved in SERAPHIN. An interesting aspect of SERAPHIN is its ability to use linguistic cues that authors often use to convey and emphasise arguments, e.g. *it is necessary to note that* ..., to refer to previous text, i.e. anaphoric and elliptic references; and to explicitly anticipate conclusions and briefings, e.g. *we have shown that...*, *we conclude our discussion...*, etc. These cues are used in computing the so-called 'importance scores' for sentences and, subsequently for creating text abstracts.

Although SERAPHIN's structure is based on the use of linguistic cues and not on any kind of linguistic resources such as dictionaries, or thesauri, or terminology data bases, we believe that its ability to use graphetic information is, indeed, one aspect of text understanding that should not be ignored. Furthermore, one can see an author's tendency to structure a text, emphasise or de-emphasise different points by using a variety of font types. With the use of mark-up languages such as the Standard Generalised Mark-up Language (SGML) (Jones, 1991) to indicate and mark explicitly the structure, content, and graphetic styles of a text, it is now possible to create computer programs that are able to parse such a text and extract various information for a number of applications, e.g. intelligent help extraction systems from on-line technical documentation.
2.7 The Text Tiling and discourse segmentation system

Statistical processing of text in information retrieval over the last 30 years has had a important impact on text linguistics research in that it has raised many open questions on how authors choose terms to describe a given topic (or topics) in a text without any ambiguity nor any discontinuity. The concept of term weights on which almost all information retrieval methodologies are based is a stochastic description of how authors write text. Although free of any sort of semantic content, term weighting has proved its efficacy in identifying relevant stretches of text amongst a lengthy text and/or a large number of texts.

One of the systems that has made use of the term weighting mechanism for performing text understanding tasks is the ‘TextTiller’ system (Hearst, 1993). TextTiller is a system that uses document similarity (Salton, 1989) to partition full-length text documents into coherent multi-paragraph units in order to reflect the pattern of subtopics and themes contained in a text. This tiling operation which can be thought of as a computer-based text segmentation into uniform and non-overlapping discourse fragments uses lexical analysis based on the \((t_f \cdot id_f)\) term weight measurement method. The intuition that discovering a text’s structure by dividing it into sentences, paragraphs, or adjacent chunks of text, and looking at how much word-overlapping occurs amongst these chunks (Skorochod’ko, 1972) is at the heart of text tiling. The simplicity of such a methodology is in its relatively easy computability. The text tiller algorithm is a two pass analyser. In the first pass, the similarities between every two directly adjacent text blocks (where text blocks are sets of a given number of consecutive sentences that form the text) are computed using the document similarity measure as described below. The weights are defined as
\[ w_{t,b} = t_f \cdot id_f = \frac{\text{Frequency of term } t \text{ in a text block } b}{\text{Frequency of term } t \text{ in the whole text document } d} \]

and the similarity between two blocks \((b_1, b_2)\) is given by the following formula:

\[
\text{sim}(b_1, b_2) = \frac{\sum_{t=1}^{n} w_{t,b_1} \cdot w_{t,b_2}}{\sum_{t=1}^{n} w_{t,b_1}^2 \cdot \sum_{t=1}^{n} w_{t,b_2}^2}
\]

where \(n\) is the total number of terms in a document (Hearst, 1993). The comparison of all text blocks shows different similarity values as shown in Figure 2-3 where the similarity is plotted against the text block numbers. A high similarity between two directly adjacent blocks indicates that the two blocks cohere well and form therefore a ‘tile’. Low similarity values indicate the gaps between blocks. High similarity values form peaks giving rise to tiles while low similarity values form what Hearst calls ‘valleys’ that indicate cuts in topic continuity or jumps from one topic to another.

Figure 2-3: The text tilling methodology adopted from Hearst (1993).
Although the text tiling method has a granularity that operates at the level of sentences as opposed to conventional information retrieval systems (where the granularity stops at the level of documents and paragraphs), text tiling consists only of breaking text into contiguous ‘tiles’ that reflect the potential topical loci but not the possible existing relations between these topics. Perhaps, if the method were based on the similarity between, not only adjacent blocks, but also non-adjacent blocks, one could measure all mutual similarities between the text blocks and rank them accordingly. The similarities between blocks of texts could then be used to build a text map (similar to the one used in the SMART system) where all similar blocks have weighted links. The links in the map and their density could be used to represent the relations between blocks of a text, and subsequently between its topics or themes.

2.8 Conclusions

The past and current text understanding systems use different approaches: AI based systems rely mostly on pre-defined frames or scripts and the analysis process means filling the slots in frames and scripts. The granularity of these systems is poor in that they tend to tackle a restricted instance of an event, i.e. a newspaper crime story, an outing for a dinner party, etc. They are therefore less concerned with the words in the text.

The SMART system is a word/phrase based text processing system in that the document similarity is measured by pragmatic parameters: the so-called term weights. Like any other information retrieval systems, SMART’s granularity does not go beyond the document and paragraph levels. The TextTiler, however, is an attempt to go to the level
of sentences by assuming that every sequence of sentences (i.e. a text block) is a document on its own and that retrieving sentences can be thought of as retrieving documents. Both SMART and TextTiller are primarily information retrieval systems and have little if no concern for the semantic and syntactic knowledge inherent in the text.

SERAPHIN relies on the presence of graphetic information and the use of phraseological markers within a text assuming, therefore, that every important sentence must have at least one these markers—something that is not always explicit in a text. It does however point out that graphetics like highlighting, and choosing distinctive type fonts, etc., which can be easily detected, are potentially interesting indicators of the authors intentions in text.

TOPIC, the cohesion index analysis, and the lexical chains systems, all rely on the notion of cohesion of text. In these systems, sentences, paragraphs, as well as words are all objects of the analysis. But most importantly, the relations that are established between pairs of words or phrases are drawn from a dictionary or a thesaurus. We have seen that these systems can be used to point at relevant passages of text or identify discourse segments. However, it is not clear to us how the TOPIC’s themes interact with each other, or how lexical chains can be used to reduce information and produce summaries.

Our text summarisation strategy involves the extraction of not only chains of cohesive words and phrases (or fragments of texts) that introduce or close topics, but whole summaries of texts at various levels of its structure, i.e. at the level of words, sentences, paragraphs, sections, and chapters of a text. We have developed a system based on
cohesion analysis for text analysis and summarisation—the TELE-PATTAN system. This system can be used to analyse the patterns of cohesion in a text and use the distribution of these patterns to categorise sentences in relation to how important they may be to the topic of the text and, subsequently use these categories to automatically generate text summaries.

Table 2-2 is an illustration of some of the systems that we reviewed and the approaches used to analyse text for the purposes of understanding, summarisation, and/or retrieval. We also illustrate in the figure the relation of our approach, i.e. through the use of TELE-PATTAN, to the other approaches discussed in this chapter.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Tele-PattAn</th>
<th>TOPIC</th>
<th>TextTile</th>
<th>SMART</th>
<th>SAM/PA</th>
<th>Lexical Chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representation (frames/scripts/plans)</td>
<td>x</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>Resources (dictionaries/thesauri)</td>
<td>√</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Methods (Frequency distribution, Relevance Judgements)</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>√</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Linguistic levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>syntax</td>
<td>√</td>
<td>√</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>semantics</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>√</td>
<td>x</td>
</tr>
<tr>
<td>cohesion</td>
<td>√</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>√</td>
</tr>
<tr>
<td>Text's Structure Levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>words/phrases</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>x</td>
<td>√</td>
</tr>
<tr>
<td>sentences</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>paragraphs</td>
<td>√</td>
<td>x</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>sections</td>
<td>√</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>full text</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>x</td>
<td>√</td>
</tr>
</tbody>
</table>

Table 2-2: Some text understanding approaches and the role of TELE-PATTAN.

Our work is an attempt to reduce the complexity of the ever challenging problem of automatic text understanding and summarisation. We believe that, with the study of cohesion patterns in text and the use of thesaural semantic knowledge, it is possible to overcome some of the problems encountered in text summarisation and retrieval,
particularly those related to the readability and consistency of the generated or retrieved summaries.

We will discuss in the next chapter the linguistic theory of cohesion in text underlying our text analysis and summarisation methodology.
Chapter 3

Lexical Cohesion and its Role in Text Understanding

3.1 Introduction

It is important to study how texts are organised at a lexical level so that this organisation can be exploited from a point of computer-based analysis of texts for the purposes of information retrieval, or abstraction, or understanding. Texts in general, and scientific texts in particular, may be characterised by the way lexical items are repeated throughout the text. The repetition of lexical items, say, words like 'neutrons' and 'protons' in a nuclear physics text, or words like 'record' and 'query' in a database management system text, are used, for instance, to reinforce a certain message, to elaborate a term, to negate an idea, to explain an observation, and so on. These repetitions, as we will see in the following sections of this chapter, range from the very simple, like using the same term or using its plural, to the complex, for example the use of a collocation of a term (cf. particles for protons). We show how to exploit lexical repetition, what a linguist would call a 'cohesion device', to establish the links between sentences in a given text, the so called 'central sentences' that might comprise the essence of a text and thus helps in the summarisation of the text. The analysis of lexical repetition patterns can be used to produce a summary of the text based on central sentences, i.e. those sentences where most repetitions occur.

The study of text is the central feature of a number of branches of linguistics. Generally, this amounts to a description of the language in what is called text linguistics—a branch of linguistics that focuses on text itself, corpus linguistics and lexicography. Any discussion of text must be preceded by a definition of what a text is. Definitions of text include references to any passage of language, spoken or written, of whatever length, that forms a unified
whole. According to Halliday and Hasan, a text is best regarded as a 'semantic unit: a unit not of form but of meaning' (1976:2). These authors define text as a 'stretch of language recorded for the purposes of analysis and description'. Essentially, texts are defined as language units that have a definable communication function and that can be characterised by principles such as cohesion and co-reference. Cohesion is 'a semantic concept' in that it refers to relations of meaning that exist within the text and that define it as a text.

Co-reference occurs when items in a text are not being semantically interpreted in their own right, but rather refer to something else for their interpretation. Items that have such a function in English are personal pronouns, demonstratives, and comparatives. These items constitute directives that are used as 'signals to retrieve information from elsewhere in the text' (Halliday & Hasan, 1976:31). Within any linguistic system, resources such as cohesion and co-reference enable a writer to produce a readable text with a message comprising ideas, thoughts, and arguments. The produced text is a coherent and cohesive 'weave' of sentences that can be related to an identifiable theme within a given domain or context.

The terms cohesion and coherence are often interrelated. Halliday and Hasan refer to a coherent text as semantic unit and to cohesion as a semantic concept. Quirk et al. regard a text as a stretch of language whose coherence depends on the appropriateness of the actual use and is measured by its external relations and consistency with the real world semantically and pragmatically—the so-called cohesion of text (1985). deBeaugrande has argued that cohesion and coherence are text centred notions that designate operations directed at the text materials and can be regarded as operational goals without which the attainment of other discourse goals may be blocked (1991). Perhaps, because of the fact that cohesion is a property and that coherence is a quality, linguists tend to emphasise the notion of cohesion
and its effect on the structure of text and the regulation of textual components more than they do for coherence. We will, therefore, use the term 'cohesion' throughout this thesis with the assumption that coherence is directly related or embedded in the notion of cohesion.

Most of the conventional natural language processing systems focus on sentence analysis whereby the objective is to identify the key roles and actors within each and every sentence. However, if we were dealing with a systematically organised collection of sentences, that is text, then not only should we consider the structure of every sentence, but also the way the sentences are interconnected together to form a smooth and uniform flow of information.

Consider the following paragraph comprising the sentences 4, 5, 6, 7, and 8 from the introductory chapter of a nuclear physics text to which we have added the sentence numbers indicating the physical order in the text (Jackson & Barret, 1977:1-3):

(4) The extent to which we are able to make precise and meaningful statements about the nuclear matter distribution and the nuclear charge distribution and the variation in both quantities from one nucleus to another reveals quite clearly the state of our understanding of much more fundamental issues, such as the nature of the interactions between various types of particles and the role of these interactions in scattering phenomena, the subtle balance between various features of the nucleon-nucleon interactions in bound states, and the difference between the average properties of nuclei described by macroscopic models and the specific nuclear structure properties described by microscopic models. (5) The study of nuclear sizes involves both the study of the nuclear charge distribution by means of processes dominated by electromagnetic interactions and the study of the nuclear matter distribution by means of strong-interaction processes. (6) By combining the information so obtained, comparison of the proton and neutron distributions can be made. (7) Most of the discussion will be devoted to the determination of the radial shape of the distributions in spherical nuclei, but the angular dependence of the shape of nuclei which are not spherical will also be considered. (8) One of our principal aims will be to try to determine and explain precisely which properties of the relevant distributions can be obtained from the various experiments and to indicate the extent to which previously published parameters are really determined by the measurements as opposed to being merely consistent with them.
The single word terms, protons, neutrons, nucleus/nuclei, scattering, and interaction, are used repeatedly in these five sentences, and indeed throughout the pages of Jackson and Barret’s book (1977). The authors also reuse complex terms, like nuclear matter distribution, nuclear charge distribution, proton/neutron distribution throughout the text. The re-use or repeated use of these terms involves the various morphological and semantic variants of these terms: one sees the use of plurals of the single and double word terms, and a range of paraphrases that are used throughout the text, as exemplified by these five sentences. For example, protons and neutrons are paraphrased nucleons and the three terms (i.e. protons, neutrons, and nucleons) are paraphrased as (nuclear) particles, (collectively electrons and nucleons are referred to as atomic particles); nuclear matter and nuclear charge distributions can be referred to as nuclear distributions and paraphrased as distributions in spherical nuclei. The introduction and re-use of a term, its morphological variants (complex repetition), and its paraphrase throughout, or in a portion of a text, helps to make the text cohesive. This cohesion is called lexical cohesion. Any two sentences that contain a term or its variants can be viewed as lexically ‘linked’ or ‘tied’ to each other.

As one peruses a text, more terms are encountered that may repeat in one way or the other. The repetition of some terms helps to make lexical links between sentences, and consequently make the text cohesive and coherent. However, not all sentences may be linked, comments in newspaper reports are examples of marginal sentences that only enhance the idea conveyed by the text but have no effect on the overall meaning if they were removed. In this type of sentences, one finds less repetitions and less re-use of terms than in other sentences in the text. Our comments do not include a discussion of either pronominal references or the use of specific or non-specific determiners. These grammatical categories do help in the establishment of cohesion in text, and such cohesion is termed ‘syntactic
cohesion'. Our interest is in lexical cohesion - the repetition and linking of lexical items, e.g. nouns.

In order to read, understand, and/or modify a given text, a reader is required to have the skill and the expertise of following through given arguments over the length of the text. Indeed, there may be many arguments in a given text. Writing a text requires skills and expertise in building arguments, and sustaining them with evidence, or counter evidence over many sentences. And, it is at that level—the inter-sentential level, that the notions of cohesion and co-reference can be seen to operate clearly.

In this chapter, we present a description of the notion of cohesion in text as defined by experts in text linguistics such as Halliday and Hasan (1976), Hoey (1991), and deBeaugrande and Dressler (1981). We focus particularly on lexical cohesion and its role in maintaining the coherence and meaningfulness of a text. We will show, with the help of some example, how lexical cohesion helps to link a sentence, a paragraph, or a section of a text to another, and how the collection of such links can be used to describe and evaluate the cohesion of the text. We discuss, at the end of this chapter, what we regard as requirements and hypotheses for the development and implementation of a cohesion based text analysis and summarisation system.

3.2 Lexical and syntactic cohesion
Cohesion is defined as those surface structures of a text which link different parts of sentences or larger unit of discourse, e.g. the cross-referencing functions of pronouns, specific and non-specific references using determiners, and some adverbs (Halliday & Martin, 1993). It represents the way certain words or grammatical features of a sentence connect that sentence to its predecessor and/or its successor sentences in a text (Hoey, 1991).
The interconnection between sentences through the use of pronouns, conjunctions, substitutions, repetitions of a lexical item, and so forth, is one way in which an author of a text ensures that information flows through the text.

Fowler has elaborated this notion of information flow by arguing that a text is essentially a 'progressive sequence of propositions' that convey ideas. He argues that 'well-formed' texts can be distinguished from other texts because well-formed texts have three essential features: cohesion, progression, and thematisation. Cohesion enables the author to focus on one topic at a time and to move to another topic in a smooth manner in such a way that ideas and thoughts are coherent and consistent. Progression is a feature of well-formed texts in that the author of such texts generally organises a logical and chronological progression of sentences. Thematisation allows the elaboration of an overall theme that the reader should not find difficult to grasp (Fowler, 1986:61).

Linguists regard cohesion, where sentences are co-ordinated together through 'links' and 'ties' (due to cohesive repetitions between sentences), as the most important feature of text organisation. A text that lacks cohesion will be simply an arbitrary sequence of sentences and makes the understanding of text arduous for the reader.

3.2.1 Syntactic Cohesion

Halliday and Hasan have been cited frequently for their work related to cohesion in general text and in English text in particular. In addition to lexical cohesion, these authors have defined four major syntactic features that are used to make a text cohesive: conjunction, reference, substitution, and ellipsis (Halliday & Hasan, 1976). We briefly discuss each of these features below.
**Conjunction:** The conjunction of sentences may be 'additive' in which case a sentence is an elaboration of the previous sentence, or it may be 'adversative' where a sentence is in a contrastive relation with the sentence before it, using connectors like 'however', 'but', 'yet', and so on. Conjunction may also be 'causal', expressing a 'cause-effect' relationship between two sentences using words like 'because', 'therefore', 'if..then..', etc.

**Reference:** In order to ensure cohesion in text, one can use pronouns in sequences of sentences as a means to save rewriting terms, and to help refer to previously mentioned items or names in previous sentences. This cohesion feature is called 'reference'. Personal pronouns such as I, you, he, she, it, and they, and demonstrative pronouns, e.g. this, that, these, those, and so on, represent frequent examples of reference in text.

**Substitution:** One can also refer to an event, an action, or an entity by a single word, e.g. 'did you study computer science? Yes I did'. Here, the word 'did' replaces a whole sentence. This feature is termed 'substitution'.

**Ellipsis:** Sometimes, parts of a sentence are simply omitted in the following sentence without affecting the overall meaning. Consider, for instance, the sentences: 'Is it raining?' No, it is not'. The word 'raining' has been ignored, but the reader still understands that it is not raining. This kind of cohesion is referred to as 'ellipsis'.

### 3.2.2 Lexical Cohesion

Syntactic cohesion has been discussed at length by Halliday and Hasan (1976: chapter 8). Although syntactic cohesion is as important as lexical cohesion, a systematic study requires an elaborate grammar model of the language of the text—a controversial subject at the best of times. Our intent is to focus on exploring cohesion for manipulating and retrieving...
scientific text, extracting terms from such texts, summarising text, and so on. Lexical cohesion, we believe, is readily exploitable for the tasks we wish to accomplish.

Although Halliday and Hasan allocate only about twenty pages for the discussion on lexical cohesion and the rest totally devoted to syntactic cohesion, instances of lexical cohesion account for more than forty percent of the total number of cohesive ties. Furthermore, the complexities that we may encounter in trying to identify instances of syntactic cohesion are overwhelming due to the complexity of language as a totally versatile system of communication. In this system, every author and every reader has his/her own way of linking ideas and thoughts in the process of generating or understanding text. We have therefore focused our research on the analysis of lexical cohesion and its effect on special language text, particularly scientific text.

Halliday and Hasan argue that lexical cohesion is a collection of semantic relationships between lexical items that they broadly cluster into two sub-classes: reiteration and collocation (1976):

*Reiteration*: This is a form of repetition of lexical items through a number of devices, namely, synonymy, near-synonymy, super-ordinate, and general-word devices which include human nouns, place nouns, fact nouns, and so on.

*Collocation*: This term often refers to the name given to the relationship between two items occurring together with a certain probability within a textual context (cf. Hoey, 1991). Some of the reiterations mentioned above like near-synonymy and super-ordinate frequently occur in the same context. There is some fuzziness in the definition of collocation as a category.
Halliday and Hasan recognise it as the most problematical part of lexical cohesion and resolve the matter by pinpointing that when these reiterations stand in some recognisable 'lexico-semantic (word meaning)' relation, they re-occur, i.e. they collocate. Items that re-occur usually include synonyms, near-synonyms, super-ordinates, antonyms, and therefore form sets of collocations.

Hasan has reassessed the problem of collocation in her later work (1984). She proposes another categorisation where she merges reiteration and collocation and produces two types of lexical cohesion: general and instantial.

Consider for instance the two sentences, respectively the fourth and the eighth sentence of the introductory chapter of a nuclear physics text (comprising about 101 sentences) by Jackson & Barret (1977). These two sentences relate to the distribution of matter and charge in a nucleus, and to how particles, neutrons and protons interact in a nucleus. The comparison of these two sentences shows the existence of both reiteration and collocation. Examples of reiterations in these two sentences are *properties-properties, interactions-interactions*, and *distributions-distributions*. And, instances of collocation are *nuclear-charge, nuclear-matter*, and *particles-nucleus*.

---

2 The general type of lexical cohesion comprises repetition (e.g. *measure, measuring, measurement*), synonymy (e.g. *determine, find out*), antonymy (e.g. *action, reaction*), hyponymy (e.g. *measure, determine*), and metonymy (e.g. *atom, proton*). The instantial type comprise equivalence (e.g. *Expressions*), naming (e.g. The *phenomenon* was named *diffusion*), and semblance (e.g. The *compound* behaves like a *gas*).
The extent to which we are able to make precise and meaningful statements about the nuclear matter distribution and the nuclear charge distribution and the variation in both quantities from one nucleus to another reveals quite clearly the state of our understanding of much more fundamental issues, such as the nature of the interactions between various types of particles and the role of these interactions in scattering phenomena, the subtle balance between various features of the nucleon-nucleon interactions in bound states, and the difference between the average properties of nuclei described by macroscopic models and the specific nuclear structure properties described by microscopic models.

One of our principal aims will be to try to determine and explain precisely which properties of the relevant distributions can be obtained from the various experiments and to indicate the extent to which previously published parameters are really determined by the measurements as opposed to being merely consistent with them.

Note here that nearly most of the reiterations tend to occur at an inter-sentential level. Whereas collocations occur mostly in a single sentence. In our particular example, we see that there are also repeating items which are compound terms such as nuclear distribution. The number of lexical links may change depending on whether we consider single or compound term repetitions or both. We will discuss the number of links and its relevance further in this chapter.

To conclude this discussion on lexical and syntactic cohesion, Figure 3-1 below shows how the sentences of a well-formed text are interconnected in a coherent and cohesive sequence—an ideal representation of text where all sentences are connected through one, or more than one lexical or syntactic cohesion link.
Two sentences may be connected, or linked together 'cohesively' by means of the two general types of cohesion discussed above. These two types are summarised in Figure 3-2.

Hoey who has also studied cohesion in text has categorised the features of lexical cohesion into different ‘lexical semantic relations’. These include antonymy, that is opposites, hyponymy, for example expressions of hierarchies, and others. He has introduced in his study a number of potentially computable notions such as links, ties, bonds, and bond networks in relation to lexical cohesion and to text organisation. We will often refer to these terms (shown in Table 3-1) in the rest of the thesis.
Repetition | The occurrence of one or more items in a sentence.
Link | A connection by repetition between two items in a text.
Tie | A cohesive feature that helps to make a connection between the current sentence and a previous sentence: 'all links are ties, but not all ties count as links'.
Bond | A connection that exists between a pair of sentences by virtue of there being an above average number of links relating them. (The requisite number of links is 3 and it is never less than 3).
Bond Network | (i) A set of interconnection bonds amongst sentences. (ii) A graphic representation of texts.

Table 3-1: Lexical cohesion patterns in text according to Hoey (1991:265-269).

3.3 ‘Patterns of lexis’ in text

Hoey has studied patterns of lexical cohesion in text and has enumerated a number of other specialised 'devices' which ensure lexical cohesion (1991). He shows in his work, which is primarily inspired from that of Halliday and Hasan (1976), that lexical cohesion can further be classified. He focuses generally on the repetition of nominal terms, both single and compound terms, that are used particularly in scientific text for 'linking' sentences or parts of sentences.

We have seen above that lexical cohesion devices may be either reiterations or collocations. Hoey argues that the two types of lexical cohesion, collocation and reiteration, can be expressed using various lexical repetition devices. He termed such devices simple lexical repetition, complex lexical repetition, simple paraphrase, and complex paraphrase. We describe each of these devices with examples taken from a nuclear physics text (Jackson & Barret, 1977).

3.3.1 Simple Lexical Repetition

The literal repetition of a term is known as 'simple lexical repetition'. The assumption here is that words keep the same meaning as they are repeated throughout a given text; this is particularly true for scientific text.
The extent to which we are able to make precise and meaningful statements about the nuclear matter distribution and the variation in both quantities from one nucleus to another reveals quite clearly the state of our understanding of much more fundamental issues, such as the nature of the interactions between various types of particles and the role of these interactions in scattering phenomena, the stable balance between various features of the nucleon-nucleon interactions in bound states, and the difference between the average properties of nuclei described by macroscopic models and the specific nuclear structure properties described by microscopic models.

The study of nuclear sizes involves both the study of the nuclear charge distribution by means of processes dominated by electromagnetic interactions and the study of the nuclear matter distribution by means of strong interaction processes.

One major problem one might encounter is that of homonyms: words that have the same spelling but may be different in meaning, origin, grammar, or pronunciation. The noun ‘charge’ and the verb ‘charge’ are homonymous of each other, as are the noun ‘bank’ and the verb ‘bank’. Polysemy can be another problem wherein one grammatical category of a word may have multiple meanings: the term ‘nucleus’ is a case in point, in (nuclear) physics, a nucleus is a part of an atom, whereas in cell biology, a nucleus is a part of a cell.

The problems of homonymy and polysemy are major issues particularly when one deals with general language texts. In scientific writing, these problems are not encountered that frequently.

In the example shown above, there are six simple lexical repetition links which are nuclear-nuclear, matter-matter, distribution-distribution, interactions-interactions, interaction-interaction, and charge-charge. If we assume that three links establish a bond, then, the two sentences 4 and 5 are bonded. If instead, a bond needs at least 7 links to exist, then the two sentences would not be bonded. Hoey argues that there should be at least three links to form a bond. Hoey’s analysis did not exceed a forty sentence text. A bond threshold of three
could well be satisfactory for a text of this length. We believe that the specification of the number of links that form a bond, the so-called bond threshold, may be subject to a number of parameters such as the length of a text, the distance between sentences, and the type of lexical repetition that form the links. Unlike Hoey, part of our research work consists of the study of lexical cohesion patterns in longer texts with variable bond thresholds and the establishment of a relationship between the bond thresholds and the distribution of bonds in a text.

3.3.2 Complex Lexical Repetition

The more complex form of lexical repetition is the so-called 'complex lexical repetition' where two sentences may contain two different lexical items that either share the same lexical morpheme but are not formally identical, or, if the items are formally identical, they have different grammatical functions. Some examples are highlighted in the following two sentences:

Sentence 8: One of our principal aims will be to try to determine and explain precisely which properties of the relevant distributions can be obtained from the various experiments and to indicate the extent to which previously published parameters are really determined by the measurements as opposed to being merely consistent with them.  
Sentence 20: The electromagnetic interaction with the nuclear magnetic moment is observable in elastic electron scattering at 180, in inelastic electron scattering, and in hyperfine splitting of certain atomic levels, and it is therefore possible to determine the magnetic moment distribution of nuclei.

In the example above, the complex lexical links found are *distribution-distributions* and *determine-determined*. If we were to consider only complex lexical repetition links and a bond threshold larger than three, then the two sentences would not be bonded. We will show further in this thesis that complex lexical repetition is less encountered in text than simple lexical repetition.
3.3.3 Paraphrase

Paraphrasing in text is defined as the act of rewriting, or rewording parts of the text. Hoey argued that there are instances of paraphrase in text that contribute to its cohesion. He identifies two types of paraphrase: simple paraphrase and complex paraphrase (1991). Simple paraphrase occurs whenever a lexical item substitutes another in a context without loss or gain in specificity and without discernible change in meaning, e.g. *particle* and *atom*.

The substitution that Hoey talks about in the context of paraphrase can in fact be more restricted. He argues that simple paraphrase can be ‘mutual’ if two words actually forming the paraphrase can be interchanged without discernible change in meaning. Simple paraphrase is otherwise partial if only one of the two words can substitute another in context but not vice-versa. For instance, the word *proton* can be used to replace the word *particle* because particles are of different types such as nuclei, protons, electrons, neutrons, etc. In the pair of sentences 9 and 11 shown below, the words *nucleons* and *electrons* are particles that are both parts of the atom. They both are under a particle category and, therefore, they share general properties of particles. In restricted cases, a discussion on electromagnetic radiation of particles for example, these two words can be a potentially mutually paraphrased.

Sentence 9: The interaction between charged leptons (i.e. electrons, positrons, and muons) and nucleons consists of an electron-magnetic and a weak term, but the latter has a completely negligible effect in the processes considered here.

Sentence 11: This was many years after the suggestion (Guth 1934) that, for fast electrons, the finite size of the nuclear charge distribution would produce large deviations from the differential cross-section (Mott 1929) for elastic scattering from a point charge.

The other type of paraphrase is that of complex paraphrase. Here, Hoey argues that words can also paraphrase in a rather complex way if they are antonyms and occur in the same context. The terms ‘cathode’ and ‘anode’ are often encountered together in text on electric...
or electronic charges. These two terms can be considered a case of complex paraphrase since they represent opposite concepts. The other instance of complex paraphrase is found in what Hoey refers to as the complex paraphrase ‘link triangle’ which is shown in Figure 3-3. In this case, a complex paraphrase link between two words results from two other links between the two words and third word (which is actually hidden)—hence the link triangle. For example, given that the two words writer and writing are connected by a complex repetition link, and that the word writer is connected to the word book by a simple paraphrase link, one deduces the complex paraphrase link between the words book and writing.

Figure 3-3: Identifying complex paraphrase.

Simple paraphrase and complex paraphrase require the knowledge of the semantic relationships between words in a text. This knowledge is usually compiled in dictionaries, thesauri, and domain-specific term bases. Simple partial paraphrase between words can only be identified by using a hierarchical semantic network that contains relations such as ‘is a’ or ‘is part of’. We will show in the next chapters that paraphrase in text, particularly simple mutual paraphrase, is insignificant compared to simple and complex repetitions.

Consider, for example, the paraphrase 'particles'. Replacing one item of the paraphrase in a sentence with another from the particle hierarchy does not change the meaning expressed by such a sentence if the order of replacement is respected. In other words, if we say that
'protons are small particles', we can equally say 'muons are small particles'. They may not be of the same size, but they are still small in that they belong to, say, a 'small particles' class. Furthermore, using paraphrase, the author of a text may navigate in the hierarchy whenever generalisation or specialisation is needed.

### 3.4 Automatic cohesion-based text analysis - Assumptions and hypotheses

The question one might ask is: how can we understand text given the fact that the production of text and its texture is largely affected by obvious, and sometimes less obvious occurrences of specific syntactic and semantic patterns—the so-called patterns of lexis? Incidentally, this is a part of more complicated questions on text related to issues ranging from organisation and storage, to the understanding and generation of text. For the information retrieval community, the approach to text understanding is based on the identification and use of individual words/terms of the text. This problem is not posed in terms of 'protocols'—'writer/reader' and 'communications', though Salton is aware that text understanding and transformations must be based on some comprehension of the contents (Salton, 1989).

The argument here is that text is patterned but does not have structure such that one can 'make predictive statements about text organisation', and that such patterning reveals 'complex ways in which topics may interrelate in their development'. Hoey asks us to observe lexical repetition and lexical patterns in text because such an observation ‘may allow us to say interesting things about it [i.e. text] and to elicit coherent subtexts from it’ (1991:26).

The analysis of cohesion patterns in text, we believe, has some bearing on understanding the structure and the meaning of text. This analysis seems daunting to the uninitiated due to the fact that one has to consider a number of parameters in the analysis such as the variety of
cohesion types, the nature of the text to analyse, and its genre. These parameters have a direct influence on the results that one may obtain. In order to develop a cohesion-based text analysis methodology, we need to, first, look at different types of cohesion, and second, choose the text that we intend to analyse. These requirements are described below.

**Selecting cohesion types:** One can identify two major criteria in the choice of types of cohesion to consider for analysis. First, looking at the frequency of occurrence of types of cohesion in a text can help to identify which types are more likely to be encountered than others. Such a criterion requires reading large amounts of text and manually identifying different types of cohesion. The second criterion is based on the hypothesis that some types of cohesion establish bonds between sentences whereas others do so between paragraphs, sections, or chapters of a text. Stoddard elaborates extensively on the potential of cohesion types and emphasises the role of pronominal references which constitute a syntactic cohesion type (Stoddard 1993). Problems that may rise if we consider pronominal references is in the resolution of such references which may require powerful natural language processing techniques. We follow Hoey’s assumption in that 40% to 50% of cohesion in text is due to lexical cohesion and base our analysis therefore on lexical cohesion only.

**Choice of a corpus of selected text:** In any text analysis work, there is a need to specify the type, domain, nature, and, perhaps, the length of the text that one attempts to analyse. Furthermore, we cannot generalise a theory about the structure or behaviour of discourse for the simple reason that we cannot generalise results to all texts. This sort of selection is mostly due to the lack of efficient standards in defining and structuring text in minute details. However, until we understand all the regularities and irregularities of text and discourse, we cannot determine such standards, at least not yet.
The restriction of the types of texts, and even domains of knowledge, seems therefore prevalent despite the fact that limited conceptions of text can lead to limited theories that lack grounds, consistency, and provability. We understand that the need for specific and appropriate textual databases and corpora for cohesion-based text analysis is evident and that, as Kroeber (1967:58) points out, this may raise critical questions about the quality of the results obtained by such an analysis:

If one assumes he understands the essential mysteries of language and can display them in brief, invented sentences, one is unlikely to exploit the opportunities provided by fine literature to discover the fullest achievements and richest potentialities of the language.

Furthermore, the selection of passages of text with a limited number of sentences and the results obtained consequently may not be enough to study all the properties of text. In an attempt to answer Kroeber’s criticism, one needs to select substantial amounts of a wide variety of raw texts that are primarily written for other purposes that do not necessarily include cohesion analysis.

The lexical cohesion analysis methodology that we have developed does not suffer from Kroeber’s criticism in that we have attempted to analyse large texts, although limited to two domains of knowledge which are ‘Nuclear and elementary particle physics’ and ‘Water engineering’. As the problem of text length is solved, the number of texts to analyse is mostly subject to availability. We would like to argue that our methodology, as we will elaborate in the next chapters, can be generally applied to non-narrative scientific texts provided that we have the appropriate linguistic tools for the analysis, e.g. dictionaries, thesauri, term bases, grammars, etc.
These assumptions that we have also adopted in our methodology, we believe, are not a limitation of the methodology as such, but rather a reduction of the problem at hand. In other words, one cannot generalise the results of any cohesion-based text analysis unless there is a clear understanding of the type, structure, and domain of the text. However, our attempt is to show that even with such a handicap, one can tackle the problem of summarising and abstracting text as we adopt a ‘divide and conquer’ strategy.

3.5 Conclusions

In this chapter, we have discussed the linguistic concept of cohesion that is used in the production of text. We have seen how, through the use of both lexical and syntactic features, a writer creates a well-formed text that is thought to be cohesive in that the reader of such a text succeeds to understand the message of the writer and the arguments being conveyed. It appears that, while cohesion is a skill that the writer uses to produce good quality text, the reader sees cohesion as a text processing operation in that he/she looks continuously for linguistic signals that link parts of the text to others.

We have seen that cohesion in text, both syntactic and lexical, is a semantic (rather than a structural) concept which is concerned with meaning and that text linguists agree on its existence in almost every text of every genre. It occurs when the interpretation of an element in context refers or depends on other elements. The dependence or reference can be ensured by using either grammatical or lexical signals. Cohesion, in summary, is the major reason why a readable text holds together and has texture.

We believe that lexical cohesion in particular, due to the simplicity of its nature, is readily exploitable and that the notions of links and bonds established between cohesive sentences of
a text can be used to understand and represent the meaning of the text, and, consequently, summarise it.

In the next chapter, we will propose and discuss in detail our lexical cohesion-based text analysis and summarisation methodology.
4.1 Introduction

The discussion thus far in this thesis has been focused on how a set of sentences are bound together, through lexical cohesion for instance, to form a cohesive text. Our intention now is to discuss ways and means, based primarily on lexical cohesion, for extracting 'key' sentences of a given text. Such an extraction methodology, that is methods, tools, and techniques, will be grounded on Halliday’s functional linguistics where the stress is on how an author of a text ensures that the text has a ‘theme’ and how this theme is maintained through lexical repetition.

We have seen that lexical cohesion patterns help to establish lexical links between sentences in a text. Such links form a lexical cohesion network of sentences connected through the repetition of a lexical item or its variants. The lexical cohesion network is a record of bonding in a given text. Hoey refers to a connection that exists between a pair of sentences by virtue of there being an above-average number of links relating them as a bond. He claims that the above-average number of links cannot be less than three and that it varies from one text to another and even within the same text. We call this number the bond threshold. The notion of bonds formed by lexical cohesion links between sentences within a text constitutes the essence of our cohesion-based text analysis and summarisation methodology.

Given the bonding information between sentences, a text comprising such sentences can be represented as a matrix whose elements are the sentences of the text. Each of these
sentences is associated with the information on its lexical cohesion bonding with the other sentences of the text. We refer to this information as the ‘bonding pair’. A bonding pair is a numerical indication of how many bonds a given sentence has with sentences that precede it and with others that succeed it. The relevance of the matrix resides in the fact that it provides a way to differentiate between different sentences in a text. In other words, the bond matrix can be used to identify sentences which have many bonds, those which have little or no bonds, and which have more bonds with successor than with predecessor sentences.

The notion of bonding pairs is illustrated in Figure 4-1. Figure 4-1a illustrates a collection of links whereas Figure 4-1b represents a collection of bonds that constitute a bond network.

The elements of the square and symmetric bond matrix are flags that indicate the presence or absence of a bond between the corresponding pairs of sentences. The ‘1’ flags indicate the corresponding two sentences are bonded and the ‘0’ flags indicate the
absence of a bond. The sum of the flags of a given row represents the total number of bonds that the corresponding sentence in that row has with successor sentences. Similarly, the sum of the flags of a given column is the total number of bonds that the sentence in that column has with previous sentences. The cohesion bonds in a text can therefore be regarded as a list of elements of the form $S(N_{\text{bef}}, N_{\text{aft}})$ where $S$ is the sentence number, $N_{\text{aft}}$ and $N_{\text{bef}}$ represent the number of bonds that the sentence $S$ has with successor and predecessor sentences. In the rest of this thesis, we will refer to this list as the ‘sentence bonding’. An example of a bond matrix is shown in Table 4-1 below:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>$Naft=\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 4-1: An example of a bond matrix.

The computation of lexical cohesion in text is a challenge that both computer scientists and linguists take despite the complexities of the endeavour. This is, perhaps, due to the immensity of the potential of cohesion, the variety of cohesion devices, and the ever growing textual resources. One can therefore restrict the research question by considering fewer cohesion types, or by analysing shorter text, or text in specific domains (cf. scientific text).
We focus in our study on non-narrative texts such as scientific texts. The term 'non-narrative' basically stands for texts that are not narrative. A narrative text is defined as a story describing events that are either real or imaginary (Wales, 1989). There are two different types of narrative text: real and fictional. Real narrative texts include newspaper stories and historical records that are usually based on real facts whereas fictional narrative texts include texts like comic strips, poetry, novels, and short stories.

In our context, the choice of texts for the analysis has been subject to two main criteria. First, we are interested in texts whose cohesion is achieved through lexical cohesion more than syntactic cohesion. Second, the study done by Meyers and Hartley (1990) on lexical cohesion in specialist and popular science articles is quite important in that it reveals interesting facts about the nature and types of cohesion found usually in scientific texts. Meyers and Hartley argue that in 'scientific texts', which they use to designate texts published in specialist journals in contrast to 'popular texts' of non-specialist magazines, cohesion is achieved mostly through reiteration of strings of characters, with little substitution, conjunction, and almost no pronouns. Based on Meyers and Hartley's findings, we believe that our methodology for lexical cohesion based summarisation of non-narrative scientific texts can be used to determine the role of lexical cohesion in this type of text and, perhaps, be a model for the development of lexical cohesion analysis of other types of texts.

The objective of this chapter is to describe our methodology for text summarisation based on lexical cohesion analysis. In section 4.2, we show how sentences are categorised on the basis of the number of lexical cohesion bonds they have with each other. The categorisation of sentences, as we will explain, can be semi-formalised by specifying a number of parameters. Section 4.3 contains a detailed description of the
role of sentence categorisation in text summarisation. We present in this section various
text summarisation methods based on the selection of different sentence categories.
Sections 4.4, 4.5, and 4.6 cover a description of our methodology. We explain how
lexical analysis can be achieved and suggest a general computational architecture for text
summarisation based on the analysis of lexical cohesion in text. Based on this
architecture, we have developed a lexical cohesion analysis and text summarisation
system TELE-PATTAN. The tools and services of TELE-PATTAN are briefly outlined
in section 4.7. Section 4.8 concludes this chapter.

4.2 Sentence categorisation and lexical cohesion

The notions of links, bonds, and bond networks discussed above are used to establish an
empirical categorisation of sentences (in a text) based on patterns of lexical cohesion.
These empirical categories are based essentially on the number of bonds a sentence has
with the preceding and the aft sentences. Hoey has argued that there are four categories
of sentences and each of these categories is distinguished from the other by the number
of bonds with its predecessor and successor sentences. The four categories are referred
to as topic-opening, topic-closing, central, and marginal sentences. We argue that the
central sentences can be subdivided into ordinary central and key-central sentences. We
outline each of these categories below and show how they can be identified using the
sentence bonding (a list of bonding pairs derived from the bond matrix).

4.2.1 Topic opening sentences

A topic opening sentence is any sentence that has considerably more bonds with
sentences after it than with sentences before it. A computer program can be used to
decide which sentences are potential topic opening sentences by ranking the sentences of
a text in a descending order according to the number of aft-bonds. One can then argue
that through an empirically determined threshold, say, $Th_{aft}$, sentences with a number of aft-bonds (i.e. $N_{aft}$) above this threshold will have a greater potential for being topic opening sentences.

4.2.2 Topic closing sentences

The sentences in a text that have considerably more bonds with predecessor sentences than with successor sentences are potentially topic closing sentences. Potential topic closing sentences can be identified by ranking all sentences of a text according to the number $N_{bef}$ in a descending order. Given a topic closing threshold, say $Th_{bef}$, sentences which have a number $N_{bef}$ above this threshold represent potential topic closing sentences.

4.2.3 Central sentences

The concept of centrality of a sentence is defined in relation to the total number of bonds the sentence has with predecessor and successor sentences within a text. Central sentences are those which are bonded to most of the sentences in the text. One way to identify this category is to rank the sentences in a descending order according to the total number of bonds (i.e. $N_{aft} + N_{bef}$). The sentences which come at the top of the list and have a total number of bonds above a certain threshold represent potential central sentences.

It is possible that some of the topic opening and topic closing sentences may have a large total number of bonds. Such sentences are to be excluded from the set of central sentences. There are two parameters that can be used as thresholds for deciding automatically as to whether a sentence is central or not. The first parameter, which we
call the difference parameter and define as \( d(=|N_{off} - N_{bef}|) \), is used on the instruction that, for topic opening and topic closing sentences, this difference \( d \) would be a large number. Therefore, for central sentences, \( d \), should not be below a certain threshold number that we refer to as the difference threshold, \( Dth \). The second parameter that we call the centrality threshold \( Cth \), is related to the total number of bonds normalised with respect to the total number of sentences \( n \) in the text. This normalisation helps to take into account the fact that for larger texts, there would be a larger number of sentences with more bonds than a corresponding text with a lesser number of sentences. Therefore, all sentences that fulfil the following centrality conditions are potentially central sentences:

\[
\frac{N_{off} + N_{bef}}{n} \geq Cth \text{ and } |N_{off} - N_{bef}| \leq Dth
\]

We believe that such conditions allow a better selection of central sentences and help to avoid possible overlapping of topic opening and/or topic closing sentences with central sentences.

### 4.2.4 Key Central sentences

Having defined the concept of centrality above, it is possible that, for a long text, the number of central sentences may be considerably high. We introduce therefore the notion 'key-centrality'. Key central sentences can be identified by determining the most central sentence in the text, that is the sentence which has the largest number of bonds and is above the difference and centrality thresholds. We know that the total number of bonds for the most central sentence is, say a number \( Max \), and that key central sentences are primarily central sentences that should be closer to the most central sentence. In
order to identify these sentences in a text, assume that all central sentences are placed on
a horizontal line whose length is the number Max as shown in Figure 4-2 below.

The positions of key central sentences on the line depend on the total number of bonds
they have. One can define a limit along this line beyond which central sentences become
key central. The limit should correspond to a fraction of the number Max. We call this
limit the ‘key-centrality threshold’ $KCth$ (whose value should be between 0 and 1). Key
central sentences should, on top of the centrality conditions, fulfil the following
criterion:

$$N_{af} + N_{bf} \geq KCth \times Max$$

The selected sentences are considered as ‘key central sentences’ and are more focused
on the main topic of the text than central sentences.

4.2.5 Marginal sentences

A marginal sentence is any sentence that has a total number of bonds that is less than a
certain number. We refer to this number as the ‘marginality’ threshold. The notion of
marginality is used to describe those sentences which have very little or no bonds at all
with other sentences in the text. Therefore, given a marginality threshold $Mth$, a
potential marginal sentence $S(N_{bf}, N_{af})$ should meet the following condition:
\[ N_{en} + N_{bef} \leq Mth \]

Intuitively, one might consider the value 0 as a pragmatic choice for the marginality threshold in a short text. However, in a long report, a book, or generally a long text, the marginality threshold may be higher, particularly if one is interested in generating summaries of long texts based on the elimination of marginal sentences.

4.3 Text summarisation

The discussion on the extraction of different classes of sentences leads primarily and rather intuitively to the idea of text summarisation. A text summary can be either an abstract or an extract. An abstract of a text is a short text that attempts to describe the content of the whole text in a very abridged form. The sentences found in an abstract may not be necessarily found in the original text. They usually are written to express the content of a text in a different and/or keyword-like phrases. An extract however, consists of a collection of sentences that are selected from the original text. Although, both abstracts and extracts represent two different types of text summaries, abstracts are more concerned with the meaning of the text, as opposed to extracts which are more concerned with its content.

TELE-PATTAN offers four methods for generating summaries based on the selection of sentences of different categories from the original text. Another method consists of extracting whole paragraphs as summaries instead of individually selected sentences. This means that TELE-PATTAN does not generate natural language sentences for a summary but rather extracts them from the text based on various selection criteria. These criteria define what we refer to as summarisation procedures. We describe below
each of the four procedures which are all based on individual sentence selection and postpone the discussion of paragraph selection based summaries till the next chapter.

Procedure 1: The first strategy of generating a text summary is to collect topic opening, central, and topic closing sentences. Such a summary would include introductory and concluding sentences in addition to central sentences. This type of summary is usually the longest amongst other summaries generated by other procedures.

Procedure 2: To generate a summary that can be used to describe the content of a text without having a lengthy introduction or conclusion, a set of central sentences could well be the desired summary. In fact, this would be a preferred summary for users who want to study a lengthy text without having to read all of it.

Procedure 3: An even shorter summary of a text can be produced by considering only key central sentences. Given the fact that the number of key central sentences is by and large smaller than the number of central sentences, the summary produced is considerably shorter. The idea here is to describe the whole content of a text in terms of one, two, or three sentences at most that should convey the gist of the text.

Procedure 4: The bond matrix of a text may contain sentences which are marginal, i.e. those which have a very small number of bonds (c. zero). Ignoring these marginal sentences would normally not affect the overall meaning of the text. Consequently, the original text without these marginal sentences can be thought of a summary. In an ordinary text, one does not expect a large number of marginal sentences. This means
that a summary based on ignoring marginal sentences may be longer than other summaries.

The discussion on any summarisation methodology is paved with questions related to the size of the summary and its readability. These amounts to the problem of evaluating automatic summary generators. In our context, the length of the summary is largely dictated by the configuration of the sentence categorisation parameters. The quality of its readability, however, depends on the opinion of the reader. So far, the summaries produced by TELE-PATTAN have a good degree of readability. One has to admit that extracting one sentence as a summary of a thousand sentence long text is unreasonable and would undoubtedly lead to a loss of information. This is to say that the readability of the summary depends on its length and the distance between its sentences. One way to improve the readability of a summary, perhaps, is to have sentences of the same paragraph appearing in the summary. We will come back to this point in the next chapter where we introduce the notion of paragraph centrality and discuss paragraph extraction based summaries.

4.4 Computing lexical cohesion: The need for linguistic knowledge resources

The extraction of lexical cohesion patterns is a process that involves the identification of lexical ties between words within sentences of a text. These ties may or may not contribute to the cohesion of the text depending on the context in which the words that form the tie occur. Furthermore, the development of a computational model for cohesion analysis requires some lexical knowledge resources related to the language in which a given text is written. These knowledge resources can be dictionaries, thesauri, or even terminology data bases which are organised repositories of terms and concepts
describing a specific domain. In Figure 4-3, we show the three lexical resources and how they may be used to extract lexical cohesion patterns in text.

![Diagram of Lexical Knowledge Resources]

**Lexical Knowledge Resources**

- Dictionary
- Thesaurus
- Termbank

Collocation  Morphology  Syntax  Related words  Synonyms  Antonyms  Related terms  Synonyms

Idioms  Compound Words

Simple Lexical Repetition  Complex Lexical Repetition  Paraphrase

Figure 4-3: The use of lexical resources for computing lexical cohesion.

The identification of simple lexical repetition patterns is relatively straightforward in that, if one excludes homonymy, it becomes a pattern matching operation. Complex repetition patterns can be identified using a knowledge base that comprises (some of) the knowledge of a given language, i.e. morphology rules, derivation rules, as well as rules dealing with irregular nouns and verbs. Identifying and extracting complex lexical cohesion patterns consists of knowing exactly all the morphological features of words in a text. These features include the derivatives and inflection forms of a given word.

Simple and complex paraphrase repetitions involve knowledge about the meaning of the words in the text and, in some cases, the pragmatics of the discourse. The need for a dictionary, or a thesaurus, is then crucial at this stage. The dictionary and the thesaurus
are resources that help to identify a semantic role of a word or an expression in a given context. This semantic role is implicit in the meaning of a word (in a dictionary), and is reinforced by the relationships that link it to other words (in a thesaurus). A dictionary is a reference book that lists (usually in an alphabetical order) and explains the words of a language or gives their equivalents in other languages. It also includes other information such as the pronunciation, the etymology, and the morphological variants of the words. A thesaurus is an extended dictionary that contains most of the words in a dictionary and 'lists' of words denoted by semantic names. The term thesaurus has two meanings. First, a thesaurus is a reference book comprising synonyms, often including related and contrasting words and antonyms; thesauri like Roget's use a set of 'concepts'—material, political, etc.—to further categorise synonyms and related words. In other words, it is an organised collection of words in general language. Second, a thesaurus is sometimes referred to as a reference book of selected words and/or concepts such as specialised vocabulary of a particular field, in short, an organised collection of special terms. Such a collection is nowadays referred to as a terminology data base or a 'term base'. In this thesis, we use the terms 'general language thesauri' and specialist-language 'term base' so as to avoid the polysemous use of the term thesaurus.

The complex and simple paraphrase identifications rely mostly on the categories that are defined in a thesaurus. These categories are defined on the basis of many criteria, e.g. synonymy, antonymy, relatedness, and collocation.

In the section that follows, we discuss a lexical cohesion-based text summarisation methodology and how we can extract some of the lexical cohesion patterns using a
thesaurus. The techniques can be used with either a dictionary or a terminology data base. Better still, the three resources that we describe above can be used simultaneously. For instance, one can use a general language thesaurus to check for semantic relations between words such as antonymy, while other relations such as synonymy and morphological variants (which are not found in a thesaurus) can be looked-up in a dictionary. The relations between special domain terms that are not found in a dictionary or thesaurus can be looked-up using an appropriate terminology data base.

4.5 A lexical cohesion-based text summarisation methodology

Lexical cohesion, and any analysis based on lexical cohesion, is perhaps the least intrusive method for analysing text in that it is less concerned with the structure of the sentences, i.e. the grammar of the language, than the methods used in conventional natural language processing systems in artificial intelligence. And, at times, it only involves pattern matching processes whereby recurrence of single or compound words are recorded. The expression 'least intrusive' means that such a method neither derives its inspiration from complex descriptions of syntax or semantics of a language nor is it based on arguments commonly found in semantics literature which are in turn based on the philosophical orientations of the semanticists. Lexical cohesion analysis has less linguistic and epistemological input as compared to syntactic and semantic analyses. This makes it a non-intrusive type of analysis.

The crucial question which is of most concern to us as computer scientists is whether we can make use of the well-written text of the author and the competence of the reader of
such text to understand its structure and content. Most importantly, can we use the notion of the oft-repeated cohesion patterns to summarise text?

Given that lexical cohesion manifests itself as lexical patterns, and that these might help in the identification of the central sentences of a given text and in the extraction of sentences that may be used to introduce or indeed to close topics, it is possible to develop a computer based methodology for identifying such patterns for a range of applications in general and for text summarisation in particular. For terminology management, particularly terminology extraction, one can argue that the so-called central sentences represent potential repositories of a large proportion of significant terms of a domain. The generation of text summaries is currently dependent on the more-intrusive techniques of syntax analysis which is in turn based on a fairly narrow and incomplete view of a given natural language. However, the use of lexical cohesion, particularly the identification of topic opening, central, and topic closing sentences provides a basis for generating full text summaries. This is perhaps the least intrusive and most practical way of generating summaries. For example, information retrieval relies on a very reductive strategy, that of keywords in context (i.e. KWIC), for indexing and retrieving text from a large database. If keywords are reductive because they are the commonest units found in a range of texts, then lexical patterns due to lexical cohesion can be used instead of keywords. Unlike keywords, lexical patterns do not ‘reduce’ text but rather help in identifying it.

It is the simplicity of the notions that underlie lexical cohesion, and, the oft-repeated observation that even for analysing small fragments of texts, one requires masses of time if such analysis is carried out by hand, that has motivated us to conceive a computer-
based methodology for the analysis of lexical cohesion in text. We present a methodology for analysing a text, extracting lexical cohesion patterns, categorising sentences, and summarising text.

4.5.1 Architecture

Figure 4-4 shows the underlying architecture of the system (TELE-PATTAN) that we have developed to analyse texts, compute lexical cohesion patterns, and produce text summaries. The methodology on which the architecture of TELE-PATTAN is based consists of selecting texts from a corpus of scientific texts, pre-processing and analysing these texts, visualising the patterns of cohesion, and interpreting these patterns. Pre-processing consists of marking-up the texts and indexing their sentences. We have used a special mark-up notation to indicate the boundaries and headers of linguistic structures such as paragraphs, sections, and chapters of a text. The mark-up is also used to indicate references and expressions that are not considered as lexical items such as mathematical equations and numerical expressions. The first task in performing text analysis is the computation and extraction of patterns of lexical cohesion. This task requires the use of morphological analysis and a thesaurus. TELE-PATTAN uses the on-line electronic Wordnet Thesaurus and has its own morphology analyser of the English language. The patterns of cohesion corresponding to the text are stored and used to build the cohesion (or bond) matrix that we mentioned earlier in this chapter. Both the bond matrix and patterns of cohesion are visualised in textual and graphical forms. Based on the bond matrix and cohesion patterns corresponding to a text, the sentences of such a text are categorised and various summaries are subsequently generated. The architecture also includes the analysis of the distribution of lexical
cohesion bonds in text and its effect on both sentence categorisation and text summarisation.

Figure 4-4: A methodology for lexical cohesion-based text analysis and summarisation.

4.5.2 Text mark-up and parsing
The task of recognising the different linguistic units in a text is not trivial. This is due to the complex structure of text. Consider the linguistic unit of a sentence for instance, it is common knowledge that a sentence starts with a capital letter and ends with a period. However, one often encounters abreviations which end with a period and proper names
which start with capital letters. This usually is one of the problems of identifying sentence boundaries. There are stylistic heuristics that one can use in identifying these boundaries in cases where such an operation is ambiguous. An example of these heuristics is that two sentences are separated by a period and two consecutive blank spaces.

The organisation of text is relevant to almost all types of text analyses. It is therefore necessary to identify the different components of a text that are either directly or indirectly related to the desired analysis. Some analyses may only use lines of words in a text, e.g. intelligent question answering systems. Other types of text analysis such as those used in terminology management and elicitation systems may use words, terms, and phrases. We note that, generally, the relevant components of a text structure that one intends to use for a given analysis depends considerably on the types of tasks that one wishes to perform on the text.

A lexical cohesion based text analysis requires almost all textual components that are known and frequently encountered in ordinary text. These include, headers, chapters, sections, paragraphs, sentences, words, figures, and tables. We have attempted, by adopting a text structure, to observe the effect of the components of this structure on the distribution of lexical cohesion links. We assume therefore that a text which is available either in a paper or an electronic form, may contain some or all of the linguistic components shown in Figure 4-5 below.
Figure 4-5: A text structure for lexical cohesion analysis.

The fork-like ends of the connections between the linguistic units, i.e. the nodes, in the figure above, indicate one-to-many relationships between these units. The diagram shown in the figure above is an entity relationship model (Bowers, 1993) comprising most of the entities that have a relation with the entity ‘text’. The word ‘entity’ here stands for a thing in the real world with a real existence (Date, 1986). The representation of these entities can be expressed in the form of rules which act as a document specification grammar as outlined below. A text document is seen as a set of entities with a title and a reference number which is used for indexing.

```
Document := <Textual_object><Object_number><Object_title>
Textual_Object := {text|chapter|section|paragraph|figure|table|sentence, }
Object_number := number|number expression;
Object_title := a string of characters;
```
The mark-up of text prior to its analysis consists of associating a set of tags with the
different entities in the text. We have developed our own parser that can extract all the
entities above. This parser can be adapted to a more general and standardised type of
mark-up such as SGML, i.e. Standard Generalised Mark-up Language (Jones, 1991).

4.5.3 Identifying and extracting lexical cohesion patterns

The identification and extraction of lexical cohesion patterns consist of comparing all
sentences in a given text and determining which pairs of words are potentially repeated.
That is to say which words, and in what sentences, have a lexical cohesion link. The
number of comparisons can be reduced by ignoring words that are frequently repeated
such as definite and indefinite articles like ‘the’ and ‘a’ which belong to what is known
as ‘closed class words’. In Table 4-2, we outline some criteria that indicate the
existence of potential lexical cohesion links.

<table>
<thead>
<tr>
<th>Type of links</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple lexical repetition</td>
<td>Words which are literally the same. Base and possessive forms of words</td>
</tr>
<tr>
<td>Complex lexical repetition</td>
<td>Singular and plural forms of words, e.g. phenomenon and phenomena; or</td>
</tr>
<tr>
<td></td>
<td>Verbs at any tense and their ‘ing’ form, e.g. find and finding;</td>
</tr>
<tr>
<td></td>
<td>Verbs of the same form at different tenses, e.g. one is the past tense of the other, e.g. find and found;</td>
</tr>
<tr>
<td></td>
<td>Words which are literally the same but have different grammatical</td>
</tr>
<tr>
<td></td>
<td>functions, e.g. to make and a make;</td>
</tr>
<tr>
<td></td>
<td>Words which share a common morpheme and have different grammatical</td>
</tr>
<tr>
<td></td>
<td>functions, e.g. founded and foundation, or cohesive and cohesion.</td>
</tr>
<tr>
<td>Simple mutual paraphrase</td>
<td>Words which are synonyms;</td>
</tr>
<tr>
<td></td>
<td>Plural forms, singular forms, and possessive forms which are synonyms;</td>
</tr>
<tr>
<td></td>
<td>Verbs at the same or different tense which are synonyms, e.g.</td>
</tr>
<tr>
<td></td>
<td>embellishing and enriched.</td>
</tr>
<tr>
<td>Simple partial paraphrase</td>
<td>Words that have an ‘ordinate-subordinate’ relationship, e.g. animal and</td>
</tr>
<tr>
<td></td>
<td>cat;</td>
</tr>
<tr>
<td></td>
<td>A word which is the plural of a simple or complex paraphrase of the</td>
</tr>
<tr>
<td></td>
<td>other, e.g. animals and cat;</td>
</tr>
<tr>
<td></td>
<td>A word which is the possessive form of a simple or complex paraphrase of the other, e.g. animal’s and cat;</td>
</tr>
<tr>
<td></td>
<td>A word is the synonym of a simple or complex paraphrase of the other,</td>
</tr>
<tr>
<td></td>
<td>e.g. beast and cat (the word beast is a synonym of the word animal);</td>
</tr>
<tr>
<td></td>
<td>Verbs at the same or different tenses and but are simply partially</td>
</tr>
<tr>
<td></td>
<td>paraphrased, e.g. contained and include.</td>
</tr>
</tbody>
</table>

Table 4-2: Criteria for the identification and extraction of lexical cohesion patterns.
Looking at the table above, the first impression one gets is that simple lexical repetition is the easiest to identify in a text compared to all the other repetitions. The identification of instances of complex repetition requires a morphological analysis which consists of extracting morphological variants. Simple paraphrase occurs when certain words in text are used to replace other words without affecting the meaning of the text and without loss or gain in specificity. It is partial if the replaced words contain (i.e. include in meaning) the others. It is mutual if both words can be interchanged. Synonymy is an example of simple mutual paraphrase. However, instances of simple partial paraphrase are not that simple to identify. For instance, one may substitute the word *particle* with *proton*, because any assertion about a particle means also the same assertion about the proton. This is possible since a proton is a particle, but all particles are not necessarily protons. Therefore, the substitution of the two words can only be partial.

4.6 Why use Wordnet?

Wordnet is an on-line lexical reference system which is a database of nouns, verbs, adjectives, adverbs, and function words organised into synonym sets representing each a different lexical concept and linked by different relations (Miller et al., 1990). Wordnet is primarily based on psycholinguistic theories of the human lexical memory. The meaning (or synonym) sets of Wordnet are linked with different semantic relations such as synonymy, antonymy, hyponymy, and meronymy. Wordnet has also a limited set of morphological relations between word forms. A detailed description of the structure of this thesaurus can be found in appendix A.

In the early stages of our research, we used the Macquarie thesaurus of Australian English that we were lucky to acquire in a machine readable form. This thesaurus also
known as 'the book of words' (Bernard, 1990), contains over 180,000 entries divided into over 800 'semantic' categories which range from 'abstinence/overindulgence' to 'animal noises', to 'sport' (mainly Australian sport), to trade unions, to water craft and work, etc. For each semantic category, there are entries under nouns, adjectives, verbs, and adverbs. Each of the grammatical categories could be divided into up to 7 or 8 subcategories. The importance of the Macquarie Thesaurus which was to motivate us to use it for lexical cohesion analysis is that it comprises relatively modern English as it is compiled in the 1990's. However, its restricted and non-clear semantic categorisation caused a major problem in identifying lexical relations between words.

Another important lexical resource that one could think of using in a lexical cohesion analysis is the Roget's Thesaurus. This thesaurus was first compiled in 1849 by Peter Mark Roget. In Roget's Thesaurus, english words and phrases are classified and arranged so as to facilitate the 'expression of ideas and assisting literary composition'. This thesaurus has been revised a number of times over the years to incorporate changes in the English spelling system. Roget's thesaurus has words classified into six 'conceptual' categories that are; a) abstract relations, b) space, c) matter, d) intellect, e) volition, and f) affections (Dutch, 1972). In this thesaurus, cross-reference related words both in terms of orthography and meaning. Notwithstanding the fact that Roget's Thesaurus was made available in machine readable form only recently, it is not clear how its semantic categories are marked. This is perhaps due to the fact that it comes as flat file structure.

In assessing Wordnet, Macquarie's, and Roget's Thesauri, and knowing that lexical cohesion analysis requires thorough and efficient search of lexical relations between
word forms and morphological variants of word forms, we clearly see the emerging advantages of Wordnet as the most suitable resource to use. In Table 4-3, we illustrate the properties of these thesauri that we discussed.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Wordnet</th>
<th>Macquarie's</th>
<th>Roget's</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Language</td>
<td>Modern</td>
<td>Modern</td>
<td>Modern</td>
</tr>
<tr>
<td>Contents size</td>
<td>164,000 entries</td>
<td>180,000 keywords entries</td>
<td>990 entries</td>
</tr>
<tr>
<td>Semantic Categorisation Type</td>
<td>Conceptual (all the synonym sets)</td>
<td>Ad-hoc 812</td>
<td>Conceptual 5x24</td>
</tr>
<tr>
<td>Semantic Relations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonymy</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Antonymy</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Hyponymy</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Meronymy</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Morphological Relations</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Syntactic categories</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nouns</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Verbs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Adjectives</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Adverbs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Cross references</td>
<td>rich</td>
<td>limited</td>
<td>rich</td>
</tr>
<tr>
<td>Machine readable forms</td>
<td>Fully organised database</td>
<td>Flat text files</td>
<td>Flat text files</td>
</tr>
<tr>
<td>Use for lexical cohesion analysis</td>
<td>Suitable</td>
<td>Limited</td>
<td>Limited</td>
</tr>
</tbody>
</table>

Table 4-3: A comparison of Wordnet, Macquarie’s, and Roget’s Thesauri.

Although there is no clear psychological evidence on which the grouping of synonyms in Wordnet is based, the authors of Wordnet make use of a lexicographic assumption that a synonym set represents a lexical concept rather than an abstract one (Miller et al., 1990). The discussions that will follow in this thesis are based on the use of Wordnet as a lexical knowledge resource.

4.7 Tools and services of TELE-PATTAN

In an attempt to implement the above methodology, we have developed TELE-PATTAN: a system that can be used to extract patterns of cohesion in text and use the distribution of such patterns for text summarisation.
LExical PAtterns ANalysis) is written in Prolog, a logic programming language. In addition to its text summarisation capability, this system can be used by linguists, terminologists, journalists, and authors, in the exploration and analysis of cohesion in texts. It is therefore an important tool to evaluate and test theories related to the notion of cohesion and coherence. The analysis of text and the extraction of lexical cohesion patterns are based on theories laid down by a number of corpus linguistics and text linguistics experts such as Hoey, Halliday, Hasan, deBeaugrande, Dressler, and so on. TELE-PATTAN is an experimental system that is primarily aimed at exploiting patterns of lexical cohesion in text and using these patterns to identify different categories of sentences. These categories represent the backbone of TELE-PATTAN’s lexical cohesion based text summarisation methods. We outline below some of the main tasks that this system can be used to achieve:

- Analysing new texts
- Building and viewing the bond matrix
- Visualising the patterns of lexical cohesion
- Extracting topic opening, topic closing, central, key central, and marginal sentences
- Viewing lexical cohesion patterns distribution
- Configuring the sentence extraction parameters
- Generating text summaries
- Extracting central paragraphs

TELE-PATTAN, shown in Figure 4-6, is configurable and can be interfaced with any machine readable lexical resource that is made available via the Internet such as dictionaries, thesauri, and terminology data bases³.

³ For a more detailed description of TELE-PATTAN’s functionality, development, and implementation, refer to TELE-PATTAN’s reference manual and user guide (Benbrahim, 1996).
4.8 Conclusions

We have attempted throughout this chapter to describe our computational methodology for lexical cohesion-based text analysis and summarisation. This methodology which is based solely on the availability of machine readable texts, a detailed thesaurus or a dictionary, and a set of morphological rules, is exploitable for the analysis and summarisation of scientific texts.

We have seen that the bond matrix reflects the distribution of bonds in a text and can be used to categorise the sentences of the text. We have discussed how these categories can be determined using semi-formal computation methods that are primarily based on the bonding pair each sentence in a text has. These categories comprise topic opening, topic closing, central sentences, key central, and marginal sentences. The semi-formal specification of these categories represents guidelines that reflect the importance of a sentence in a text and its role in establishing a cohesive sequence of arguments and ideas, or alternatively, its marginality with respect to the subject matter of a text.
In the next chapter, we will present and discuss the analysis of various texts of different sizes and demonstrate how summaries based on patterns of lexical cohesion are generated.
Chapter 5

Lexical Cohesion-based Text Summarisation: Case studies

We have discussed earlier the distribution of lexical cohesion links and bonding in a text. The bond matrix that describes the types and number of these links is the major ingredient in identifying bonded and non-bonded sentences. The bonding is established between the sentences as a result of the existing number of lexical cohesion links that in turn forms bonds. Such a bonding can be used to identify the different categories of sentences such as topic opening, topic closing, central, key central, and marginal sentences. These sentences can then be used to generate text summaries of different types.

In the course of this chapter, we will discuss the results of the lexical cohesion analysis of texts of varying lengths. We will present a detailed discussion on how the different types of lexical cohesion contribute to the overall cohesion of a given text. We then show how the four summarisation procedures that we discussed in the previous chapter can be used to produce summaries of a given text based on the distribution of lexical cohesion links within the text. Each of the summarisation procedures is used to generate summaries based on the selection of (highly-cohesive) single sentences. Such a clamping of sentences may read as rather odd. Intuitively, it appears that instead of single sentences, one might use the neighbouring sentences also. Taken to its logical conclusion, one can argue that a summary might consist of a series of highly connected paragraphs. We discuss the feasibility of generating such summaries in a later section of this chapter.
In this chapter, we present a few case studies on lexical cohesion based text summarisation. We outline the preliminaries of this methodology in section 5.1 where we show how a text is marked-up and how lexical cohesion links are extracted leading to the formation of bonds, and subsequently the formation of the bond matrix. We discuss here the distribution of lexical cohesion bonds in the first text comprising about 100 sentences. The results of the analysis of four other texts are discussed in section 5.2. The sections 5.3 and 5.4 consist of a discussion on cohesion-based summary production and the evaluation of such summaries. In section 5.5, we raise the issue of the readability cohesion of the summary. We will discuss the problems associated with single sentence extraction and suggest in this respect an alternative summarisation method which consists of extracting central paragraphs. Finally, we conclude the chapter in section 5.6.

5.1 Methodological preliminaries

In this chapter, we present the results of the analysis of five texts, four in nuclear and elementary particle physics and one in water engineering. The first text that we analysed is the introductory chapter of Jackson and Barret's book on nuclear sizes and structure which comprises 101 sentences. Three other texts are learned papers acquired from the CERN preprints virtual library on the Internet\(^4\). The fifth text is a report on a feasibility study of a harbour construction. These texts are listed in the Table 5-1 below.

<table>
<thead>
<tr>
<th>Label</th>
<th>Text type</th>
<th>Domain</th>
<th>Number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-1</td>
<td>Book chapter</td>
<td>Nuclear and Elementary Particle Physics</td>
<td>101</td>
</tr>
<tr>
<td>Text-2</td>
<td>Learned paper</td>
<td>Nuclear and Elementary Particle Physics</td>
<td>204</td>
</tr>
<tr>
<td>Text-3</td>
<td>Learned paper</td>
<td>Nuclear and Elementary Particle Physics</td>
<td>216</td>
</tr>
<tr>
<td>Text-4</td>
<td>Learned paper</td>
<td>Nuclear and Elementary Particle Physics</td>
<td>613</td>
</tr>
<tr>
<td>Text-5</td>
<td>Report (Feasibility Study)</td>
<td>Water engineering</td>
<td>2671</td>
</tr>
</tbody>
</table>

Table 5-1: A selection of texts used in the analysis.

\(^4\) The World Wide Web address to this library is: \texttt{http://cern.preprints.cgi-bin/}
Most common lexical cohesion links are those due to simple repetition which amounts for 70-85% of bonds that exist in all texts. Thus, an analysis based simply on simple repetition can help in the generation of summaries. We discuss cohesion patterns in the Jackson and Barret text at length; the results obtained from other texts are remarkably similar.

5.1.1 Text mark-up

The text is first manually marked-up and then parsed to extract the different textual components such as words, sentences, paragraphs, sections, figures, tables, chapters, etc. A mark-up example is shown below.

```
<Text><1><NUCLEAR SIZES AND SHAPES>
<Authors><JACKSON, D. & BARRET, R.>
<Chapter><1><Introduction>
<Section><1><INTRODUCTION AND DEFINITIONS>
<Paragraph>
(1) It will be seen that this theory makes the radius of the uranium nucleus very small, about for [7 x 10^-13 cm]. (2) It sounds incredible but may not be impossible (Rutherford 1929).#
<EndParagraph>
<Section><1.1><INTRODUCTION TO THE STUDY OF NUCLEAR SIZES>
<Paragraph>
(3) The determination of nuclear shapes and sizes is one of the traditional problems of nuclear physics. ..... 
<EndParagraph>

Our mark-up is similar to Standard Generalised Mark-up Language (SGML). Such protocols help in encoding the layout structure of text together with its other attributes like author, title, etc. Parsing a marked-up text is an operation that involves extracting the different textual components and storing the results in an indexed file that contains indexed references to sentences, sections, paragraphs, tables, and figures in the text. Once a text is parsed and indexed in the above manner, one can use a summarisation procedure not only at a sentence level, but also at paragraph, section, and chapter levels.
```
5.1.2 Extracting patterns of cohesion

We have mentioned in Chapter 3 that lexical cohesion in text is achieved by a number of repetitions, particularly simple lexical repetition, complex repetition, and paraphrase.

The instances of simple lexical repetition can be extracted by producing a frequency list corresponding to the given text. A simple repetition of a word or a term is established if the word or the term occurs more than once in the text. If the word ‘distributions’, for example, occurs in the sentences [6,7,8], it establishes the simple lexical links 6-7, 6-8, and 6-7. The instances of complex repetition and paraphrase are extracted by comparing the words of a sentence with those of another. Subsequently, one needs a frequency list containing all the words in a text to be able to extract the lexical cohesion links between the sentences of such a text. Consider for example the frequency list shown in Table 5-2.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word/Term</th>
<th>Sentence Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>distribution</td>
<td>[4, 5, 11, 19, 28, 31, 35, 41, 42, 45, 46, 47, 58, 59, 61, 62, 63, 66, 67, 72, 75, 76, 78, 80, 84, 88, 91, 92, 94, 96]</td>
</tr>
<tr>
<td>21</td>
<td>nuclear</td>
<td>[3, 4, 5, 10, 11, 12, 13, 14, 16, 17, 18, 46, 47, 56, 73, 76, 91, 96, 97, 98, 99]</td>
</tr>
<tr>
<td>21</td>
<td>charge</td>
<td>[4, 5, 10, 11, 12, 13, 16, 17, 20, 22, 28, 46, 47, 56, 73, 76, 91, 96, 97, 98, 99]</td>
</tr>
<tr>
<td>17</td>
<td>distributions</td>
<td>[6, 7, 8, 12, 22, 28, 41, 43, 44, 55, 66, 72, 76, 80, 81, 82, 99]</td>
</tr>
<tr>
<td>4</td>
<td>determination</td>
<td>[3,7,10,16]</td>
</tr>
<tr>
<td>4</td>
<td>average</td>
<td>[4,92,96]</td>
</tr>
<tr>
<td>4</td>
<td>study</td>
<td>[5,12,35,66]</td>
</tr>
<tr>
<td>3</td>
<td>make</td>
<td>[4,22,73]</td>
</tr>
<tr>
<td>3</td>
<td>role</td>
<td>[4,87]</td>
</tr>
<tr>
<td>3</td>
<td>means</td>
<td>[5,13,36]</td>
</tr>
<tr>
<td>2</td>
<td>strong</td>
<td>[5,23]</td>
</tr>
<tr>
<td>2</td>
<td>shape</td>
<td>[7,48]</td>
</tr>
<tr>
<td>2</td>
<td>determine</td>
<td>[8,19]</td>
</tr>
<tr>
<td>1</td>
<td>shapes</td>
<td>[3]</td>
</tr>
<tr>
<td>1</td>
<td>features</td>
<td>[4]</td>
</tr>
<tr>
<td>1</td>
<td>difference</td>
<td>[4]</td>
</tr>
</tbody>
</table>

Table 5-2: A word frequency list.

In addition to the list shown in the table above, morphological analysis reveals that word pairs like determine and determination, and shape and shapes are instances of complex lexical repetition. The instances of simple mutual paraphrase are identified by checking
pairs of words for possible thesaural relationships, particularly synonymy. The existing lexical links between words lead to the establishment of lexical cohesion links between all sentences containing these words. Examples of these links are shown in Table 5-3 below.

<table>
<thead>
<tr>
<th>Types of links</th>
<th>Words/Word pairs</th>
<th>Established links between sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple lexical repetition</td>
<td>distributions, charge</td>
<td>6-7, 6-8, 7-8, 6-12, ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4-5, 4-10, 5-10, 5-12, 10-11, ...</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Complex lexical repetition</td>
<td>distributions-distribution,</td>
<td>4-6, 4-7, 4-8, 5-6, 5-7, ...</td>
</tr>
<tr>
<td></td>
<td>determine-determination,</td>
<td>3-8, 3-19, 7-8, 7-19, 10-8, ...</td>
</tr>
<tr>
<td></td>
<td>shape-shapes</td>
<td>3-7, 3-48, ...</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Simpl mutual paraphrase</td>
<td>shapes-make, make-shape,</td>
<td>3-4, ...</td>
</tr>
<tr>
<td></td>
<td>determine-study</td>
<td>4-7, ...</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>5-8, ...</td>
</tr>
</tbody>
</table>

Table 5-3: The extraction of lexical cohesion patterns.

Note that most of links shown in the table above are simple lexical repetition links. In fact, this observation becomes clearer as we show the analysis of all sentences in the text. We will come back to this point later in this chapter. Note also that the automatic use of the thesaurus can sometimes lead to the extraction of links that are rather ambiguous and misleading. We refer in this respect to the simple mutual paraphrase links *make-shape* and *shapes-make*. The relationships between these pairs of words are very suspect. The word *shapes* is treated as a noun in the text whereas the word *make* is actually used as a verb. Ambiguities like these are likely to occur during the analysis of a text and cannot be avoided unless one is equipped with a grammar parser that identifies the syntactic roles of every word in a sentence—something which is rather arduous and, at times, can lead to more ambiguities due to multiple parse trees. Nevertheless, suspect links and the like are not harmful. We will show later in this
chapter that, indeed, simple mutual paraphrase accounts for very few links compared to complex and simple lexical repetitions to the extent that one can ignore it altogether.

5.1.3 Building bonds and the bond matrix

The above discussion on lexical cohesion links is aimed at generating the bond matrix and the bond networks corresponding to the text. We mentioned earlier that in order to have a bonded pair of sentences, the two sentences must at least have a minimum number of links (i.e. the bond threshold) between themselves. Consider for instance a threshold value of two. In other words, a bond is established between two sentences that have at least two lexical cohesion links. Table 5-4 shows examples of bonded sentences corresponding to this threshold.

<table>
<thead>
<tr>
<th>Bonded sentences</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-4</td>
<td>1(smp), 1(sr)</td>
</tr>
<tr>
<td>3-7</td>
<td>1(sr), 1(cr)</td>
</tr>
<tr>
<td>4-5</td>
<td>3(sr)</td>
</tr>
<tr>
<td>4-7</td>
<td>1(cr), 1(smp)</td>
</tr>
<tr>
<td>7-8</td>
<td>1(sr), 1(cr)</td>
</tr>
<tr>
<td>5-8</td>
<td>1(smp), 1(cr)</td>
</tr>
</tbody>
</table>

Table 5-4: Some of the bonded sentences in the Jackson and Barret text (1977). Only sentences that have two or more links are shown. (sr: simple repetition, cr: complex repetition, smp: simple mutual paraphrase)

To see the effect of the variation of the threshold value over the bonding in these sentences, we illustrate bond networks for different threshold values. The results are shown in Figure 5-1. Recall that in a bond network, the nodes represent sentences and the arcs represent bonds between pairs of sentences.
Figure 5-1: Bond networks for different threshold values.

Figure 5-1 shows that the threshold value plays an important role in bonding sentences. At a threshold value of one, every lexical link is a bond and therefore most of the sentences appear to be bonded. As the threshold increases, fewer and fewer sentences remain 'bonded'. Thus, a central sentence at a threshold value of unity, may yet become marginal at higher values of the threshold. The choice of the threshold value may not be arbitrary but rather deliberate. We believe that such a choice should be based primarily on the length of the text and, perhaps, on its type.

The bond threshold is not the only factor involved in establishing bonds between sentences. The types of the lexical cohesion links have a major effect on the number and existence of these bonds. We elaborate on this point in the next section.
5.1.4 Relative contribution of lexical cohesion links

We have demonstrated how links are formed by selecting a small chunk of text comprising the first 8 sentences. We show here that the results can be scaled up to longer texts. In section 5.2, we discuss the results of the analysis of the other four texts and compare these results with those obtained from the first text.

The discussion of cohesion has thus far dealt with three major cohesion links: simple and complex repetition and simple mutual paraphrase. Note that the contribution of the links becomes more complex, starting from the simple frequency count for simple lexical repetition to the matching of morphological variants of the same word for complex repetition links. The most complicated computation is that of simple mutual paraphrase which may involve hyponymous relationships between two words (e.g. *mammal* and *cat*, *particle* and *proton*) on the one hand and idiomatic paraphrase on the other. Simple lexical repetition does not require any lexica, whereas for complex repetition, one needs a lexicon as comprehensive as Wordnet complete with a range of morphosyntactic data about thousands of words. And, for simple mutual paraphrase, we need a semantic thesaurus, a term base, or an encyclopaedic dictionary.

Now, if all these links make equally significant contributions to cohesion in a given text, and by implication to the summarisation procedures outlined in chapter 4, then our methodology would be quite resource intensive in that it would require lexica and encyclopaedias. This, however is not borne out by our investigation in that simple lexical repetition links contribute quite significantly to cohesion in text followed by complex repetition links. Simple mutual paraphrases play an important role at lower thresholds and, at higher thresholds, these effects seem to be vanishingly small.
Before we look at our results in some detail, a word about how bonding in text has been computed by TELE-PATTAN in the light of the fact that sentences may be linked to each other through more than one cohesion link (cf. Table 5-4). We have seen that sentence 3 is bonded to sentence 4 through one simple mutual paraphrase link and one simple repetition link, whereas sentences 4 and 5 are bonded through three simple lexical repetition links).

If we looked at only one type of lexical cohesion link, say simple lexical repetition, and ignore the complex repetition and simple mutual paraphrase, then the total number of bonds due to simple lexical repetition links would be smaller in magnitude in contrast to a situation where all links were counted. Suppose that the bond threshold is set at two links, and that we only count simple lexical repetition links, then only one pair of the sentences shown in Figure 5-1 would be bonded (sentences 4 and 5, because all other pairs have only one or no simple lexical repetition links). However, if one counts complex repetition together with simple lexical repetition links, then at a threshold value of 2, we would have three bonds (i.e. sentences 3 and 7, sentences 4 and 5, and sentences 7 and 8). Similarly, an increase in the number of bonds is noted when one allows simple mutual paraphrase to be counted. Table 5-5, based on Table 5-4, sums up the above discussion.
Table 5-5: An example of bond formation at a bond threshold of 2.
(\(\checkmark\) indicates the corresponding sentences are bonded, and the symbol \(\times\) indicates no bonding because the number of links is less than the threshold).

<table>
<thead>
<tr>
<th>Bond Thresholds</th>
<th>Number of bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2813 1886 626 3427 3064 2343 3628</td>
</tr>
<tr>
<td>2</td>
<td>1530 710 71 2109 1633 810 2204</td>
</tr>
<tr>
<td>3</td>
<td>695 162 16 1013 767 218 1101</td>
</tr>
<tr>
<td>4</td>
<td>417 80 2 663 460 101 702</td>
</tr>
<tr>
<td>5</td>
<td>232 29 3 383 259 33 417</td>
</tr>
<tr>
<td>6</td>
<td>165 19 1 267 170 22 276</td>
</tr>
</tbody>
</table>

Table 5-6: The distribution of bonds in the 101 sentence text.
(sr: simple repetition, cr: complex repetition, smp: simple mutual paraphrase)

The number of bonds formed by complex repetition and simple mutual paraphrase decreases drastically as the bond threshold increases. The formation of bonds by simple repetition in conjunction with complex repetition (i.e. sr+cr) is more significant than that with simple mutual paraphrase (i.e. sr+smp). Perhaps, this means that authors tend to repeat morphological variants more than they do synonyms. Furthermore, the most...
significant formation is the global one which takes place when all lexical cohesion links are taken into consideration (i.e. sr+cr+smp). Although, there may not be much difference between the contribution of sr+cr+smp links and that of simple lexical repetition (i.e. sr), the use of morphological variants and mutual paraphrasing has an important role in building a cohesive text, and hence the formation of more bonds.

Consider now the relative contributions of lexical cohesion types to the formation of bonds as compared to the global formation (i.e. that of sr+cr+smp). The percentages shown in Table 5-7 below are computed from the numbers in Table 5-7. We can clearly see that the contribution of simple lexical repetition, either individually or in conjunction with other links, is very significant and that it tends to be the in the same range regardless of the bond threshold. Simple mutual paraphrase (smp), which may lead to the appearance of ambiguous links as we have pointed out earlier in the case of the shape-make link, has little or no contribution particularly at higher threshold values.

<table>
<thead>
<tr>
<th>Bond Thresholds</th>
<th>Percentages of contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sr</td>
</tr>
<tr>
<td>1</td>
<td>78%</td>
</tr>
<tr>
<td>2</td>
<td>69%</td>
</tr>
<tr>
<td>3</td>
<td>63%</td>
</tr>
<tr>
<td>4</td>
<td>59%</td>
</tr>
<tr>
<td>5</td>
<td>56%</td>
</tr>
<tr>
<td>6</td>
<td>60%</td>
</tr>
</tbody>
</table>

Table 5-7: The individual and joint contributions of lexical links to the formation of bonds. (sr: simple repetition, cr: complex repetition, smp: simple mutual paraphrase)

We have mentioned previously that the bond threshold is a measure of the cohesion of a given text which we believe can be related to the length of the text. Consequently, for an even longer text where the threshold value should preferably be 3 or more, the
contribution of simple mutual paraphrase is relatively negligible compared to that of simple lexical repetition or complex repetition.

Generally, one can see that the number of bonds decays as the threshold decreases for all lexical cohesion links and that this decay is sharper for simple mutual paraphrase. This is illustrated in Figure 5-2.

![Figure 5-2: The individual contributions to the bond formation.](image)

The contribution of simple lexical repetition and complex repetition links (sr+cr) and the contribution of all lexical cohesion links (sr+cr+smp) have a similar trend of variation and decay. However, they are higher than the contribution of simple lexical repetition links. Figure 5-3 shows the difference in the contribution and decay as a function of the bond threshold for all lexical cohesion links both individually and collectively. This figure shows that the individual contribution of simple lexical repetition (sr), the sr+cr contribution, and the sr+cr+smp contribution are the most significant factors responsible for establishing the cohesion of the 101 sentence text.
Figure 5-3: The overall contributions of lexical cohesion to the formation of bonds.

The above discussion shows that the maximum number of bonds is formed when all the lexical cohesion links are taken into account. We have argued in chapter 4 that the notion of sentence categorisation that constitutes the basis for cohesion-based text summarisation approach relies on the presence of lexical cohesion links and, particularly, on the number of bonds that sentences of a given text have with each other. Subsequently, in our discussion on text summarisation in this thesis, we consider the contribution of all the lexical cohesion links, i.e. simple lexical repetition, complex repetition, and simple mutual paraphrase, knowing that some of these carry more weight than others.

5.2 Experimental results: An analysis of four texts

We have shown, in our discussion of the results of the analysis of the first text, that simple lexical repetition contributes the most to the formation of bonds as compared to complex repetition and simple mutual paraphrase. We have also shown that the contribution of simple repetition, complex repetition, and simple mutual paraphrase produces the highest number of bonds. The question now is: what does the analysis of other longer texts reveal?
Figure 5-4 shows the contribution of all the lexical cohesion types to the formation of bonds at bond thresholds ranging from 1 to 6 for all the five texts that have been analysed.

Figure 5-4: The variation of the number of bonds formed by the joint and individual contributions of lexical cohesion types in the five different texts. (a) Text-1, (b) Text-2, (c) Text-3, (d) Text-4, (e) Text-5.

The graphs shown in the figure above indicate clearly that the most significant contribution to the formation of lexical cohesion bonds in the five texts is that of 'sr+cr+smp', i.e. the contribution of simple lexical repetition, complex repetition, and simple mutual paraphrase. The contribution of simple lexical repetition and complex repetition is the next most significant. It is actually almost the same as that of
'sr+cr+smp'. The graphs reveal also that the simple mutual paraphrase contribution to the bond formation is negligible in almost all five texts except, perhaps, in the situation where the bond threshold is one or two.

Furthermore, the number of bonds shown in Figure 5-4 appears to follow the same trend in all five texts regardless of the lexical cohesion types. A best-fit that is adequate for such a trend has the following numerical form:

\[ \text{Number of Bonds} = \alpha \times 10^{4+\beta x} \]

where \( x \) represents the bond threshold and the parameters, \( \alpha \) and \( \beta \), depend on the size of the text and the lexical cohesion types involved in the bond formation. Typical values of these parameters are \((\alpha=1.6)\) and \((\beta=0.5)\). This estimation is important because it helps in choosing the bond threshold. We mentioned earlier that the bond threshold may vary depending on the size of the text, but we did not say how. We argue that this threshold should preferably not exceed the value of 10. This is notably the limit at which the number of bonds starts vanishing.

In order to illustrate the relative contributions of the different lexical cohesion types to the formation of bonds in text, every contribution is compared to the global contribution which is that of 'sr+cr+smp' by evaluating the ratio of the number of bonds they both form. The results obtained for the five texts are shown in Figure 5-5. The purpose of illustrating the relative contributions is to establish an order of importance of the lexical cohesion types in the formation of bonds.
The figure above clearly indicates that there is a regular order for almost all the lexical cohesion types which is the following one:

1. simple lexical repetition+complex repetition+simple mutual paraphrase
2. simple lexical repetition+complex repetition
3. simple lexical repetition+simple mutual paraphrase
4. simple lexical repetition only
5. complex repetition+simple mutual paraphrase
6. complex repetition only
7. simple mutual paraphrase only

The ‘pecking order’ obtained above can be used to judge the degree of cohesion of a given text. In other words, the more a text contains all of the three lexical cohesion
types, the more cohesive it is. The ordering also reveals that one can actually ignore the simple mutual paraphrase altogether as, in most cases, its individual contribution to the formation of bonds is insignificant and its contribution with other links has a minor effect.

5.3 Summary production

The objective of our research is to establish a methodology for analysing and, most importantly, summarising text using the notion of lexical cohesion and patterns of lexis in text. We have argued that it is possible to generate text summaries by extracting sentences that represent the significant body of the text. The phrase ‘significant body of the text’ in our context means the most cohesive part of the text. With the observation that authors have the tendency to repeat words, expressions, and phrases in the process of producing a text, our study of such repetition patterns reveals that there are indeed central sentences as well as potentially marginal sentences within a text.

In this section, we describe the process of generating summaries using procedure-4 which consists of ignoring marginal sentences. We show the sentence bonding at four thresholds and see how the number of marginal sentences increases as the threshold value increases, leading subsequently to the production of shorter and shorter summaries. Then, we show how to produce a shorter summary by extracting only the ‘key central’ sentences from a text (i.e. a procedure-3 summary). We will discuss, at the end of this section, the potential problem with adopting a ‘single sentence’ extraction strategy to produce summaries.
5.3.1 Procedure-4 summaries

Marginal sentences, we recall, are those sentences in the bond matrix corresponding to a given text which have no bonds with any sentence in the text. We mentioned earlier that the extraction of these sentences can be tuned by using a marginality threshold which usually is a small value (c. zero). This threshold which indicates the total number of bonds that a potential marginal sentence has, should not to be confused with the bond threshold which specifies the number of lexical cohesion links needed to form a bond between two sentences.

**Bond Threshold = 1**

This value of the bond threshold means that one lexical cohesion link between two sentences is sufficient to establish a bond between them. The sentence bonding corresponding to this threshold is shown in Table 5-8 below. The table shows that all sentences have at least one bond. There are therefore no marginal sentences, and subsequently the summary is literally the same as the original text.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.75</td>
<td>2</td>
<td>0.38</td>
<td>2.93</td>
<td>2.78</td>
<td>2.81</td>
</tr>
<tr>
<td>7</td>
<td>6.86</td>
<td>8</td>
<td>4.103</td>
<td>6.73</td>
<td>10.78</td>
<td>11.87</td>
</tr>
<tr>
<td>13</td>
<td>13.77</td>
<td>14</td>
<td>14.58</td>
<td>15</td>
<td>15.13</td>
<td>16</td>
</tr>
<tr>
<td>19</td>
<td>22.81</td>
<td>20</td>
<td>18.65</td>
<td>21</td>
<td>10.26</td>
<td>22</td>
</tr>
<tr>
<td>25</td>
<td>25.74</td>
<td>26</td>
<td>22.64</td>
<td>27</td>
<td>31.62</td>
<td>28</td>
</tr>
<tr>
<td>31</td>
<td>34.77</td>
<td>32</td>
<td>34.72</td>
<td>33</td>
<td>33.40</td>
<td>34</td>
</tr>
<tr>
<td>37</td>
<td>30.27</td>
<td>38</td>
<td>33.18</td>
<td>39</td>
<td>15.16</td>
<td>40</td>
</tr>
<tr>
<td>43</td>
<td>51.54</td>
<td>44</td>
<td>39.53</td>
<td>45</td>
<td>37.36</td>
<td>46</td>
</tr>
<tr>
<td>49</td>
<td>37.36</td>
<td>50</td>
<td>17.29</td>
<td>51</td>
<td>27.5</td>
<td>52</td>
</tr>
<tr>
<td>55</td>
<td>29.8</td>
<td>56</td>
<td>36.37</td>
<td>57</td>
<td>47.19</td>
<td>58</td>
</tr>
<tr>
<td>61</td>
<td>53.45</td>
<td>62</td>
<td>41.37</td>
<td>63</td>
<td>46.29</td>
<td>64</td>
</tr>
<tr>
<td>67</td>
<td>10.2</td>
<td>68</td>
<td>61.26</td>
<td>69</td>
<td>45.23</td>
<td>70</td>
</tr>
<tr>
<td>73</td>
<td>33.26</td>
<td>74</td>
<td>44.21</td>
<td>75</td>
<td>51.21</td>
<td>76</td>
</tr>
<tr>
<td>79</td>
<td>60.15</td>
<td>80</td>
<td>67.22</td>
<td>81</td>
<td>41.10</td>
<td>82</td>
</tr>
<tr>
<td>85</td>
<td>66.14</td>
<td>86</td>
<td>63.14</td>
<td>87</td>
<td>55.8</td>
<td>88</td>
</tr>
<tr>
<td>91</td>
<td>23.4</td>
<td>92</td>
<td>60.6</td>
<td>93</td>
<td>98.7</td>
<td>94</td>
</tr>
<tr>
<td>97</td>
<td>6.0</td>
<td>98</td>
<td>102.3</td>
<td>99</td>
<td>75.4</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5-8: The sentence bonding at a bond threshold of 1.
**Bond Threshold = 3**

Increasing the bond threshold to a value of three means that a bond is formed out of at least three lexical cohesion links. In this case, the sentence bonding which is shown in Table 5-9 indicates that there are eight marginal sentences.

<table>
<thead>
<tr>
<th></th>
<th>1 [0,8]</th>
<th>3 [0,24]</th>
<th>4 [3,64]</th>
<th>5 [2,60]</th>
<th>6 [0,10]</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>[3,20]</td>
<td>[1,18]</td>
<td>[2,15]</td>
<td>[4,38]</td>
<td>[11,7,4]</td>
</tr>
<tr>
<td>13</td>
<td>[5,27]</td>
<td>[14,7,17]</td>
<td>[16,9,15]</td>
<td>[17,9,14]</td>
<td>[18,9,13]</td>
</tr>
<tr>
<td>19</td>
<td>[10,35]</td>
<td>[20,9,23]</td>
<td>[21,4,6]</td>
<td>[22,22,47]</td>
<td>[23,8,12]</td>
</tr>
<tr>
<td>25</td>
<td>[16,18]</td>
<td>[26,13,28]</td>
<td>[27,12,13]</td>
<td>[28,15,33]</td>
<td>[29,17,11]</td>
</tr>
<tr>
<td>31</td>
<td>[19,28]</td>
<td>[32,20,23]</td>
<td>[33,21,13]</td>
<td>[34,5,7]</td>
<td>[35,20,25]</td>
</tr>
<tr>
<td>37</td>
<td>[15,6]</td>
<td>[38,18,3]</td>
<td>[39,3,0]</td>
<td>[40,2,6]</td>
<td>[41,15,28]</td>
</tr>
<tr>
<td>43</td>
<td>[20,18]</td>
<td>[44,7,7]</td>
<td>[45,22,15]</td>
<td>[46,10,2]</td>
<td>[47,29,16]</td>
</tr>
<tr>
<td>49</td>
<td>[0,8]</td>
<td>[50,3,0]</td>
<td>[51,6,0]</td>
<td>[52,6,0]</td>
<td>[54,14,1]</td>
</tr>
<tr>
<td>55</td>
<td>[15,0]</td>
<td>[56,2,3]</td>
<td>[57,11,4]</td>
<td>[58,9,2]</td>
<td>[59,0,2]</td>
</tr>
<tr>
<td>61</td>
<td>[25,16]</td>
<td>[62,2,9]</td>
<td>[63,6,4]</td>
<td>[64,4,3]</td>
<td>[65,18,6]</td>
</tr>
<tr>
<td></td>
<td>[68,28,12]</td>
<td>70 [0,8]</td>
<td>71 [2,2]</td>
<td>72 [0,2]</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>[2,6]</td>
<td>[74,14,9]</td>
<td>[75,33,10]</td>
<td>[76,2,0]</td>
<td>[77,0,1]</td>
</tr>
<tr>
<td>79</td>
<td>[21,5]</td>
<td>[80,10,0]</td>
<td>[81,2,0]</td>
<td>[82,18,6]</td>
<td>[83,9,0]</td>
</tr>
<tr>
<td>85</td>
<td>[6,2]</td>
<td>[86,52,9]</td>
<td>[87,3,2]</td>
<td>[88,9,2]</td>
<td>[89,6,0]</td>
</tr>
<tr>
<td>91</td>
<td>[4,0]</td>
<td>[92,2,0]</td>
<td>[93,27,3]</td>
<td>[94,41,2]</td>
<td>96 [19,2]</td>
</tr>
<tr>
<td>98</td>
<td>[41,4]</td>
<td>[99,16,0]</td>
<td>[100,6,0]</td>
<td>[101,15,0]</td>
<td></td>
</tr>
</tbody>
</table>

Table 5-9: The sentence bonding at a threshold value of 3.

The marginal sentences which represent the empty cells are [2, 15, 53, 66, 67, 69, 95, 97]. A summary produced by ignoring this small set of marginal sentences would be slightly shorter than the original text itself (exactly by 8 sentences).

**Bond Threshold = 6**

Consider now a bond threshold of six. All bonded sentences in this case are shown in Table 5-10 below. Note that there are more marginal sentences (shown here in empty cells) than in the case of a threshold value of three. The following list includes all the 42 marginal sentences found.

[1, 2, 14, 15, 21, 39, 40, 46, 49, 50, 52, 53, 56, 57, 59, 60, 62, 63, 64, 65, 66, 67, 69, 70, 71, 72, 73, 76, 77, 80, 81, 83, 84, 85, 88, 89, 90, 91, 92, 95, 97, 100].

- 103 -
The exclusion of the marginal sentences leads to the production of a summary which is 59 sentences long (i.e. 101-42). This summary consists of the following sentences:

\[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 41, 42, 43, 44, 45, 47, 48, 51, 54, 55, 58, 61, 68, 74, 75, 78, 79, 82, 86, 87, 93, 94, 96, 98, 99, 101\]

This summary is clearly shorter than the one mentioned above (threshold = 3). However, this summary can still be considered long. To reduce its size, one can either increase the bond threshold to a higher value (>6), or raise the marginality threshold to 1 or even 2. A marginality threshold of 2, for instance, means that every sentence in the text which has 2 bonds or less can be considered marginal.

**Bond Threshold = 9**

When we increase the bond threshold to a value of 9 and chose a marginality threshold of 2, we obtain the sentence bonding shown in Table 5-11 below. For these threshold values, 81 sentences are considered marginal.

<table>
<thead>
<tr>
<th>Bond Threshold</th>
<th>3 [0,5]</th>
<th>4 [2.47]</th>
<th>5 [2.25]</th>
<th>6 [0.3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>37 [1,0]</td>
<td>38 [2,0]</td>
<td>41 [3,0]</td>
<td>42 [10,2]</td>
<td></td>
</tr>
<tr>
<td>51 [4,0]</td>
<td>58 [2,0]</td>
<td>54 [1,2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55 [6,0]</td>
<td>61 [0,4]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68 [5,0]</td>
<td>74 [0,5]</td>
<td>75 [0,3]</td>
<td>78 [18,6]</td>
<td></td>
</tr>
<tr>
<td>79 [1,0]</td>
<td>82 [1,0]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86 [11,5]</td>
<td>87 [2,0]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>93 [7,0]</td>
<td>94 [9,0]</td>
<td>96 [2,0]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98 [22,2]</td>
<td>99 [2,0]</td>
<td>101 [1,0]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5-11: The sentence bonding at a threshold value of 9.

<table>
<thead>
<tr>
<th></th>
<th>4 [0,25]</th>
<th>5 [1,13]</th>
<th>11 [1,3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 [4,2]</td>
<td>33 [8,0]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>74 [0,3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86 [3,0]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98 [5,0]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We can reduce this summary even further by tuning the thresholds. The question is: to what extent? Considering sentences which have, say 2 or 3 bonds, as marginal may be acceptable for a 100 sentence long text. It may also be acceptable for longer texts. However, for shorter texts, higher marginality thresholds may lead to the exclusion of sentences which have less bonds but are crucial to the understanding of the topic of the text. Rather we would be in favour of tuning the bond threshold in that this threshold actually enhances the notion of cohesion between the sentences that one deems appropriate for a summary. Even then, one has to limit the bond threshold because, as we pointed out before, the total number of bonds in text tends to drop drastically as the bond threshold increases.

The definition of the bond threshold, that is the minimum number of lexical cohesion links that form a bond, implies an inverse relationship between the threshold and the number of bonds that may exist within a text. Given that the size of a summary produced using the procedures mentioned above depends upon the number of bonds, it is perhaps obvious that the size of such a summary will be a function of the bond threshold. Indeed, such a systematic variation has been observed by us for low bond thresholds as shown in Table 5-12 below.

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>Summary Sizes (in sentences)</th>
<th>Original Text sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Text-1</td>
<td>101</td>
<td>98</td>
</tr>
<tr>
<td>Text-2</td>
<td>202</td>
<td>188</td>
</tr>
<tr>
<td>Text-3</td>
<td>214</td>
<td>208</td>
</tr>
<tr>
<td>Text-4</td>
<td>610</td>
<td>582</td>
</tr>
<tr>
<td>Text-5</td>
<td>2627</td>
<td>2615</td>
</tr>
</tbody>
</table>

Table 5-12: The variation of the size of procedure-4 summaries at low thresholds.
It is important to note that we have confirmed Hoey’s empirical observation that summaries should be produced using a bond threshold of three links. Our evaluation shows this to be the case.

Our experiments for higher bond thresholds (between 3 and 9) show that the inverse relationship between the summary size and the bond threshold still holds but it becomes highly non-linear and varies with the text size; this is not the case for lower bond thresholds. Note that the summaries of text-2 (which contains more equations than all the other texts) show irregularities even at lower thresholds.

Given the fact that the number of bonds decreases as the bond threshold increases, one can argue that the size of the summaries produced using either of the previously mentioned procedures also decreases. Furthermore, our experimental results obtained from the analysis of the five texts indicate that for bond threshold values between one and three, the sizes of procedure-4 summaries appears to follow the same trend. However, this may not be true for threshold values higher than three.

5.3.2 Producing short summaries

We have mentioned previously that it is possible to produce a summary out of central sentences by extracting the ‘key central’ sentences. A example of this type of summary is shown below:

22 It will, however, be necessary to use the information so far obtained for nucleons in order to make a connection between the predictions of theories for nuclear distributions due to point nucleons and the observed nuclear charge and magnetic moment distributions due to nucleons with finite electromagnetic size.

78 For the charge and matter distributions the value of the volume integral will be \(|Z|\) or \(|A|\), respectively, or unity, depending on the normalization condition, and this determines the physical significance of the uniform radius \((U>\) as the radius of the uniform sphere which contains the same amount of matter or charge as the real distribution.

86 Elton (1961b) introduced a basic length which is related to the uniform radius by \(\text{eq. } 1.34\) and uses the symbol \(R\) to denote the radius of the uniform distribution which has the same r.m.s radius as the true distribution so that \(R = Q\).
The notion of key centrality deals with the extraction of sentences that are thought to be the most central sentences in text. Key central sentences have the highest cohesive bonding in a text, and subsequently, may include the key points, that the author intends to convey. Summaries based on the extraction of this type of sentences are by and large shorter than all the summaries produced using other procedures and may be considered as short indicative abstracts.

5.4 The evaluation of summaries

The evaluation of summaries, particularly automatically generated summaries, is an open question and has not been given much attention from workers in artificial intelligence and in information retrieval. This is, perhaps, due to the fact that the evaluation itself requires deep understanding of natural language and usually requires subjective matters of taste and style, etc. The information retrieval community is now starting to put more emphasis on questions related to the evaluation of retrieval techniques besides effectiveness measurement methods such as precision and recall (Salton, 1989). Automatic evaluation of summaries may be approached from an information retrieval perspective, particularly in the case of summaries that are primarily extracts from the original texts.

In our case, we have conducted an experiment whereby working scientists\(^5\) were asked to read to the summaries and the original texts and answer a number of questions. We asked three of these scientists who are physicists to read the summaries of the four texts whose subject matters cover different aspects of nuclear and elementary particle physics.

\(^5\) Dr. J. S. Al-khalili, Dr. J. A. Tostevin, and Dr. L. Tostevin (specialists in Nuclear and Elementary Particle Physics - Physics Department, University of Surrey); Dr. Roland Price (a specialist in Engineering and Applied Mathematics - Wallingford Software Hydraulics Research Ltd., Howbery Park, Wallingford, Oxon, OX10 8BA).
The fourth, an engineering scientist, was asked to read a summary of a report on a feasibility study related to the construction of a harbour (the longest text we have analysed). The questions were subdivided into three questionnaires. The first set of questions (Questionnaire-1) required reading the summary, then answering a set of questions related to the readability, fragmentation, content, and accuracy of the summary. In the second questionnaire, the scientists were asked to read the original text first, then answer another set of questions related to the readability of the original text and familiarity with its information content. The third questionnaire consisted of only one question which was a concluding comment on the accuracy of the summary after both the summary and the original text were read.

5.4.1 Evaluating the summary of the first text

The first text and its summary were read by three nuclear physicists. Their answers to the three questionnaires are outlined in Table 5-13 and Table 5-14. There are two major points that emerge from the answers to these questionnaires. First, we noted that two of the three readers finished reading the summary in a third of the time it took them to read the original text. The third reader took half the time in that respect.

Second, the three of them agreed that the summary included a topic on ‘the size of particles and nuclei’. Two of them identified three other topics that are ‘charge distribution’, ‘matter distribution’, and ‘scattering’. There were other topics such as ‘the importance of nuclear and particle structure’, ‘parametrisation’, and ‘models’ (of nuclear sizes and structure) that each of the readers thought were included in the summary. The title of Jackson and Barret’s book from which the first text was taken is ‘Nuclear Sizes and Structure’. One can see that the common topics identified in the
summary are closely related to the title of the book, and that these topics are expressed in almost the same wording as the title. Concerning the logical presentation of arguments, the readers agree that the summary needed some improvements. They all mentioned that there were very few pronouns in the summary.

<table>
<thead>
<tr>
<th>Questionnaire-1</th>
<th>Human Readers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1. How long did it take to read the summary?</td>
<td>20 minutes</td>
</tr>
<tr>
<td>2. What do you think were the main topics of discussion in the summary?</td>
<td>-Finite size of particles and nuclei. -Importance of their structure. -Nuclear potentials</td>
</tr>
<tr>
<td>3a. Are the arguments being presented logically?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Fragmented</td>
</tr>
<tr>
<td>3b. Did you find pronouns in the text that had no meaning?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Lost of them (&gt;5)</td>
</tr>
<tr>
<td>3c. Were you concerned that the summary did not have equations, graphs, or tables?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Not really</td>
</tr>
<tr>
<td>3d. How well does the summary read?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Fragmented</td>
</tr>
<tr>
<td>4. When you read a learned paper, do you expect that a) the abstract of the paper is</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>An excellent guide to the Content of the paper</td>
</tr>
<tr>
<td>5. How accurate was the summary in terms of subject content?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Pretty good</td>
</tr>
</tbody>
</table>

Table 5-13: The opinions of three nuclear physicists on the summary of the introduction chapter of Jackson and Barret's Book (Questionnaire-1).
The absence of equations, graphs, and tables from the summary was very irritating for the third reader and slightly so for the second. The third one said it was not really irritating. In general, the readers agreed that the summary was, in average, readable and that it had a quite accurate content in terms of subject matter.

The answers to questionnaires 2 and 3, shown in Table 5-14, revealed that the three readers who are familiar with the subject matter of the original text found that such an original text does not have to have a perfect readability to be understood. Two of them agreed, after reading both the original text and the summary, that the summary was adequate. The third reader, who found that the absence of equations, tables, or figures from the summary made it irritating to read, said that there were many points ignored (i.e. those equations).

<table>
<thead>
<tr>
<th>Questionnaire-2</th>
<th>Human Readers</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>How long did it take you to read the original text?</td>
</tr>
<tr>
<td></td>
<td>40 minutes</td>
</tr>
<tr>
<td>7</td>
<td>In your opinion, the original text is</td>
</tr>
<tr>
<td>a)</td>
<td>Readable</td>
</tr>
<tr>
<td></td>
<td>Has an average readability</td>
</tr>
<tr>
<td></td>
<td>Difficult to read</td>
</tr>
<tr>
<td>b)</td>
<td>Very long</td>
</tr>
<tr>
<td></td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>Okay</td>
</tr>
<tr>
<td>8</td>
<td>Does the text have information with which you are</td>
</tr>
<tr>
<td></td>
<td>Very familiar</td>
</tr>
<tr>
<td></td>
<td>Generally aware</td>
</tr>
<tr>
<td></td>
<td>Have lay persons’ knowledge</td>
</tr>
<tr>
<td></td>
<td>Not interested</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Questionnaire-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 5-14: The opinions of the three nuclear physicists on the accuracy of the summary of the introduction chapter of Jackson and Barret’s Book compared to the original text (Questionnaires-2 and 3).
5.4.2 Evaluating summaries of other texts - key results

Three other texts and the corresponding summaries were read by one nuclear physicist. His general comments on the three summaries were that they lacked co-ordination mainly because they did not contain equations. He mentioned however, that they were readable and that their subject content was quite accurate. The summaries were read in a third of the time it took to read the original texts. The topics that he identified in these summaries were almost the same as those reported in the original texts. Furthermore, in the case of the third text (i.e. a learned paper comprising more than 600 sentences), six topics were identified in the summary: ‘charge symmetry’, ‘charge symmetry breaking’, ‘meson exchange’, ‘quark mass difference’, ‘neutron rich nuclei’, ‘binding energies’, and ‘Coulomb’s interaction’. The keywords reported in the original paper were ‘charge independence’, ‘charge symmetry’, ‘mass difference’, ‘electromagnetic interactions’, and ‘symmetry breaking effects’. Note that the topics that reader identified in the summary amount to more that than the keywords themselves. For instance, the summary contained the topics ‘Coulomb’s interaction’ and ‘meson exchange’ which do not appear in the abstract nor the keyword list of the original paper. This observation indicates that our summaries can be used as text surrogates and even replace the abstract and keyword lists that are attached to papers and reports.

The fifth text comprising more than 2600 sentences and its summary were also read by one reader (i.e. the engineering scientist). His general comments on the summary were that it was very readable and that the information contained in it was indicated with a good accuracy. He was quite sure that the summary, which took him 10 minutes to read, could even be reduced in size and still be readable.
5.5 Problems with single sentence extraction summaries

The text summarisation methods which we have discussed so far are examples of what are now known as sentence extraction systems. This type of text summarisation, which is widely used in information retrieval and document relevance measurements, is based on term weighting methods. Our cohesion based sentence categorisation and extraction does perhaps share one common problem with conventional sentence extraction systems in the information retrieval community. This problem relates to the readability of the final extract. In our context of text summarisation, we have to evaluate the readability and cohesion of the extract. However, in information retrieval, the main concern is how relevant the extract is.

Cohesion based text summarisation which is based on sentence extraction has a potential problem related to the cohesion and readability of the extract. That is to say, how well the extracted sentences cohere with each other. Single sentences may belong to any paragraph within the text and may therefore lead to the production of discontinuous summaries. Perhaps, the only assurance of producing a cohesive extract is the knowledge that most of the sentences that form the extract have most of the lexical cohesion bonds found in a text.

One common problem in this respect is that of pronominal references. In writing a text, authors often make use of pronouns and demonstratives for the purpose of referring to items across sentences within a text. Summaries of a text based on single sentence extraction may contain sentences which are crucial for these references. However, previous studies have revealed that, in scientific texts, the use of pronouns is not as dominant as it is in narrative texts (Meyers & Hartley, 1990). Consider the following
central sentences that were obtained from the first text (i.e. Jackson and Barret' introductory chapter) at a high bond threshold value:

25 Following the development of the cyclotron, the energy dependence of α-particle scattering from heavy nuclei was studied up to 40 MeV (Farwell and Wegner 1954) and the abrupt departure from pure Coulomb scattering beyond a critical energy was interpreted in terms of a radius parameter (Blair 1954).

26 This radius parameter cannot be directly interpreted in terms of a nuclear matter radius, since the range of the potential must be connected with the finite range of nuclear forces and the size of the projectile.

28 They succeeded in reproducing neutron scattering data up to a few MeV with a complex square-well potential, but fits to angular distributions for 20 MeV protons required a potential with a diffuse surface resembling the surface of the nuclear charge distribution (Woods and Saxon 1954).

29 For scattering at ~100 MeV, Serber (1947) had suggested that the collision of an incident nucleon with nucleus could be interpreted in terms of collisions with individual nucleons, and Fernbach, Serber and Taylor (1949) analysed total neutron cross-sections effectively with a square-well potential whose imaginary (absorptive) part was related to the total cross-sections for nucleon-nucleon scattering.

31 The scattering of nucleons and pions from nuclei at energies ~1 GeV were interpreted in terms of a potential which was related in a fairly intuitive way to the nuclear matter distribution (Coor et al. 1955, Williams 1955, Abashian, Cool, and Cronin 1956).

47 In many situations involving electromagnetic interactions we actually require the nuclear charge distribution [*ch(r)*] of the nucleus instead of the distribution of point protons.

75 We have used the symbol U and term uniform radius in preference to Süssman's charge radius [R] in order to make the definition more general and avoid confusion with other quantities.

Note that sentences 28 and 75 actually begin with a pronoun or a demonstrative. Sentence 75 comprises a discourse-structuring recapitulating verb phrase ‘we have used’.

Now, if these sentences were used to produce a summary, then there would be unresolved pronominal and demonstrative references. Some discourse structure information would be missing. In order to compensate for the unresolved pronouns and demonstratives, the system TELE-PATTAN uses a very simple heuristic: each sentence that has an above threshold number of lexical cohesion links is included in the summary.

5.5.1 From central sentences to central paragraphs

In the attempt of solving the summary cohesion and readability problem, we believe that if there are central sentences in the text, then there should be central paragraphs and indeed central sections, etc. Producing summaries based on the extraction of paragraphs is, perhaps, preferable than selecting individual sentences from distant paragraphs in that these summaries would have a better readability.
The concept of keywords was a first step for producing summaries of documents. The concept of key sentences is partially based on keywords and partially based on the notion of cohesion. Given the concerns about the readability of the summaries produced using lexical cohesion analysis, namely the rather disjointed nature of the summaries produced and the related problems of pronominal references, it is probably in order here to discuss the notion of ‘central paragraphs’.

In one sense, the notion of ‘central paragraph’ is a bootstrapping on the notion of central sentences which in turn were bootstrapped on keywords and their repetitions. Intuitively and with some justification, one can argue that a paragraph containing, say $C$ central sentences, would carry much more information about the content of a document than a paragraph carrying none.

The number $C$ can be unity and, if assigned a value greater than one, then we can bring in the very important notion of the bond thresholds which played an important role in producing the summaries described in this thesis. One can extend the notion of central paragraphs by suggesting that we would regard a paragraph to be central if it contained either a $C$ number of central sentences, or a $TO$ number of topic opening sentences, or a $TC$ number of topic closing sentences. This argument can be extended further by noting that in some paragraphs the author may wish to open a new topic and use a previous topic to reinforce his/her argument. Alternatively, the author may, in deciding to close a topic, wish to introduce a new one.

The notion of central paragraphs may be used in a negative sense that is by arguing that if a paragraph contains $M$ or more marginal sentences, then that paragraph may be
ignored. Furthermore, one can argue that the presence of marginal sentences in a paragraph, despite the fact that it may comprise central, topic opening, and/or topic closing sentences, will reduce the relevance of such a paragraph.

It should be noted that we have not found any reference to the notion of central paragraphs in the literature on cohesion, although workers in information retrieval have toyed with this idea. The whole notion of linguistic units, be it words, single or compound, be it phrases, simple or complex, be it sentences, is a very rich one. The interdependence of constituent units with their encompassing structures, i.e. the words and their relationships to the sentence they form, has been and still is a source of fascination and curiosity for linguists and philosophers. Similarly, the notion of 'paragraph', itself a compound word (defined as a noun which indicates 'a division of a piece of writing which is made up of one or more sentences and begins with a new line'), and its relationship to its constituent sentences, has excited many literary critics, writers, experts, and so forth. Thus, our notion of central paragraphs should be seen as speculative. However, it nonetheless yields interesting results.

5.5.2 Extracting central paragraph using a possible weighting method
Central paragraphs may contain some or all of the sentence categories discussed earlier. The number of instances of these categories, as we have speculated, may affect the relevance of a given paragraph to the topic of the text. TELE-PATTAN can be used to specify a set of weights to evaluate the centrality of a paragraph and subsequently its relevance to the topic(s) of the text. The idea here is to associate a score to a paragraph depending on the number of central, topic opening, topic closing, and marginal sentences it contains. Assuming that the numbers of these sentences in a paragraph $p$
are respectively \( C_p, TOp, TCp, \) and \( M_p \), we can argue that the centrality score \( S_p \) of such a paragraph may be evaluated as follows:

\[
S_p = w_c \times C_p + w_{to} \times TOp + w_{tc} \times TCp + w_m \times M_p
\]

where \( w_c, w_{to}, w_{tc}, \) and \( w_m \) are arbitrary weights that are associated respectively with central, topic opening, topic closing, and marginal sentences. The weights are all positive except the one associated with marginal sentences. This restriction is deliberate since we intend to focus on central sentences rather than marginal sentences.

One way to compute the score of a given paragraph is to find out first the number of sentences of each of the four categories, i.e. topic opening, topic closing, central, marginal, that belong to such a paragraph. In Table 5-15 below, the paragraphs of the first text that we have analysed are shown with the different corresponding sentences from each category. The last column of the table shows the length and content of each paragraph.
Table 5-15: Computing paragraph centrality.

(with $w_c = 0.9$, $w_{to} = 0.5$, $w_{tc} = 0.5$, and $w_m = -0.25$)

The table above shows that the paragraphs of the first text have scores ranging from the value of -0.5 to 5.65. The paragraphs which contain more central sentences than others tend to have a higher score, e.g. paragraphs 5, 7, 8, etc. This score is even higher if the paragraphs contain topic opening or closing sentences in addition to central sentences as it is the case for paragraph 4. There are other paragraphs in the table that have a zero score. These paragraphs comprise sentences which do not fall under any category. The sentences of these paragraphs have smaller numbers of bonds but not small enough to be considered marginal. Some of the zero score paragraphs are 15, 16, 28, and 29. Note
that these paragraphs are relatively shorter than others. There are only two paragraphs which comprise simultaneously both central and marginal sentences, e.g. paragraphs 4 and 28. Paragraphs like these, we believe, may be found in cases where authors mention a comment or a reference in the middle of an elaboration paragraph.

5.5.3 Text summarisation based on paragraph extraction

The paragraphs of a text can be ranked in a descending order according to their centrality scores. This ranking shown in Table 5-16 below indicates which paragraphs have more central sentences, more topic sentences, and fewer marginal sentences. This table is derived from Table 5-15.

<table>
<thead>
<tr>
<th>Paragraphs</th>
<th>Length-Sentences</th>
<th>Score</th>
<th>Paragraphs</th>
<th>Length-Sentences</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>8-[9,10,11,12,13,14,15,16]</td>
<td>5.65</td>
<td>26</td>
<td>4-[82,83,84,85]</td>
<td>0.90</td>
</tr>
<tr>
<td>8</td>
<td>5-[27,28,29,30,31]</td>
<td>4.50</td>
<td>25</td>
<td>3-[79,80,81]</td>
<td>0.90</td>
</tr>
<tr>
<td>13</td>
<td>7-[43,44,45,46,47,48,49]</td>
<td>3.60</td>
<td>20</td>
<td>1-[68]</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>4-[23,24,25,26]</td>
<td>3.60</td>
<td>18</td>
<td>3-[63,64,65]</td>
<td>0.90</td>
</tr>
<tr>
<td>30</td>
<td>5-[93,94,95,96,97]</td>
<td>2.70</td>
<td>17</td>
<td>4-[59,60,61,62]</td>
<td>0.90</td>
</tr>
<tr>
<td>11</td>
<td>4-[36,37,38,39]</td>
<td>2.70</td>
<td>10</td>
<td>1-[35]</td>
<td>0.90</td>
</tr>
<tr>
<td>5</td>
<td>3-[17,18,19]</td>
<td>2.70</td>
<td>32</td>
<td>1-[101]</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>3-[20,21,22]</td>
<td>2.30</td>
<td>29</td>
<td>1-[92]</td>
<td>0.00</td>
</tr>
<tr>
<td>23</td>
<td>3-[74,75,76]</td>
<td>1.80</td>
<td>28</td>
<td>2-[90,91]</td>
<td>0.00</td>
</tr>
<tr>
<td>12</td>
<td>3-[40,41,42]</td>
<td>1.80</td>
<td>22</td>
<td>2-[72,73]</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>3-[32,33,34]</td>
<td>1.80</td>
<td>16</td>
<td>1-[58]</td>
<td>0.00</td>
</tr>
<tr>
<td>31</td>
<td>3-[98,99,100]</td>
<td>1.40</td>
<td>15</td>
<td>2-[56,57]</td>
<td>0.00</td>
</tr>
<tr>
<td>27</td>
<td>4-[86,87,88,89]</td>
<td>1.40</td>
<td>21</td>
<td>3-[69,70,71]</td>
<td>-0.25</td>
</tr>
<tr>
<td>24</td>
<td>2-[77,78]</td>
<td>1.40</td>
<td>14</td>
<td>6-[50,51,52,53,54,55]</td>
<td>-0.25</td>
</tr>
<tr>
<td>3</td>
<td>4-[5,6,7,8]</td>
<td>1.40</td>
<td>1</td>
<td>2-[1,2]</td>
<td>-0.25</td>
</tr>
<tr>
<td>2</td>
<td>2-[3,4]</td>
<td>1.40</td>
<td>19</td>
<td>2-[66,67]</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

Table 5-16: A paragraph centrality ranking.

The highly ranked paragraph which is paragraph 4 is the most central as it comprises the highest number of central sentences. Other higher ranking paragraphs are 8, 13, 7, 30, 11, 5, 6, 23, 12, 9, 31, 27, 3, 2, 26, 25, 20, 18, 18, 17, in a decreasing order of centrality. Paragraphs with a zero or a negative score are less central, and if they contain marginal sentences and have a negative score, they tend in turn to be marginal. The scores for all the paragraphs of the text are shown in Figure 5-6 below.
Text summaries can now be generated by selecting the most, or, some of the most central paragraphs in the text. A first candidate summary of the text is the selection of all paragraphs that have centrality scores higher than zero. This summary contains the paragraphs 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 18, 20, 23, 24, 25, 26, 27, 30, and 31 comprising in total 79 sentences. This is quite a long summary given the size of the original text (101 sentences). Another choice would be to include the paragraphs having the five highest centrality scores which are actually 4, 8, 13, 7, and 30. The produced summary includes the sentences [9, 10, 11, 12, 13, 14, 15, 16, 23, 24, 25, 26, 27, 28, 29, 30, 31, 43, 44, 45, 46, 47, 48, 49, 93, 94, 95, 96, 97] and is outlined below.
This summary contains topic opening sentences that are needed for a short introduction, central sentences, and topic closing sentences that close the summary. The difference between this summary, i.e. what we call a 'procedure-5 summary', and other summaries is that it is more cohesive and more readable in that full paragraphs are selected rather than individual sentences. This summary shown above comprises three marginal sentences which are 15, 95, and 97—a fairly negligible number given the size of the summary. These sentences would not harm the summary if they were removed.

5.5.4 Central paragraphs in longer texts
We present here a comparative discussion on the distribution of central paragraphs in the five texts that we have analysed. Figure 5-7 below illustrates the paragraph centrality for these texts. In the first text (i.e. Figure 5-7a), the most central paragraph is the fourth one. Then, in a decreasing order of centrality, paragraphs 8, 13, 7, and so on, follow. Note that paragraphs 3 and 30 have lower centrality values. These two
paragraphs, being physically located at the beginning and end of the text and being less central, can be thought of as topic opening and a topic closing paragraphs respectively. In the middle of the text, the centrality of paragraphs varies. One can argue however that the gist of the text may be expressed in those few paragraphs in the beginning of the text, e.g. 3, 4, 8, other paragraphs from the elaboration sections of the text, e.g. 11, 12, 13, 23, and the few closing paragraphs, e.g. 27, 30, and 31.

Figure 5-7: Central paragraphs in texts. (a) Text-1, (b) Text-2, (c) Text-3, (d) Text-4, (e) Text-5.
Text-2 and text-3 shown in Figure 5-7(b) and Figure 5-7(c) appear to have a wider range of central paragraphs than in the case of text-1. Many paragraphs have negative scores. In case (d), the central paragraphs appear to congregate in the middle of the text between paragraphs 20 and 75. Paragraph 3 has a centrality value of almost 1 which indicates subsequently the presence of topic opening sentences. Such a paragraph plays the same role as paragraph 3 in the text of case (a). In the fifth text shown in Figure 5-7(e), the appearance of a clear cut between positive and negative centrality scores indicates perhaps that the most important part of the text includes paragraphs between 240 to 420.

The location of central paragraphs in a text may be uncertain at this stage. This is perhaps due to the fact that it is the selection of the weights that are associated with the different categories of sentences which dictates what paragraphs may be considered central. The production of procedure-5 summaries (i.e. summaries generated using central paragraphs), therefore, depends on the selection of these weights. For instance, to ensure that a procedure-5 summary contains an introduction and a conclusion, the weights associated with the topic opening and topic closing sentences should have values closer to those of the weight associated with central sentences.

Producing text summaries using central paragraphs, we believe, has two major advantages. First, there is obviously no discontinuity between sentences of the same paragraph as they are originally cohesive solving therefore (at least partially) the readability problem of the summary. Second, this method is pragmatically plausible given the common knowledge that in writing a text, one has the tendency to provide an introduction or a conclusion in the form of paragraphs rather than sentences.
5.6 Conclusions

Our exemplar analysis of the nuclear and elementary particle physics text of more than 100 sentences and other longer texts has shown that patterns of lexis which ensure text cohesion do indeed exist, and at times, extensively. The distribution of these patterns varies depending on their types. It is apparent to us that simple lexical repetition in text produces more lexical cohesion links than complex lexical repetition and simple mutual paraphrase.

The lexical cohesion links relate pairs of words and are the components needed for the formation of bonds between pairs of sentences. We have shown that the individual and joint contributions of simple lexical repetition to the formation of bonds in a text appears to be the most significant as compared to other contributions. The analysis of the five texts reveals the pecking order of these contributions which is shown below.

1. simple lexical repetition+complex repetition+simple mutual paraphrase
2. simple lexical repetition+complex repetition
3. simple lexical repetition+simple mutual paraphrase
4. simple lexical repetition only
5. complex repetition+simple mutual paraphrase
6. complex repetition only
7. simple mutual paraphrase only

Most importantly, we note that the contribution of simple lexical repetition, be it individual or in conjunction with other lexical cohesion repetition types, has indeed the lion’s share of the number of bonds in a text and therefore assures of up to 90% of the lexical cohesion of the text. Although it is preferable that one should include complex lexical repetition and mutual paraphrase in the analysis, we believe that it is sufficient to consider simple lexical repetition, particularly if one is limited in machine readable resources such as thesauri or dictionaries. Even more striking is the point that, with the
exception of complex and simple mutual paraphrase, the analysis can be multi-lingual
and does indeed raise the possibility of summarising different texts written in languages
other than English.

We found out that the distribution of bonds follow an apparently regular pattern which
indicates an exponential decay. The regular distribution of bonds allows us to set an
upper bound to the bond threshold. We also found out that this upper bound should
preferably not be more than 10. Hoey first argued that a bond between two sentences
can be significant only if it is formed of at least three links setting therefore a lower
bound to the bond threshold. Our argument is that, for a bond to exist and be
significant, it should have from 3 to 10 links.

The purpose of this chapter was to demonstrate the use of lexical cohesion analysis for
the automatic generation of text summaries. It is important to evaluate the quality of
these summaries. For this purpose, we have conducted an experiment that consisted of
asking working scientists to read the original texts and their summaries and express their
opinions about the readability, content, and accuracy of these summaries. The results
that we obtained were quite encouraging in that most of the summaries were judged as
readable and that their contents in terms of subject matter were quite accurate. There
was, however, some mitigation about whether the lack of equations, tables, or graphs
would harm the summaries or not. The general opinion about this point was that the
absence of equations, or references to equations, would not harm the readability or
content of the summary. It would however cause information fragmentation as far the
logical presentation of arguments is concerned. This information fragmentation may, as
a consequence, lead to a poor summary readability.
We have introduced the notion of paragraph centrality in an attempt to solve the problem of summary readability. We have argued that this notion can be used to indicate the most relevant part of a text at a higher level than that of sentences. We have seen how, in a given text, there are topic opening and closing paragraphs which tend to introduce and conclude a text in, perhaps, the same way that topic opening and closing sentences do. The most central paragraphs of a text can be found in the beginning, middle, or even close to the end of the text. This reflects the emphasis of the author as he/she moves from one theme or argument in a text to another. Furthermore, the notion of central paragraphs can be extended to that of central sections, central chapters and even to higher levels. Summarising a text using central paragraphs, sections, or chapters can in fact be regarded as a method to produce indicative outlines of the text that are richer than the the familiar structural outlines (containing only section numbers) which are familiar to us. This means that, provided with the lexical cohesion bonding information between sentences of the text, one can, not only generate summaries, but also produce outlines of the corresponding text.

In the next and last chapter, we will conclude the results of our methodology and propose further related research work.
Chapter 6

Conclusions and Future Work

Our work on analysing and summarising text using the notion of cohesion, particularly lexical cohesion, we believe, has thrown some light over the potential of this approach. This approach is different from existing and conventional text analysis and processing systems in many ways. The most important aspect of the cohesion-based approach is perhaps its artificially intelligent orientation in that the analysis is closer to the meaning of the text than to its structure. We believe that a text summarisation system based on lexical cohesion links that exist between words, sentences, and paragraphs, and using the semantic knowledge of relationships between words from a thesaurus, can make such a summary more realistic in that lexical cohesion allows one to talk about topic opening, topic closing, central, key-central, and marginal sentences. Such judgements about the writer’s intent can be made just by looking at simple lexical repetition for instance.

There are a number of problems one can solve by using lexical cohesion analysis: summarisation we have discussed at length, other areas of language engineering, like terminology elicitation, and information retrieval, can also benefit.

6.1 What have we achieved?

We have attempted to investigate the linguistic notion of cohesion that text linguistics experts regard as an important aspect of a well-written text. By ‘well-written’, we mean a text that is readable in that the arguments, thoughts, and ideas of the author make a harmonious continuity: a weave (=text) of thoughts and ideas knitted together with arguments.
We have examined theories related to cohesion in text presented by Halliday and Hasan (1976) and Hoey (1991), and demonstrated that patterns of cohesion do indeed exist in text. Furthermore, we have argued that the notions of links and ties due to cohesion patterns between the sentences of a text indicate the existence of a continuous and homogenous flow of information. We have demonstrated that Hoey's claims concerning lexical cohesion are valid over much larger tracks of text than those he used for his study.

We agree with Hoey when he says that lexical cohesion patterns in text can be used as a metaphor for sentence categorisation; such a metaphor constitutes the backbone of our cohesion-based text summarisation methodology. As an extension to, and perhaps, a refinement of Hoey's theory, we have defined in a semi-formal way the categories of sentences that are labelled as topic opening, topic closing, marginal, and central sentences. Moreover, we have suggested the key-central sentence category which stems from a further restriction on the selection of central sentences.

The experiments that we have conducted over several texts shows that, not only patterns of lexical cohesion exist in text, but also that the number of links and bonds formed by these patterns follow a trend that indicates decay as the bond threshold is increased. In his first study, Hoey mentioned that a bond between two sentences is significant only when the sentences share at least three lexical cohesion links. We have demonstrated that this number, to which we referred as the bond threshold, should not only have a lower bound but, perhaps more importantly, it should also have an upper bound. The decay of the number of bonds in various texts indicates that it reduces to almost zero as the bond threshold exceeds the value of 10.
Furthermore, the analysis of the various texts reveals that the different types of lexical cohesion which establish links and bonds between sentences, paragraphs, and sections have distinct effects on the overall cohesion of the text. We have elaborated on this observation by comparing the number of bonds produced by the lexical cohesion types either individually or collectively. We have actually established a ‘pecking order’ of these types and noted that simple lexical repetition is the most important type and that simple mutual paraphrase is the least important. Moreover, simple mutual paraphrase appears to contribute little or nothing at all to the formation of lexical cohesion bonds in text as compared to simple and complex lexical repetitions. The ‘pecking order’ of lexical cohesion types, based on the analysis of the distribution of links, is the following one.

1. simple lexical repetition+complex repetition+simple mutual paraphrase
2. simple lexical repetition+complex repetition
3. simple lexical repetition+simple mutual paraphrase
4. simple lexical repetition only
5. complex repetition+simple mutual paraphrase
6. complex repetition only
7. simple mutual paraphrase only

Our methodology has allowed us to test both Halliday and Hasan’s (1976) and Hoey’s (1991) theories of cohesion and show that such theories can be used to build a framework that, in turn, can be used to develop text summarisation systems. We have conducted lexical cohesion text analysis experiments over a large number of lexical tokens with the help of the on-line thesaurus Wordnet which is widely recognised as a rich lexical resource.
In order to evaluate the quality of our summaries, we have conducted an experiment that consisted of asking four working scientists to read the summaries and the corresponding original texts and express their comments by answering a number of questions related to the readability, content, and quality of the summaries. The results of the evaluation were positive. The four scientists who participated in the experiment found that the summaries were, in average, readable, and that the information content of these summaries was reported quite accurately. With the exception of one summary which had a few references to mathematical equations, the scientists agreed that the summaries had a quite logical presentation of arguments and that, the presence of very few pronominal references caused little fragmentation in the summaries. The general comment which emerged was that the summaries were short, readable, and quite accurate, particularly in their contents of the main topics found in the original text.

6.2 Text summarisation: Is cohesion analysis a plausible method?

We have reviewed earlier in this thesis some of the cohesion-based text understanding systems such as the TOPIC system (Reimer & Hahn, 1990), the lexical chain analysis system (Moris & Hirst, 1991), and the cohesion index analysis system (Stoddard, 1993). It is not clear to us how themes in TOPIC interact, nor is it clear how Morris and Hirst’s ‘lexical chains’ establish the chaining of themes from that of lexical items. However, the attempts made in developing these systems to understand the themes or topics of a text indicate a developing consensus amongst computer scientists and text linguists that the analysis and exploitation of cohesion helps in summarising text.

Our methodology for automatically producing text summaries is entirely based on lexical cohesion. We have argued earlier that the concept of sentence categories, namely topic
opening, topic closing, central, key central and marginal sentences can be used for generating text summaries.

It is the simplicity of the notions that underlie lexical cohesion in text, and the oft-repeated observation that even for analysing small fragments of text, one requires masses of time if such an analysis is carried out manually, that make our methodology for analysing and summarising text distinct from others. Other methodologies rely on syntactic analysis and, in some instances, are based on semantic hypotheses for paraphrasing text that lead to an inevitable interference with the text itself. The notion of counting links and bonds in our methodology is clearly simple and has a minimal of interference with the text as it embodies simple counting operations to categorise sentences and consequently summarise text. The operationalisation of lexical cohesion analysis is, therefore, less problematic in that it relies on the simple notion of counting text tokens and variants of tokens. Dictionaries like Wordnet are essential for detecting these variants.

Furthermore, text summarisation strategies in our methodology consist of selecting individual central sentences, and even individual central paragraphs for an improved readability of the summary. This means that these strategies do not interfere with the text in that sentences of the text are not reduced or transformed in any way. These sentences are rather judged as central or relevant to the subject matter of the text, and then, extracted as a summary. The difference, perhaps, between our work and others in this context is that our methods produce summaries ‘from the text’ whereas others produce summaries ‘of the text’. According to Paice’s categorisation of automatic
summaries (1993), we should distinguish summaries 'from the text' as indicative and summaries 'of the text' as informative.

We have discussed the readability of the summaries produced by TELE-PATTAN. We insisted that single sentence extraction may lead to the production of discontinuous summaries that may have a poor readability. An alternative method that may be used to overcome this problem is to produce summaries using central paragraphs. We introduced this notion of paragraph centrality and argued that there are two important advantages here. First, summaries based on the extraction of central paragraphs will have less discontinuity, fewer pronominal references, and therefore better readability. Second, this notion can be scaled up to the levels of sections and chapters of the text. Central paragraphs, or central sections, or even central chapters of a given text may be regarded as an outline of such a text. Such an outline may be richer and more meaningful than the conventional text outlines which consist of section numbers and headers only.

The notion of sentence categories, which can be used to indicate the topics or themes of a text, makes our summaries act as 'a way into the text'. In other words, our summaries can be seen as outlines of the different topics expressed or reported in the text.

6.3 Future Work

The experiments that we have performed in the course of this work are based on non-narrative texts. We believe that our methods and techniques can be used to analyse all types of text including narrative and fiction texts, but perhaps not to the same degree of success as with non-narrative scientific texts. However, the analysis of these texts may
need the inclusion of syntactic cohesion in addition to lexical cohesion. The small narrative text that Hoey analysed appears to contain many instances of syntactic repetition like reference, anaphora, ellipsis, and substitution (Hoey, 1991). The analysis of both syntactic and lexical cohesion in narrative texts may need the use of elaborate and sophisticated methods to resolve instances of references and anaphora. In this section we discuss the limitations of our methodology and how it can be improved. We then raise further research questions concerning a possible use of lexical cohesion analysis in two areas which are information retrieval and terminology elicitation.

6.3.1 Limitations of and improvements to our methodology

There are three issues related to the improvements of our methodology that we would like to point out. The first one is on the analysis of syntactic cohesion and its role in the cohesion of scientific text as compared to its counterpart, lexical cohesion. The second one is related to the use and exploitation of the features of Wordnet as a valuable on-line thesaurus to improve the extraction of simple partial paraphrase where there is a need for semantic hierarchies. The third one is on automatic evaluation of cohesion based text summaries.

Accounting for syntactic cohesion

Our cohesion based text analysis and summarisation methodology which we tested on a collection of texts is based on lexical cohesion. Our study does not include syntactic cohesion types such as ellipsis, conjunction, and reference which is achieved through the use of pronouns. Although we believe that in scientific text, lexical cohesion appears to be prevalent, we are unable to make similar claims about syntactic cohesion. One can however use existing methods (Grosz & Sidner, 1986; Leass, 1991) for resolving
references in text and see how syntactic cohesion, particularly that due to reference, affects the cohesion of a text in comparison with lexical cohesion.

*Efficient exploitation of Wordnet’s knowledge base for cohesion analysis*

Wordnet’s source files which were written by lexicographers who, after a detailed of analysis of lexical semantics, have produced a variety of lexical and semantic relations that can be used to represent and organise Wordnet’s lexical knowledge. One of these semantic relations is that of ‘relational pointers’ which represent semantic links between word forms and synonym sets. Examples of these relations are those linking the noun and verb forms of words. Most importantly, one of the relational pointers is that of meronymy. Wordnet provides three types of meronymy pointers which are: ‘part of’, ‘substance of’, and ‘member of’. These types of relations can be used to identify many instances of simple partial paraphrase. We mentioned earlier that simple paraphrase can be mutual and partial, and that, although simple mutual paraphrase can be identified using synonymy, one needs knowledge of a semantic hierarchy to identify instances of simple mutual paraphrase. Perhaps, the only complication that one may encounter is that, because Wordnet is based on word-forms rather than concepts, the meronymy pointers may not indicate a proper semantic hierarchy. Furthermore, the lexical relations between words and their morphological variants can be used to solve the problem of linking homonyms like ‘bank’ and ‘bank’ where one word is a verb and the other is a noun. These relations can also be used in lexical cohesion analysis to avoid the identification of suspect links such as that of ‘make’ and ‘shape’. We have shown in our study that simple mutual paraphrase has almost no effect on the cohesion of a text. However, the inclusion of simple partial paraphrase may reveal something different.
Cohesion-based text summarisation - towards automatic evaluation

Currently, the evaluation of cohesion based summaries relies on the use of feedback assessment which consists of studying the readability of the summary in comparison with the original text by human subjects who are familiar or acquainted with the domain knowledge of the text. It is, perhaps, as complicated to evaluate text understanding, text summarisation, and text retrieval systems, as the development of such methods. The complexity of evaluation, particularly that of summarisation, is due to the complexity of language itself in that one has little means to measure the quality of a text or its author’s success in conveying the message it embodies. Our judgement on text quality can be seen to be often subjective in that every reader has a specific objective in reading a given text. And, if one’s background is not related to the subject matter of a text, such a text is judged as either irrelevant or inappropriate.

In order to avoid subjective judgements on the quality of a text summary, we believe that the process of evaluation should be tied to the original text. In the context of cohesion-based summarisation, one way to automatically evaluate a summary is to compare the number of terms they both have. This is based on the assumption that summaries and abstracts contain most of what is widely known as ‘keywords’ which, in the context of terminology elicitation, represent terms. Summaries that have most of the terms that are found in the original text can be considered, perhaps, ‘good’ or ‘acceptable’ summaries. Other summaries that have a poor content of terms tend in turn to be ‘weak summaries’. But again, the size of the summary is a determiner of the number of terms. Questions abouts automatically evaluating summaries remain unsolved and methods used nowadays are still evolving.
6.3.2 Cohesion-based text analysis: Towards semi-intelligent information retrieval systems

Current information retrieval systems rely on the reductive technique of using keywords in contexts for cataloguing texts and subsequently using the keywords to retrieve texts. Consequently, a number of broad terms may be used as keywords and a diverse range of texts can be retrieved using these keywords. For instance, the keyword nuclear energy can be used to catalogue pro- and anti-nuclear power texts: the use of keywords of the kind helps to retrieve a whole range of documents the bulk of which might be of interest to lexicographers only! Abstracts of documents are used as surrogates for information retrieval purposes. Again abstracts, unless well-written, contain dense information which sometimes may be opaque to many texts in the same specialism. The information retrieval literature does address questions related to 'beyond keywords...', and most of the answers relate to refining and restricting the use of certain keywords. In our view, lexical cohesion-based texts extracts (i.e. summaries comprising topic-opening, central, and topic closing sentences) that are retrieved as result of a given query will be more informative and, perhaps, would free information retrieval systems from the exclusive reliability on the 'keywords in context' (KWIC) method.

Using cohesion analysis in conjunction with the conventional term frequency methods in information retrieval, we believe that the retrieved documents will have, besides similarity, a degree of overlap with the query. This is a claim that may find some support from workers in the information retrieval milieu particularly those working on the application of artificial intelligence techniques in the development of document retrieval engines.
6.3.3 Terminology elicitation: A lexical cohesion analysis method

The extraction of text summaries comprising, for instance, topic opening, central, and topic closing sentences, can be used to identify the key terms of the specialist domain described in such a text. The text summary can be used as a precursor to terminology extraction from texts and can possibly be used in the selection of texts that are rich with terms and the rejection of those that have a poor content of terms. Systems that deal with the extraction, rejection, and organisation of terms are known as ‘terminology management systems’.

In a term base, the definitions associated with terms are excerpts from the text in the corpus, i.e. they are groups of selected sentences. The selection of these sentences to construct term definitions is empirical and pragmatic at the same time. We believe that either central or key central sentences can be used for this purpose whereby a definition of a single or a compound term in a text may comprise a single key central sentence, or a collection of central sentences.

Terminology management systems, like the University of Surrey’s System Quirk, provide access to text corpora that may contain terms, their descriptions and elaboration. Such systems lack (lexical-) cohesion analysis tools and thus programs like TELE-PATTAN will be an interesting addition to terminology management systems.
Central and key central sentences can be used to define terms assuming they are already identified. One way to extract and identify these terms is to use the lexical chain extracting tool which is part of TELE-PATTAN. These chains which were first used for discourse segmentation in Morris’ and Hirst’s work (1991) represents lists of potential domain specific terms.

Furthermore, in a terminology management system where corpus is regarded as one of the main resources which needs organisation and maintenance, cohesion analysis can be used to locate the appropriate corpus to store a text of a given domain. Similarly, it can be used to extract the appropriate text from a number of candidate texts in a corpus. This operation, which can be regarded as a selection constraint, consists of evaluating the degree of cohesion between the topic of interest and the candidate texts in the corpus.
In summary, the use of cohesion analysis to extract summaries, to identify domain specific terms, or to retrieve documents may be consolidating for establishing a framework in which authors of different texts are compared according to the distribution of cohesion patterns in their texts. Such comparisons may be used to add, say, a 'cohesion attribute', to the authors' profiles.
Appendix A: Wordnet — An online electronic thesaurus

Wordnet is an on-line lexical reference system which is a database of nouns, verbs, adjectives, adverbs, and function words organised into synonym sets representing each a different lexical concept and linked by different relations (Miller et al. 1990). Wordnet is primarily based on psycholinguistic theories of the human lexical memory.

The approach to the design of Wordnet is based on the observation that Murray’s Oxford English Dictionary, though it has a clear notion of word use and sense priority, its compilation based on historical (diachronic) evidence did neglect issues and problems concerning the synchronic organisation of knowledge. In an attempt to resolve such problems, the study of the factors influencing the contemporary (synchronic) structure of knowledge, particularly lexical knowledge became an important issue in lexicology and gave rise to what is now known as psycholinguistics (Miller, 1976). The emphasis in psycholinguistics is to determine the necessary information that needs to be included in a lexicon to account for the phonetics, syntax, and semantics of English. Wordnet is the result of such an investigation explored by a group of linguists and psycholinguists at the University of Princeton (Miller et al. 1990).

The difference between Wordnet and other dictionaries is that it is based on word meanings rather than word forms. That is to say, the principal entries are not alphabetical patterns (i.e. words) but rather synonym sets. This property makes Wordnet a thesaurus.

The organisational structure of Wordnet is illustrated using the lexical matrix (Miller et al., 1990) in Table A-1. The list \([F_1, F_2, F_3, \ldots]\) represents word forms, i.e. the lexical entries of Wordnet. The list \([M_1, M_2, M_3, \ldots]\) represents the set of meanings that Wordnet covers. An entry \(E_{ij}\) in a given row and column position is an indication that the word form \(F_j\) can have the meaning \(M_i\). If there is more than one entry in one column, the corresponding word form has more than one meaning and is therefore polysemic. Similarly, all entries in the same row indicates the synonym list containing different word forms.

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Word Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F_1)</td>
</tr>
<tr>
<td></td>
<td>(F_2)</td>
</tr>
<tr>
<td></td>
<td>(F_3)</td>
</tr>
<tr>
<td></td>
<td>(\ldots)</td>
</tr>
<tr>
<td></td>
<td>(F_n)</td>
</tr>
<tr>
<td>(M_1)</td>
<td>(E_{1,1})</td>
</tr>
<tr>
<td>(M_2)</td>
<td>(E_{1,2})</td>
</tr>
<tr>
<td>(M_3)</td>
<td>(E_{2,2})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(E_{3,3})</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(E_{m,n})</td>
</tr>
<tr>
<td>(M_m)</td>
<td>(E_{m,1})</td>
</tr>
</tbody>
</table>

Table A-1: Wordnet’s lexical matrix.

In the table above, the words \(F_1\) and \(F_2\) are synonyms. The word \(F_2\) has more than one meaning and therefore belongs to two synonym sets \(M_1\) and \(M_2\). The order of word forms is irrelevant here as it is only important to know the synonyms sets which can be accessed from any word form in Wordnet. This explains the ‘many to many’ relationships between word forms and meanings.

Although there is no clear psychological evidence on which the grouping of synonyms is based, an interim solution used in Wordnet is to consider a lexicographic assumption that a synonym set
represents a lexical concept rather than an abstract one (Miller et al. 1990). This is one of the reasons that Wordnet has a clear organisation that makes it readily exploitable for any lexical analysis of text.

The meaning (or synonym) sets of Wordnet are linked with different semantic relations such as synonymy, antonymy, hyponymy, meronymy. Morphological relations between word forms which were added to Wordnet at a later stage add a powerful capability to deal with inflectional morphology that is often a major problem in text analysis.

Although Wordnet is a highly organised database meant for general use, there was a need to configure such an organisation for our lexical cohesion analysis methodology. We did use its semantic relations and contented ourselves to develop our own morphological analyser.

In the early stages of our research, we used the Macquarie thesaurus of Australian English that we were lucky to acquire in a machine readable form. This thesaurus edited by Bernard (1990) and also known as 'the book of words' contains over 180,000 entries divided into over 800 'semantic' categories which range from 'abstinence/overindulgence' to 'animal noises', to 'sport' (mainly Australian sport) to trade unions, to water craft and work, etc. For each semantic category there are entries under nouns, adjectives, verbs and adverbs. Each of the grammatical categories could be divided into up to 7 or 8 subcategories. The importance of the Macquarie Thesaurus which was to motivate us to use it for lexical cohesion analysis is that it comprises relatively modern English as it is compiled in the 1990's. However, its restricted and non clear semantic categorisation showed a major problem in identifying lexical relations between word forms.

Another important thesaurus that we should not ignore and that one could think of using in lexical analysis is the Roget's thesaurus. This thesaurus was first compiled in the 1849 by Peter Mark Roget. In Roget's thesaurus, English words and phrases are classified and arranged so as to facilitate the 'expressions of ideas and assisting literary composition'. This thesaurus has been revised a number of times over the years to incorporate changes in the English spelling system. Roget's thesaurus has words classified under a number of 'conceptual' categories that are; a) abstract relations, b) space, c) matter, d) intellect. e) volition, and f) affections. One can cross-reference related words not merely in terms of orthography but also in terms of meanings. This thesaurus is also available in machine readable form. The Roget's Thesaurus classifies words into four grammatical categories; nouns, verbs, adjectives and adverbs. This is very useful for analysing scientific writing because one way of expressing scientific ideas is to discuss a phenomena in terms of activities or actions. These activities and actions are used as new information and represented in a common form into subsequent sentences. Notwithstanding the fact that the Roget's Thesaurus was made available in machine readable form only recently, it is not very easy to decipher the various categories in which this thesaurus is organised only for the reason that this comes as fully flat file structure.

In assessing Wordnet, Macquarie's, and Roget's thesauri, and, knowing that lexical cohesion analysis requires thorough and efficient search of lexical relations between word forms and morphological variants of word forms, we see the emergence of Wordnet as the suitable resource to use. This was mentioned earlier in chapter 4 and is reported here in the table A-1 for convenience.
### Table A-2: A comparison of thersaural resources.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Wordnet</th>
<th>Macquarie’s</th>
<th>Roget’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Language Quality</td>
<td>Modern</td>
<td>Modern</td>
<td>Modern</td>
</tr>
<tr>
<td>Contents size</td>
<td>164,000 entries</td>
<td>180,000 keywords</td>
<td>990 entries</td>
</tr>
<tr>
<td>Semantic Categorisation Type</td>
<td>Conceptual (all the synonym sets))</td>
<td>Ad-hoc 812</td>
<td>Conceptual 5x24</td>
</tr>
<tr>
<td>(Categories/Concepts)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Relations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonymy</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Antonymy</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Hyponymy</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Meronymy</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Morphological Relations</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Syntactic categories</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nouns</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Verbs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Adjectives</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Adverbs</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Cross references</td>
<td>rich</td>
<td>limited</td>
<td>rich</td>
</tr>
<tr>
<td>Machine readable forms</td>
<td>Fully organised database</td>
<td>Flat text files</td>
<td>Flat text files</td>
</tr>
<tr>
<td>Use for lexical cohesion analysis</td>
<td>Suitable</td>
<td>Limited</td>
<td>Limited</td>
</tr>
</tbody>
</table>

One notices a clear advantage in using Wordnet for a text analysis that requires both syntactic and semantic knowledge about words. We have indeed used Wordnet in the latest stages of our research and the discussion that will follow in sections to come are based on this lexical knowledge resource.
MATERIAL REDACTED AT REQUEST OF UNIVERSITY
AHMAD, KHURSHID (1993)
Terminology and Knowledge Acquisition: A Text Based Approach. In KLAUS, DIRK SCHMITZ (Ed.).
TKE '93: Terminology and Knowledge Engineering. Proceeding of the Third International

ALTERMAN, RICHARD (1992)

BARNET, J., MANI, I., KNIGHT, K., & RICH, E. (1990)
Knowledge and Natural Language Understanding. Communication of the ACM. 33(8). pp. 50-71.

BARR, AVRON & FEIGENBAUM, EDWARD (1981)

BARTSCHI, MARTIN (1985)

BENBRAHIM, MOHAMED (1996)
Department of Computing and Mathematical Sciences. University of Surrey. Guildford, UK.

BENBRAHIM, MOHAMED & AHMAD, KHURSHID (1994)
Department of Computing and Mathematical Sciences. University of Surrey. Guildford, UK.

BERNARD, JOHN R. L. (Ed.) (1990)

SERAPHIN: an automatic system for main sentences extraction. Proceedings of the International
Conference on Recent Advances in Natural Language Processing. pp. 329-336. Tzigov Chark,
Bulgaria.

BLAIR, D. C. (1992)
Information Retrieval and The Philosophy of Language. The Computer Journal. 35(3). pp. 268-
278.

BONZI, SUSAN (1990)
Syntactic Patterns in Scientific Sublanguages: A Study of Four Disciplines. The Journal of the

BOOSE, JONES H. (1992)
Knowledge Acquisition. In STUART C. SHAPIRO (Ed.). The encyclopaedia of Artificial Intelligence.
pp. 719-742.

BOWERS, DAVID S. (1993)
From Data to Databases, 2nd Ed. London: Chapman and Hill.

An Algebra for Structured Text Search and a Framework for its Implementation. The Computer
Journal. 38(1). pp. 43-56.
COOPER, W. S. (1971)

CULLINGFORD, RICHARD (1986).
SAM. In (Eds.). BARBARA J. GROSZ, KAREN SPARCK JONES, & BONNIE LYNN WEBBER. Readings in Natural Language processing. California: Morgan Kaufman Publishers Inc.

DATE, C. J. A. (1986)

deBEAUGRANDE, ROBERT A. (1991)

deBEAUGRANDE, ROBERT A. & DRESSLER, WOLFGANG (1981)
Introduction to Text Linguistics. Singapore: Longman.


DEJONG, G. (1979),

DUTCH, ROBERT A. (1972)
Roget's Thesaurus of Synonyms and Antonyms. Framingham, Mass: Donnison Manufacturing Co.


FOWLER, ROGER (1986)

FREEDLE, ROY O. & CAROLL, JOHN B. (Eds.)(1972)

FRIES, CHARLES CARPENTER (1952)
The structure of English. London: Longman.

GAZDAR, GERALD & MELLISH, CHRIS (1989)
Natural Language Processing in Prolog. Wokingham: Addison & Wesley.

GONNET, GASTON H. & TOMPA, FRANK WIN (1987)


GROSZ, B. J. & SIDNER, C. L. (1986)
Hahn, Udo (1990)

Hahn, Udo & Reimer, Ulrich (1985)

Halliday, Michael A. K. (1961)


Cohesion in English. London: Longman.

Writing Science: Literacy and Discursive Power. London: Falmer Press.

Harris, Roy (1988)

Harris, Zellig S. (1982)

Harris, Zellig S. (1991)

Hayes, Philip J. (1992)

Hearst, Marti (1993)


Hoey, Michael (1991)

Holmes-Higgin, Paul & Ahmad, Khursid (1992)

Jackson, Daphne F. & Barret, Roger C. (1977)

JACOBS, PAUL S. (Ed.) (1993)

JANAS, JURGEN M. & SCHWIND, CAMILLA B. (1979)

JOHNSON, B. (Ed.) (1994)

JONES, S. (1991)

KATAMBA, FRANCIS (1993)

KITTREDGE, RICHARD & LEHRBERGER, JOHN (Eds.) (1982)

KNUTH, DONALD E. (1979)

KROEBER, KARL (1967)

LEASS, HERBERT (1991)

LEHNERT, WENDY G. (1986).


LEWIS, DAVID D. (1992)

MEYERS, GREG & HARTLEY, TONY (1990)
Modelling Lexical Cohesion and Focus in Written Text: Popular Science Articles and the Naive Reader. In ULRICH SCHMITZ, RÜDIGER SCHÜTZ, & ANDREAS KUNZ (Eds.). Linguistic Approaches to Artificial Intelligence. Frankfurt am Main: Verlag Peter Lang.

MILLER, G. A. (1985)
MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D. & MILLER, K. J. (1990)

MINSKY, MARVIN (1975)
A framework for representing knowledge. In WINSTON, P. H. (Ed.). The psychology of computer

MORRIS, JANE & HIRST, GRAEME (1991)
Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text.

NOBLE, H. M. (1988)

O’CONNOR, J. (1975)
Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching. Information
Processing and Management. 11(5/7). pp. 155-164.

O’CONNOR, J. (1980)
Answer Passage Retrieval by Text Searching. Journal of the American Society for Information
Science. 32(4). pp. 227-239

Constructing Literature Abstracts by Computer: Techniques and Prospects. Information

QUIRK, R., GREENBAUM, S., LEECH, G., & SVARTVIK, J. (1985)
A Comprehensive Grammar of the English Language. Harlow (Essex): Longman Group Ltd.

REIMER, ULRICH & HAHN, UDO (1990)
An Overview of the Text Understanding System TOPIC. In ULRICH SCHMITZ, RUDIGER SCHÜTZ, &
ANDREAS KUNZ (Eds.). Linguistic Approaches to Artificial Intelligence. Frankfurt am Main:
Verlag Peter Lang.

RICH, ELAINE & KNIGHT, KEVIN (1991)

ROTHKIGEL, ANNELY (1993)
Text Knowledge and Object Knowledge. London and New York: Pinter.

RUQIAN, LU & CUNGEN, CAO (1990)
Towards Knowledge Acquisition from Domain Books. In BOB WIELINGA, JOHN BOOSE, BRAIN
GAINES, GUUS SCHREIBER, & MARTEN VAN SOMEREN (Eds.). pp. 289-301.

SALTON, GERARD (1989)
Automatic Text Processing - The Transformation, Analysis, and Retrieval of Information by
Computer. Reading (UK): Addison-Wesley.

SALTON, GERARD & ALLAN, JAMES (1994a)
Automatic Text Decomposition and Structuring. In Proceedings of the RIAO’94 Conference. pp. 6-
20. New York: Rockefeller University.

264(3). pp.1421-1426.

SALTON, GERARD & MCGILL, M. J. (1983)
SCHANK, ROGER (1980)

SCHANK, ROGER & ABELSON, ROBERT (1977)

SCHMIDT, GABRIELE & SCHMALHOFER, FRANZ (1990)
Case-Oriented Knowledge Acquisition from Texts. In BOB WIELINGA, JOHN BOOSE, BRAIN GAINES, GUUS SCHREIBER, & MARTEN VAN SOMEREN (Eds.), Current Trends in Knowledge Acquisition. pp. 302-312.

SCHWARTZ, CHRISTOPH (1990)

SHAPIRO, STUART C. (Ed.)(1991)

SIMMONS, ROBERT F. (1972)

SINCLAIR, JOHN MCH. & COULTHARD, R. M. (1975)

SKOROCHOD'KO, E. (1972)

SMEATON, ALAN F. (1992)

SOWA, JOHN F. (1984)
Conceptual Structures-Information processing in mind and machine. Reading (Mass): Addison-Wesley.


SPARCK JONES, KAREN (1992)

STODDARD, SALLY (1993)

TESKEY, F. N. (1982)

Van Reisbergen, C. J. (1979)

Wales, Katie (1989)


Wilensky, Robert (1986).

Winograd, Terry (1983)
Language as a cognitive process: Syntax. Reading (MA): Addison-Wesley.

Willet, Peter & Rasmussen, Edie M. (1990)

Winter, Eugene (1982)

Zhengxin, Chen (1993)