1998

The Generation of Compound Nominals to Represent the Essence of Text The COMMIX System

Norris, Jennifer Vivien

The Generation of Compound Nominals to Represent the Essence of Text

The COMMIX System

Jennifer Vivien Norris

A thesis submitted in partial fulfilment of the requirements of the University of Brighton for the Degree of Doctor of Philosophy

August 1998

The Information Technology Research Institute,

University of Brighton

Abstract

This thesis concerns the COMMIX system, which automatically extracts information on what a text is about, and generates that information in the highly compacted form of compound nominal expressions. The expressions generated are complex and may include novel terms which do not appear themselves in the input text.

From the practical point of view, the work is driven by the need for better representations of content: for representations which are shorter and more concise than would appear in an abstract, yet more informative and representative of the actual *aboutness* than commonly occurs in indexing expressions and key terms. This additional layer of representation is referred to in this work as pertaining to the *essence* of a particular text.

From a theoretical standpoint, the thesis shows how the compound nominal as a construct can be successfully employed in these highly informative representations. It involves an exploration of the claim that there is sufficient semantic information contained within the standard dictionary glosses for individual words to enable the construction of useful and highly representative novel compound nominal expressions, without recourse to standard syntactic and statistical methods. It shows how a shallow semantic approach to content identification which is based on lexical overlap can produce some very encouraging results.

The methodology employed, and described herein, is domain-independent, and does not require the specification of templates with which the input text must comply. In these two respects, the methodology developed in this work avoids two of the most common problems associated with information extraction.

As regards the evaluation of this type of work, the thesis introduces and utilises the notion of *percentage attainment value,* which is used in conjunction with subjects' opinions about the degree to which the *aboutness* terms succeed in indicating the subject matter of the texts for which they were generated.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my thanks in particular to Colin Beardon and Donia Scott. Colin was my Director of Studies for the period over which this work was largely developed. Donia, who has been one of my supervisors all along, took over the role of Director of Studies when Colin left Brighton for pastures new. They have both provided lots of help and encouragement over the course of this work, and have also given me a lot of patient support in general.

I would also like to express my thanks to a number of other people who have helped by giving comments or ideas, or just plain, much needed support:
to Ken Turner, for reading and commenting on the earlier work; to Lucia Rino, for encouragement and many useful discussions; to Heather Downie, for useful discussions and help with arranging meetings with abstractors; to colleagues and friends at the Information Technology Research Institute, some of whom agreed willingly to be consulted as subjects in the evaluation stage of this work; to the other people who agreed to act as subjects in the evaluation; and to Angie, who has always been there with a happy greeting in the morning, and who has rescued my plant from near death on numerous occasions.

My final thanks goes to members of the Norris and Brown families, who have given tremendous encouragement when it has been most needed, and who have put up with lots of moans and groans without (too much) complaint.

The Tale of Fat Harry and his Handbag

Once upon a time there was a famous acronymaniac called Fat Harry, who was never to be seen without his great fat bulging handbag in which he carried, it was rumoured, vast amounts of useful information. Well, some said it was useful, but then no-one really knew whether it was useful or useless, because Fat Harry never let anyone look inside. In fact, he was so jealous of his handbag of information that he protected it with an enormous padlock and used to wake at night in a cold sweat for fear of losing the key.

Fat Harry was proud of his collection: so proud, in fact, that he let it be widely known that the handbag contained two compartments. One was full of information that was relevant to his interests and the other was packed with gems that he thought he might find relevant one day. It seemed clear to all who heard of him that Fat Harry was a very 'interested' man. But he was not a happy man.

The trouble was that he awoke each day with a new interest on his mind, so that each day began with a grand re-sorting of all the bits of information between the two compartments of his handbag. Everything had to be sorted by hand, of course, and read in full before he could decide whether or not it was relevant to that particular day's interest. This meant that almost all his waking hours were spent with his head buried inside his handbag. Needless to say, Fat Harry never got much exercise (except for his arms and eyeballs, of course), and he saw precious little of the sun.

This is the story of a well-wisher's attempts to help Fat Harry with his handbag problem and to get him out and about in the sunshine.

Chapter One

1. Introduction

This chapter has four main sections: the first sets the scene, and places the thesis in the general areas of Information Retrieval and Indexing. Section 1.2 introduces the related linguistic topics, and Section 1.3 discusses the motivation behind the work from both theoretical and practical perspectives. Finally, Section 1.4 presents an overview of the rest of the thesis.

1.1 Setting the Scene

Try to imagine a library of today, coping with the amounts of information to which they are expected to have access, but without any computer support at all. At best we picture row upon row of filing cabinets, and rack upon rack of shelves covered with innumerable boxes brimming with little brown filing cards. At worst, the chaos does not bear thinking about.

1.1.1 Access to Information

We have come to expect rapid access to certain types of information: in a public library, for example, we now expect computerised cataloguing to provide, at the very least, immediate information concerning the location and loan status of particular items.

The advent of automatic Indexing and Information Retrieval systems offers the user access to the information they require, rather than just letting them know where to find it. The advantages of such systems are well established: they offer the potential of rapid access to enormous amounts of relevant information whilst taking up little physical space. There is a wide range of information now available electronically, and databases commonly give access not only to bibliographic information and specialist areas of interest, but also to abstracts of papers or articles which appear in the more commonly read journals or newspapers. Often complete texts can also be accessed on-line. The advent of the worldwide web, which renders potentially accessible a vast amount of information, has raised the expectations of information seekers sky-high.

It is the problems associated with accessing the relevant material effectively and efficiently that provides the practical incentive for this work. The next section discusses some of these problems, from an information-seeker's point of view. Although the thesis is not directly concerned with this 'user' stage of the Information Retrieval process, a consideration of the user's viewpoint will help to place the ensuing discussion in the overall field, as well as to give a clearer idea of the incentive behind the specific direction of the work.

1.1.2 Getting at the Information you want

So how does an information-seeker get at the specific information they need?

A user with a particular query wants to be able to access relevant information quickly and straightforwardly. Ideally they would be able to request information of a retrieval system using natural language, but this aim is precluded by the current state of Natural Language Processing (NLP). A user therefore needs to express their query to the system in a way that can be utilised by the system to distinguish between material which is relevant to the query and that which is not.

There are two clear pitfalls to be avoided in the search for relevant information. On the one hand, the aim is to ensure that the user is informed about all the information that may be appropriate, and that relevant material is not missed. Consider, for example, a user who wants to know which company spends the most money on advertising their video games products. It is no good if the system fails to recognise as relevant a text which discusses the marketing of video games, simply because the word 'marketing' appears in the text whereas 'advertising' has been used in the query. There needs to be a way of avoiding such 'under-recall' of information, where relevant information is missed by the system being used.

On the other hand, we wish to ensure that such a system does not 'over-recall', with the result that the user is presented with a large amount of information which is irrelevant to the query. In the example above, it is clearly a failing if the user is presented with all material dealing with anything related to video games, or all that which discusses video games and money, or even a text which discusses video games and the advertising of something else. We might imagine, for example, a text which discusses the times when video games are played, reporting that one such

time is during the advertising slots, or commercial breaks, which interrupt television programmes. Sorting through related information which turns out to be irrelevant to the query is extremely time consuming and leaves the user frustrated, tired and often no better informed.

We are thus faced with two problems to minimise: firstly, the **failure to access** relevant information (the 'fatari' problem - which we refer to here colloquially as the (sic) Fat Harry problem); and secondly, the **access of irrelevant information** (or the 'accessoiri' problem - colloquially referred to as the (sic) Handbag problem). Readers with a particular aversion to the misguided use of acronyms may prefer the terms 'under-recall' and 'over-recall' problems respectively.

It is important to note at this point that there are two measures by which the performance of Information Retrieval (IR) systems are traditionally judged: namely *precision* and *recall.* These measures are based on the two problems described above. The precision of an IR system is a measure of the proportion of retrieved documents which are relevant to the query. The recall represents the proportion of relevant documents occurring in the database which are actually retrieved. Thus, a system with a high precision measure is one in which a large number of the documents which are retrieved by that system are relevant to the query, and that the amount of irrelevant material retrieved is low. Conversely, a system with a high recall rating is one which accesses a high number of the documents which are relevant, and minimises the number of relevant documents that are missed.

There are three aspects involved in getting the best values for both precision and recall in a retrieval system. The first lies in the terms selected by the user to express the query to the system. The second lies in the way in which query terms are processed by the system and matched to relevant items held in the database. And the third, which constitutes the main area of interest in this thesis, is ensuring that documents have associated with them some information which usefully describes what their contents are about.

1.1.2.1 Asking a System for Information

A user of an IR system has in mind some central topic about which they wish to obtain more information, and for which they need to formulate a query term. It may be a topic for which a lexical item already exists, or it may be a topic which can only be expressed as a phrase. For example, the topic may be adequately expressed by the word 'trees', or it may require a phrase, such as 'trees which bear fruit', or even 'trees which have cones and evergreen leaves' if the user is not familiar with the lexical item 'conifer'. There is a sense in which it is useful to distinguish between types of topics, depending on whether they may be expressed by a lexical item or whether they require a phrase. The linguistic notion of **head**[1] is of use here. This is a grammatical term used to refer to the central component of a phrase, which can occur in the same environments as the whole phrase. The head of each of the above queries is thus 'trees', whereas the topics differ according to the **modification** of the head ('which bear fruit', 'which always have leaves'). Where a topic is fully expressed by a head (i.e. where no modification is required), we may be justified in describing it as a simple topic. In contrast, where some modification of the head is required to fully express the topic, the term 'complex topic' seems justified. This is an oversimplification of the notion of complexity in relation to type of topic, but is a useful distinction to make. It allows us to distinguish between general topics, for which a lexical item will suffice as a description, and more specific ones which require more elaborate representation.

For simplicity, then, the initial distinction is made between two types of queries, which correspond to the two types of topic described above. The term 'simple query' is used in reference to topics which may be expressed as lexical items, and which could thus be expected to occur as independent items in a standard full (i.e. not concise) dictionary. The term 'complex query' refers to a query relating to a topic whose expression requires some modification of a head.

1.1.2.1.1 Simple Queries

A user who wishes to obtain information from a database sometimes has a general query and wants to see the range of related material available. This type of query is

---

[1] Items which appear in bold type are discussed in later chapters. Those occurring in this section are discussed in Chapter Three.

fairly straightforward to formulate and to process. Consider someone who wants to access any information concerning video games[2]: they are not interested in any one specific aspect, and the precise term used for a general query (such as 'video games') is likely to occur itself as an exactly matching term in related items in the database.

Current IR systems and methodologies cater fairly well for users with general queries, and searches based on single lexical items are straightforward. However, the main problem encountered by a user with a general query results from the information overload usually associated with the use of a general query. The solution lies in there being provision for dealing with queries which are far more specific, and it is this provision that is of interest here.

1.1.2.1.2 Complex Queries

A person who has a more specific topic about which they require information, would benefit from being able to present a retrieval system with a more complex query to express that topic. There are two aspects involved here: firstly, the user must have a way of adequately expressing the complex query in a form which is usable by the system. Secondly, the system must be equipped to utilise the query so as to access the relevant information reliably, and so maximise its precision and recall. It is the latter of these two strands related to complex topics, i.e., providing a facility which will equip systems to utilise such complex queries, that forms the practical incentive behind this work.

1.1.2.2 Expression of Complex Queries

The issue here is that a person with a complex query has a lot of detail to convey in their query term, and they must formulate an expression to convey it in as concise a way as possible. For example, the user of a retrieval system may want to know how much money is spent annually on the industry associated with video games. If they were to ask a direct natural language question of the system, (i.e., "How much money is spent annually on the industry associated with video games?") this would require at least some degree of parsing, and would still not provide the system with any usable concise term to use as the basis for a search of its database.

---

[2] Note that although there are two lexemes involved here, the term 'video game' functions as a word in the same way as 'washing machine'. Such items have become lexicalised and appear as independent items in a standard dictionary.

The standard methods used in IR systems require the user to search a database via a mixture of key words and phrases, authors and titles, in which the queries must match wholly or partially to the titles or authors. Some systems include abstracts of the documents listed, but the procedures for matching queries to text are the same. Those users with highly specific queries are required to formulate possibly relevant subsets using a variety of 'sub-queries', and then manipulate these sets using Boolean operators. Thus, in the above example, the user would need to search and obtain sets of database documents related to each of the following: 'video games' plus 'electronic games', intersected with 'industry', with the resulting set then intersected with 'advertising' and 'marketing' respectively. The search may cover other possible terms which the user must guess the author to have used, such as 'spending', 'money', 'pounds', 'dollars' and so on. They would then need to hope that all likely synonyms have been included in the search terms used, and that authors would have avoided terms like 'measures to increase sales' in place of 'marketing'. This method is time-consuming, fairly haphazard, expensive and frustrating, since much of the information accessed turns out not to be relevant to what the user actually wants.

1.1.2.3 A More Useful Approach: Complex Queries as Compound Nominals

Rather than trying to concoct a Boolean search term replete with logical operators, it is useful for someone to conceptualise this query as a 'thing' or topic about which more information is required. This is not as complicated a task (for a human) as it might at first seem, and a user of a retrieval system could formulate a term of reference for their query by having in mind the wording of the directive: "Tell me what you know about 'query' ", which in the above example might yield the query term 'video games industry spending'. This is the kind of language we see in telegraphic speech and is typical of newspaper headlines. It allows a high degree of conciseness and facilitates the succinct expression of a complex topic.

This type of nominal expression has received much attention over the past twenty years or so, and has attracted various terminology, including 'complex nominals', 'compound nouns' and 'nominal compounds'. It is not clear, however, that these different terms are used consistently, partly due to a lack of clear definitions of the linguistic phenomena under study. For this reason the term used in this work is **compound nominal** and the phenomenon is addressed in detail in Chapter Three.

Compound nominals, then, offer a linguistic mechanism for expressing complex topics in a way which is concise, and requires their expression in the form of nouns. They are not difficult for people to construct, and a person with a complex topic in mind may be advised to express it in the form of the general "Tell me what you know about 'topic' " instruction.

1.1.3 Requesting Information: a Specific Database

This section introduces the database which is used as the basis for the practical aspect of the work described in the thesis. After a brief introduction to the database itself, there is a description of the current means of passing queries to it, in order to access information. The final part of this section then suggests how the methodology might be improved, and places particular emphasis on providing a more representative means of describing what the texts themselves are about, which would then provide a foundation from which to utilise the kinds of complex query terms advocated above.

1.1.3.1 The Database Used in this Work

In this work we are dealing with a database which is a commercially available academic index[3]. It comprises a listing of a number of documents, along with abstracts relating to those documents, and their titles and authors. The coverage is the hundred most frequently read periodicals (including daily and weekly newspapers) as judged by undergraduate usage. Within each issue, coverage is determined solely by length, so that all articles above a minimum word number are recorded. This means that the area of coverage is extremely broad, in terms of subject matter, domain, style and genre of the original texts.

An abstract for each article has been written by a professional abstractor, and (at the time of writing) these are recorded onto CD-ROM, arranged by publication and date. There is no categorisation according to subject of coverage. The abstractors are required to select, from a predefined list, appropriate 'key terms' which are listed at the beginning of each abstract[4].

---

[3]The General Academic Index, produced by the Information Access Company.

[4] Further details of the abstracting process, along with specific abstracting constraints, appear in Chapter Two.

1.1.3.1.1 Current Methods of Accessing the Information

Someone requiring information from this database provides a query (a word or phrase) and may search according to title or author. They may restrict the search to key terms only (which must exactly match those selected and assigned by the abstractor), or they may opt to search an entire disc for a direct match of the particular search term used.

If the user's query is a general one, then they could expect a large number of relevant items to be matched, as mentioned above. If, however, the user has a specific query (such as 'video games industry spending'), there is no mechanism by which the current access procedures can utilise the query to the full, in the absence of a direct match of the whole term. The user will be notified that there is no direct match and partial matching will be offered. This will allow searching of all the constituent items of the original search term (e.g., 'video', 'video games', 'industry' etc.) but does not reflect the essence of the query as being about a particular type of spending.

1.1.3.1.2 A Better Method of Accessing Complex Information

This database provides a typical example of the lack of facility for dealing with complex queries, and provides a good basis for the discussion and implementation of a mechanism for improving the representation of what each text is about, which is the direct subject of this thesis.

What would ideally be required, then, is a user with a complex query expressed in the form of a compound nominal, and a system able to deal with queries of this type. We would not hope for such a term to always match directly with extracts from a text, even if the text could be appropriately described as being about the topic expressed in such a query term. Highly compacted compound nominal terms tend not to appear in normal text. (Even if the author of the original text had used such a term, restrictions placed on the abstractors by the parent company producing this database would not allow more than three consecutive words to be directly quoted, for fear of accusations of plagiarism.) Although it is reasonable to assume that an abstract will contain terms which are more concise than those appearing in the extended text, we rarely find such a degree of conciseness as would be typical of a highly specific compound nominal query term.

The complex topic itself (about which a user has a query), then, will not necessarily be expressed in relevant texts (or their abstracts) in terms of that specific topic, in the form of a compound nominal, but rather as information about some simpler topic. Thus, for example, an article about video games may refer to various pieces of information about video games. It may tell us that the associated industry is growing, or that it is taking over the toy market, or that specific amounts of money are spent annually on it by specific companies. But it is unlikely to utilise highly specific terms of reference for the information it covers. Thus, the text is unlikely to contain such terms as *'video games industry growth'*, *'toy market domination'* (or indeed the highly specific *'video games industry toy market domination'*) or *'video games industry spending'*, even though each of these topics may be judged as being what the text is at least partly about.

It is the specification of 'what a text is about', or its **aboutness** (discussed in Chapter Two) in concise terms that has the potential to make for improved access to relevant information. At present, the key words or phrases provided manually by human abstractors and indexers are often restrictive. A desirable scenario would be a full IR system which included the facility to generate a list of terms representing the *aboutness concepts* of a text. These could then be matched, in whole or in part, to the query term provided by the user.

1.1.4 The COMMIX System

As has been mentioned above, compound nominals facilitate the concise expression of elaborate topics of the type required to represent the aboutness of a piece of text. This work describes the COMMIX System, which automatically generates compound nominal expressions, as representations of complex topics which reflect the aboutness of the texts to which they refer.

The task required of the COMMIX system developed in this work is therefore to generate compound nominal expressions which represent in a highly informative way what an input text is about. As we discuss in Chapter Seven, the incorporation, or extension, of this facility into a full retrieval system would give the equivalent of highly informative indexing terms for each document. This would facilitate the matching between a compound nominal query term provided by the user, and those generated by the system as representations of the aboutness of the text. The process

could either be applied to all texts in the database, or perhaps more usefully, it could be applied to just the initial set of documents matching a general query, and hence used to enhance the precision at this stage.

Such a scenario, however, would constitute a fully integrated IR system, which is not the aim of the work described in this thesis. Rather, this work concentrates on the main facility which is currently missing from that scenario: that is, the generation of informative and representative compound nominal expressions to represent the aboutness of an input text.

The overall approach adopted throughout the development of COMMIX has been to minimise the role of syntactic processing required to generate aboutness terms for the texts, whilst maximising the semantic processing, at a shallow level, by utilising the semantic information available from an on-line dictionary. Section 4 of this chapter discusses the theoretical motivation for this approach, whilst a detailed description of the system and the methodology are the subject of Chapter Four.

1.2 Related Linguistic Issues

This section introduces the relevant linguistic issues associated with this work, which relate in particular to the compaction of information into concise form, and to factors associated with nominal expression.

1.2.1 Compound Nominals

This work utilises the fact that compound nominals have the potential to be the linking agent between the user's query and the (automatically generated) expression of the content or aboutness of an abstract (and therefore of the document to which it refers). This work is concerned with the automatic generation, or production of compound nominal terms, by a system which takes normal text as input.

Although this work specifically does *not* approach the compound nominal from an analytical point of view, we nevertheless need to look at some of the theoretical issues associated with compound nominals, identifying aspects involved in their generation which have bearing on this work. In this regard, we see that there are two features of compound nominals which are crucial to their adoption as a means of expression of aboutness in this work. Firstly, the emphasis is on conciseness of the

expression; secondly, the terms are always expressed in the form of nouns, albeit of a non-simple variety. There are a number of interesting factors related to each of these features. These are introduced below, and discussed further in Chapter Three.

1.2.1.1 Nominal Expression.

The first 'crucial feature' of compound nominal expressions is the truism that they are nominal in character: that is, that they function as nouns. The traditional definition of a noun as 'the name of a person, place or thing' has the concomitant problem of vagueness associated with the notions of 'name' and 'thing': the more favoured definition being in terms of syntax and morphology. There is, however, something to be gained from adopting the traditional definition when dealing with compound nominals. These expressions exhibit such a variety of syntactic form that it is useful to 'step back' and consider them in terms of the traditional definition: the notion of a noun being the linguistic representation of something which has the conceptual status of a thing, or an **entity**.

The advantage of looking at a noun in this way is that it enables us to see even the most complex of nominal expressions as representing an entity, be it a concrete 'thing' or a non-concrete 'topic', about which something further can be said, or about which related information can be requested. An advantage of using the compound nominal rather than a more expanded noun phrase is that we avoid the problem of the reader getting lost during the parsing of an extended phrase. Thus, for example, if we wish to say something about the fact that the toy market has become dominated by the video games industry (for example, that it is a significant factor in the decline of children's health), then it is much easier for the reader/listener to follow the flow if the 'fact' itself is presented as a **given** topic, or an entity, about which something **new** is to be said. The same content could be expressed by using a compound nominal: 'the *video games industry toy market domination* has contributed to the decline in children's health'.

Although there are advantages to this kind of compaction into nominal form, there must be limits to the length of an expression, beyond which interpretation becomes difficult due to the accumulation of 'missing' (or, rather, implicit) information and the associated **ambiguities**. In addition to the primary use of COMMIX to generate aboutness expressions for texts, this work provides a tool that could be used in the

subsequent investigation of the **restrictions** which govern the production of compound nominals in respect of **length,** as well as those operating on the **ordering** of the constituents of the compounds.

There is a further aspect relating to the presentation of information in nominal form, which is relevant to this work, and this concerns the treatment of verbal information. Although we might reasonably ignore closed class words when identifying aboutness concepts, the information content of verbs is far less expendable. 'Active' information about some topic is given in the form of verbs: for example, we might be told that *'the video games industry is growing'*. We might then want to incorporate this active information (associated with the verb 'growing') within a complex topic, and then express it as part of a nominal expression, such as the compound nominal *'video games industry growth'*. This would have the advantage of allowing us to go on to say something else about this new, more complex, topic.

The process by which verbal information associated with 'grow' is translated into the form of a noun ('growth') is referred to as **verb nominalisation**: it presents as a fairly simple task for humans, but is highly problematic for mechanical systems. At the time of development of COMMIX there was no (electronic) dictionary which comprehensively listed the noun forms of verbs. COMMIX therefore utilises a rather over-simplified methodology for nominalising verbal information, which is described in Chapter Four. It should be noted, however, that since then such a facility has become available (NOMLEX, McLeod et al, 1998). COMMIX could usefully be extended to incorporate this kind of facility, although we would still be faced with the problem of **multiple noun forms**, which would require the specification of which noun form to use in which situation. As the NOMLEX system (ibid) shows, this would add the requirement for the subcategorisation of verbs, and inclusion of specific associated knowledge, as well as a more detailed parsing of the text.

The general point here is that much detailed information can be expressed in the form of a noun. The term **nominalisation** varies in its usage: it is sometimes used to refer to the formation of a noun from a different word class, (e.g., 'about' + 'ness') and sometimes to the derivation of a noun phrase from a clause (e.g., *the growth of the industry* from *the industry is growing*). This latter usage applies partly here, but does

not cover the more general process of compacting and translating information into nominal form which this work involves. To avoid the confusion which would result from extending the usage of the term 'nominalisation', the more general term 'nominal expression of information' is used in this work at points where confusion may otherwise arise.

1.2.1.2 Conciseness and Compaction

The second specific feature of compound nominals to be discussed concerns the compaction of information into concise form. There is one main concern associated with conciseness: 'expressing much in a few words' implies that something is in a sense omitted from the concise expression that would be included in a more elaborate expression. If we consider the concise term 'video game', we find that the Collins English Dictionary (2nd Edition, 1986) lists it as an individual entry, whose definition is given as *'any of various games that can be played by using an electronic control to move points of light or graphical symbols on the screen of a visual display unit'*. Whenever we use the compound term we are assuming that the reader/hearer has assimilated the full meaning, as appears in the dictionary definition (or something close to it), into the concise term. In the case of 'video game', people who encounter the term do not have to 'fill in' any information, such as that given in the definition: a video game is simply a video game. Its independent listing in the dictionary indicates that the compound form has, for most native speakers, undergone the process of **lexicalisation** to become an independent lexeme in its own right.

If we consider for a moment someone for whom the compound 'video game' has not become an independent lexeme, this illustrates another issue associated with compounding. The person may know the individual lexemes 'video' and 'game' of which the compound is composed, but not know the meaning of the compound term 'video game'. The meaning of the whole compound is not indicated fully by the meanings of its composite parts. In linguistic terms, the compound 'video game' does not exhibit the phenomenon of **compositionality**. There is some information which has become assimilated into the meaning of the compound term when lexicalised, which is not discernible by virtue of having a knowledge of the meanings of the constituents.

A person who is unfamiliar with such a compound term is left to try to fill in what for them is 'missing' information. But what kind of information is missing? Since the composite terms 'video' and 'game' have been 'glued' together to form a compound it must be assumed that there is some specific relationship that holds between the two elements. There are, however, various possible relations that might hold between the two. Is a video game a game which involves doing something with video tapes, such as throwing them around? Does it involve video cameras, or video recorders? There is nothing in the compound that tells the unfamiliar person that it is a game 'to be played on' a video machine (i.e., on a machine which shows a televised image). It is worth noting at this point that the problems typically associated with 'what has been left out' apply mainly to the interpretation perspective. From the point of view of their production, or generation, the fact that this information is only implicitly available after the compound has been formed need not necessarily be problematic.

Much of the past work (over the last thirty years or so) relating to compounds of a nominal type has concentrated on the specification of the types of **semantic relations** that hold between their composite elements. Examples of such relations include the 'part-whole' relation (e.g., 'chair leg'), the 'purpose' relation (e.g., 'on/off switch'), and the 'place' relation (e.g., 'blackboard notice'). A host of different authors have discussed the compound nominal (complex nominal, compound noun, etc.), a large number of whom have concentrated on the semantic relations holding between constituents of the compounds (e.g. Lees, 1970; Downing, 1977; Levi, 1978; Finin, 1980; Sparck Jones, 1983; Leonard, 1984; Lehnert, 1988). There have been attempts to specify an exhaustive set of such relations (e.g., Levi, 1978), and opinions differ as to the degree to which this is a realistic or worthwhile task, as there are often problems associated with the classifications (a criticism made by, e.g., Downing, 1977). Thus, considering the above example of *'video game'* as a *'game to be played on a video machine'*, are we to classify the 'to be played on' as a 'purpose' relation, a 'location' relation or some other type? Attempts to be rigorous about the classification of semantic relations are not convincing, particularly as judgement often varies substantially from one author to the next.

One such relation which is particularly difficult to identify unambiguously is the predicating relation, which we might refer to as the 'is at the same time a' relation. The example 'robot engineer' is ambiguous between the predicating interpretation

(an engineer who is at the same time a robot) and the non-predicating reading (an engineer who deals with robots). The assignment of the correct relation is dependent on the context of the text, specification of which in mechanical systems requires huge amounts of specific domain knowledge.

There clearly is a potential ambiguity of semantic relations involved in the *interpretation* of an unfamiliar compound nominal, whether it be one which is largely lexicalised amongst the population of native speakers, or a novel one such as the complex terms discussed above. From the *generation* perspective, however, these ambiguities are not necessarily problematic.

Taking a look at a complex compound nominal, it becomes clear that there is another type of ambiguity from which longer compound nominals may suffer: that of **ambiguity of structure**. Looking at the complex term above, *'video games industry toy market domination'*, we see that there are different 'units' within the compound. Thus, we might initially note that 'video games industry' is a unit, or that 'video games' is a unit. A closer look shows that different readings of the compound link different neighbouring items, so that the boundaries between units appear in different places. Is the reference to some toy produced by the video games industry, which is dominating some (unspecified) market? Or is it to the domination of some (unspecified) market by an 'industry toy' which is (possibly) a type of video game? Or is it the domination of the toy market by the video games industry? There are additional, less plausible readings, but it should be clear that the linkage between constituent units is potentially highly ambiguous, and that this is a different source of ambiguity than that arising from semantic relations.

There are other types of ambiguity (relating to **word sense** and **word class**) involved in the interpretation of compound nouns, and these are discussed briefly in Chapter Three. It is important to note, however, that as with the issues relating to compositionality (mentioned above), such ambiguities need not necessarily be problematic from the point of view of the construction and generation of compound nominals, particularly when their generation is from short texts, which is the case in this work.

1.3 The Motivation behind the work

This section presents the motivations behind the work which led to the production of the COMMIX System. There are two main strands to the motivation which underlies this work: the 'theoretical' and the 'practical'. The intention from the outset has been to pursue a specific theoretical viewpoint within the context of a practical system. In regard to the theoretical side, the aim has been to investigate the limits to which we can usefully push a methodology for the generation of novel compound nominal aboutness terms from text, which is based almost exclusively on a shallow semantic technique. From a practical point of view, the aim has been the production of a workable system which is useful in its own right, as a generator of aboutness terms, as well as providing a tool which could be used for the subsequent study of the constraints which govern the use of compound nominal expressions, at least within the context of abstracts.

1.3.1 Theoretical motivation

The theoretical thrust of this work concerns the following hypothesis:

shallow semantic processing based on the method of lexical overlap, applied to the definitions of distinct terms occurring in an existing text, constitutes an effective means of generating novel compound nominal expressions to represent the aboutness of the text.

It has been an aim throughout this work to investigate the extent to which the above hypothesis holds, when used as the basis for establishing the semantic relatedness between distinct items in a text. The COMMIX system combines this approach with a methodology which identifies and pools all information pertaining to related items, expressing such information as modifying constituents, which are placed in front of a head 'salient' noun (or nominalised verb). The head, according to this method, is taken as being whichever of the related items is the most salient in the input text.

In practical terms, COMMIX takes as input the text of an abstract, and generates as output a number of novel compound nominals intended to represent the aboutness of that text (i.e., the main concepts developed within the text, which reflect the essence of what the text is about). The practical research aim has been to use this system to investigate the degree to which the compound nominals generated by this process are representative of what the input texts are about. The system may also be

used as a tool for investigating the subsequent specification of general pragmatic constraints on the production of compound nominals for the expression of aboutness.

In terms of linguistic theory, the topics for consideration are those associated with the use of compound nominals. The various related issues which have been highlighted above are discussed in Chapter Three, with particular regard to the *generation* (rather than the interpretation) of compound nominals.

The essence of the approach adopted here is that, at least from the perspective of the generation of compound nominals, semantic information can remain implicit. This is contrary to popular methods of representing semantic and pragmatic knowledge, which rely on the explicit specification of all relationships that hold between items linked in a network, in combination with inference rules relating to the concepts in the network. It may be that this disposition can be traced back to Schank's 'Conceptual Dependency' (Schank, 1973), which required that concepts differing in meaning have different (and therefore unambiguous) conceptual representations. Although not all information about particular concepts is necessarily made explicit, there is no scope for information remaining implicit: everything must be either stated explicitly or derivable via inference rules.

The fact that ambiguity can be tolerated by COMMIX is not based on any cognitive claims about the nature of conceptual representations in humans. The method of deliberately leaving information implicit (rather than explicitly specified) is thus not necessarily cognitively motivated but can nonetheless be exploited in the generation of concise indicators of aboutness.

1.3.2 The Application Area: Practical Motivation

It should be clear from Section 1 that the main practical motivation has been to provide an informative, automatic means of representing what a text is about. If such a facility were successfully provided, it would take us a step nearer to providing radically improved access to relevant information, by facilitating the use of complex query terms in a fruitful way. It would also reduce the amount of time and effort taken to access information related to complex queries.

This motivation has resulted in the development of COMMIX, which automatically generates novel compound nominals from abstracts, to give a representative description of concepts which the abstract, and therefore the original text, are about. In the longer term, this facility would enable specific 'aboutness' terms (generated by the system) to be matched against specific user-defined search terms. It would also decrease the amount of reliance on 'direct match' techniques and Boolean operations on sets (the Boolean Combination Problem), and would improve the ease of access to relevant information.

1.4 Overview of the Thesis

Chapter One

Chapter One has given an introduction to the field of enquiry which is the basis for the work described in this thesis. It has presented an overview of the problem, and mentioned the main relevant areas of linguistic enquiry, which are discussed in more detail in Chapter Three. The overall motivation behind the work has been presented, from both a theoretical and a practical perspective.

Chapter Two

Chapter Two discusses the different ways in which the information contained within a text may be re-presented. It pays particular attention to the fields of abstracting and indexing, and identifies a significant gap in the types of re-presentation modes used, emphasising the requirement for an additional mode which represents the *aboutness* of a text.

The process of abstracting is discussed, with particular reference to the database used for this work, and some of the problems encountered by human abstractors are also addressed. The chapter reviews the main approaches to automatic summarization and indexing, with reference to some particularly relevant work by other authors. It discusses the methodological approach of lexical overlap, by which semantic relatedness between distinct items may be identified, and distinguishes work which is similar in overall approach from that which utilises similar shallow semantic processing.

Chapter Three

Chapter Three describes the linguistic phenomenon of compound nominals and discusses their essential features of conciseness and nominal expression of information. It emphasises the importance of adopting a broad conceptual view, particularly with regard to the notion of *necessary* nominalisation, by which complex topics may sometimes need to be viewed as individual concepts, and represented as nouns. This chapter also includes some reference to other approaches to the phenomenon of the compound nominal, but stresses the fact that much of this past work has been from the point of view of their *interpretation* rather than their *generation*.

Chapter Four

Chapter Four describes the COMMIX system, and discusses this particular implementation of the shallow semantic method of lexical overlap. After some examples of the application of this method in other fields, there is a discussion of its employment in this work. For clarity, the methodology is explained using a worked example, which indicates the main principles by which the system works. This chapter also specifies and discusses the key assumptions which underlie the approach taken. An assessment of the validity of some of the assumptions in relation to the initial stage of the implementation forms the basis of some specific improvements to the system, the incorporation of which is then described in the final section of the chapter.

Chapter Five

Chapter Five to a large extent stands alone. It gives an example of how the system may be used as a tool, and describes some investigations into the effect, on the performance of the system, of replacing pronouns in some of the texts by their referent types. It also includes some discussion of the different ways of dealing with multiple-term lexemes (such as '*video games*') and the repercussions of adopting the different treatments on the aboutness expressions generated by the system.

Chapter Six

Chapter Six develops and describes the procedure used to evaluate the performance of the COMMIX system over the data set of 20 texts. It discusses the traditionally-used evaluation measures of precision and recall, explaining why these measures are

not appropriate here. The notion of percentage attainment is introduced, which is the degree to which the aboutness terms achieve their maximum potential in terms of the degree to which they are judged to be representative of the texts for (and from) which they were generated. The results are analysed in 3 mutually compatible ways, and the findings are presented in detail.

Chapter Seven

Chapter Seven discusses the implications of the findings of the evaluation described in Chapter Six and summarises the contributions made by this thesis to the area of content representation, within the overall field of Information Retrieval. It discusses the benefits of the approach taken, with its emphasis on shallow semantic processing, and mentions some specific drawbacks of this approach. It suggests some particular performance enhancing improvements which could be made by the integration of the system with some additional facilities. The final section discusses the future directions and applications of the work.

Chapter Two

2. Abstracting and Indexing

The purpose of this chapter is to discuss the different ways in which the information contained within a piece of running text may be more concisely represented. It also identifies a gap in the set of re-presentation modes commonly used.

The first section provides a discussion of the different modes, which are distinguished according to their function. We make the functional distinction between, on the one hand, modes of *reproducing* the **gist** of a text, and, on the other, different ways of *indicating* what a text is about. The need for an additional mode of *indicating* text content is discussed, particularly in relation to the notion of the *essence* of a text, which is described here as one means of expressing the **aboutness** of a text. This is a more concise mode of re-presentation of text content than that provided by the abstract, yet more informative than indexing terms or key words.

Section 2.1 also discusses an analogy between this functional distinction (i.e. between the reproduction and the indication of text content) and the Abox/TBox distinction made in the field of Knowledge Representation. Whilst there is a discussion of the different modes of re-presentation of text content, we concentrate specifically on the abstract, which is the starting point for this work, and constitutes the level of the textual input to the COMMIX system.

Section 2.2 discusses some of the other approaches taken to the automatic re-presentation of content, concentrating specifically on the generation of summaries, and of indexing or key terms. In particular, this section discusses similarities with other work, some of which is concurrent with the presentation of this thesis.

The final section discusses at greater length the field of abstracting, and presents some relevant points in relation to the production of the abstracts used as the input data in this work. This section also sets the scene in regard to the practical aspect of the work described in this thesis, which aims at formalising and automating the specification of what a text is about.

2.1 The Re-presentation of Textual Information: Gist and Aboutness

Any piece of text encodes information of some kind, which can be described at a number of different levels of specificity. A reader of a text can re-present what they see to be the main points of the text content, and in doing so they are likely to alter and condense the wording appearing in the text in order to get the message across.

In the ensuing discussion we consider the overall aim of re-presenting the content of a text from two different functional perspectives. The distinction is made between, on the one hand, modes commonly used to *reproduce* text content, and on the other, modes used to *indicate* what the text is about. This functional difference between reproducing and indicating the content of text is reflected in the corresponding distinction we make here between the notions of *gist* and *aboutness*, which we can say are the corresponding representations at a conceptual level.

The term 'gist' has been used (e.g., Scott, 1993; Rino, 1996) to refer to the salient information and arguments presented in a text. There is, however, a relevant distinction to be made between this term, and the 'aboutness' of the same text. The gist of a text refers both to what a text is about, and to the arguments contained within any exposition, and any conclusions that may be drawn from those arguments. It is a restatement of the main points and the 'thrust' of the text, and gives an idea of the flow of information through the text. As such, the gist cannot be represented as an entity, or a 'thing', and cannot therefore simply consist of a nominal expression, however complex.

On the other hand, the aboutness of a text, or 'what the text is about' serves to indicate the content of the text. The term 'aboutness' was used by Hutchins (1977a) in relation to document analysis, but seems to have been little used since then. However, concurrently with this thesis, it is now becoming more commonly adopted (e.g., Boguraev and Kennedy, 1997a, b; Wacholder, 1998). The aboutness need not tell us anything about the arguments or conclusions behind an exposition, and it does not convey any information about how information flows, or is developed, throughout the course of the text. 'Aboutness' just tells us what the text is about, which may be one or more topics.

There are a number of different modes commonly used to represent the gist and the aboutness of text, and they convey different amounts of information and serve

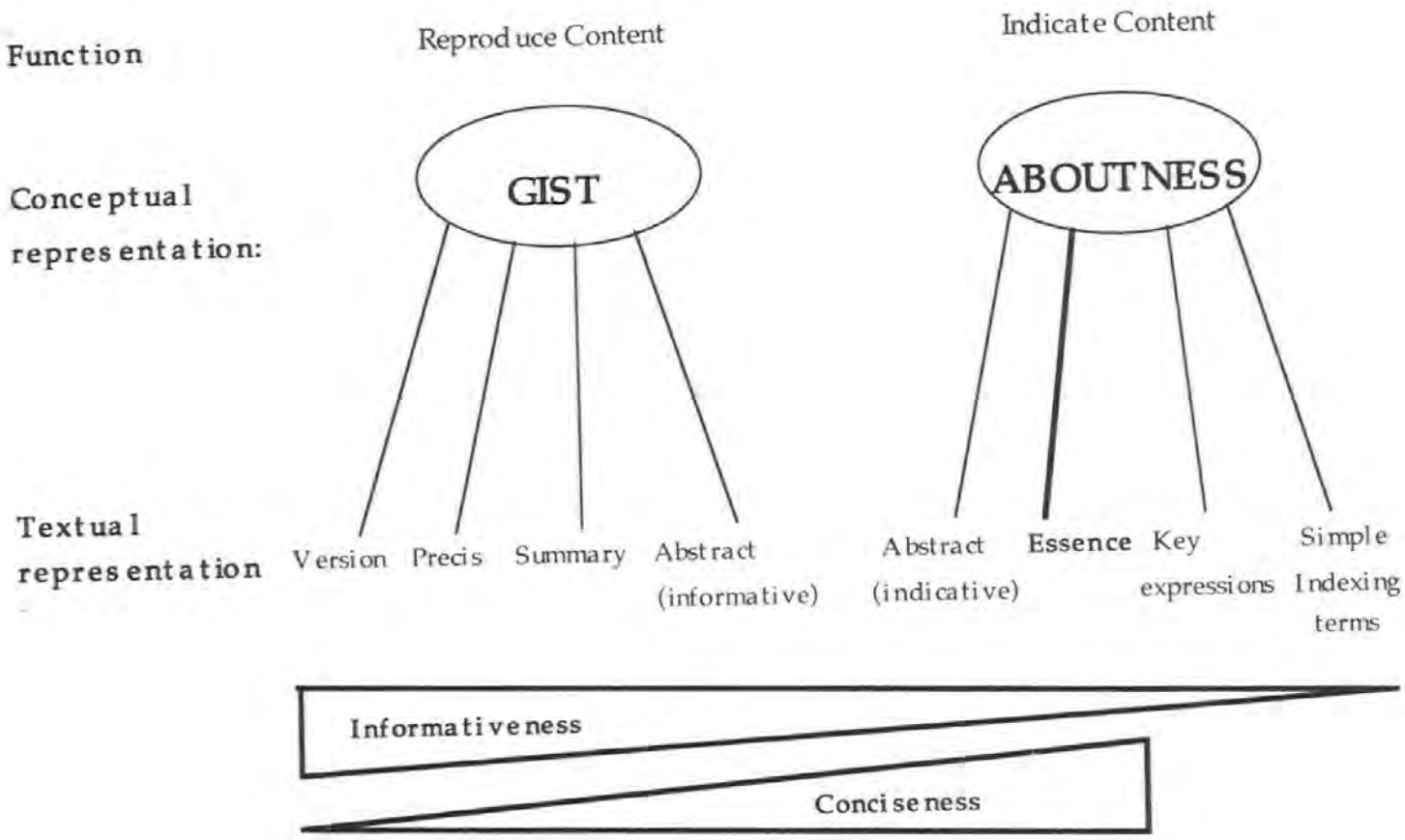


Figure 2.1. Modes of Re-presenting Textual Information

different purposes. **Figure 2.1** shows the way in which the different modes relate to one another, and shows their classification as either reproducers of the gist, or as indicators of the aboutness of a text.

2.1.1 An Analogous Distinction

Before continuing with a discussion of the different modes of reproducing and indicating content, there is a distinction worthy of mention here, which relates to that made above between the *gist* and the *aboutness* of a text. This distinction is one commonly referred to as being between the *Abox* and the *Tbox*. It is a distinction which is referred to in different ways by different authors, but the basis lies in acknowledging the difference between the *terminology* (or, for example, *topics, concepts, objects, terms*) and the *assertions* made about some world, based on the terminology. The distinction was popularised in the field of Knowledge Representation in the late '70s, specifically in work on KL-ONE (Brachman,1979), in which the terms 'Tbox' and 'Abox' were used to refer to the distinction between terminology, (referring, we might say, to 'building blocks') and knowledge pertaining to the assertions that can be made about them and the relationships between them.

This distinction bears some analogy to the one we make here between, at the functional level, the *reproduction* versus the *indication* of the content of text; and at the conceptual level, the *gist* versus the *aboutness* of a text. We can utilise this terminology in relation to **Figure 2.1**. In this way we can say that everything which relates to 'aboutness' in the diagram could be said to be *terminological*[5], in that it relates, at the conceptual level, to the topics, concepts and objects occurring (and being developed) in a text. Correspondingly, at the textual level, 'aboutness' is represented by terms which describe the conceptual entities in the text. As we discuss later (in Chapter Three), the essential feature of textual representations of aboutness (with the exception of indicative abstracts) is that they are nominal in character. On the other hand, everything that relates to the gist of a text could be said to be in some way *assertional*, relating to what is said about the topics, or entities, and the relationships between them. The remainder of this section discusses the different

[5] With the possible (partial) exception of the indicative abstract, but see discussion of abstracts which follows.

modes of re-presentation of content, and shows how the different modes may be viewed in the light of the Abox/TBox distinction. A gap in the modes currently used to indicate aboutness is identified and discussed, and is shown in **Figure 2.1** (in bold typeface) as being a representation of the *essence* of a text. The Figure shows how the different modes discussed in this section relate to one another, and to the notions of gist and aboutness, and repeated reference to this Figure is recommended. Although mention is made to all the modes shown in this diagram, the emphasis of this work is on modes of *indicating* the content, since this is the thrust of the work described in this thesis.

2.1.2 Textual Representations of Gist

We begin with the different modes of reproducing the gist, which we could say relates to 'Abox information'.

2.1.2.1 The Version

In re-presenting the content of a text, the writing of a different version has little purpose in terms of condensing the text. Rather, the normal reason for writing a different version of a text would be for the purposes of presenting the information in a different style or register, for a different audience or a different purpose. Indeed, the writer of a different version may well need to expand the text in order to improve its clarity, depending on the specific audiences for whom it was originally prepared, and for which its preparation is sought. The re-writing of text in relation to style and register, however interesting, are not specifically relevant to this work, and will not be discussed here, but may be pursued independently (e.g., Williams, 1990; Enkvist, 1973; Ghadessy, 1988; 1993).

2.1.2.2 The Precis

The term 'precis' is occasionally used to refer to a shorter version of a text, which includes all the main points. It does not seem to be often used in literature, and when occurring in speech, is usually synonymous with 'summary'. The Collins English Dictionary (3rd Edition, 1991) defines it as '*a summary of the essentials of a text*'. It is an impression of the author's that this term is in decline, although it might once have been used to indicate a highly informative summary. For this reason it is included in **Figure 2.1**, after the 'version', and is considered to be the 'most informative'

shortened version used in the representation of the gist of a text. We acknowledge, however, that in terms of its current usage, it does not necessarily constitute a largely different mode of representation from the summary.

2.1.2.3 The Summary

The distinction is made in this chapter between the abstract and the summary, since they constitute re-presentations of the information content to different degrees of conciseness and detail. However, the distinction is often not made, and some authors adopt either term to refer to a shorter representation of a text. The terms are sometimes treated as synonymous, for example, Sparck Jones, (1993a) is entitled "Discourse Modelling for Automatic Summarizing" and begins "Automatic abstracting is a challenging task ...", whilst Sparck Jones (1993b) has a sentence in (its) section 2.1 which begins "Summaries (i.e. abstracts) ... ". We must therefore note that literature relating to the summarizing process is not necessarily exclusive of the abstracting process. Similarly, some work which is said to relate to abstracting may often also be referring to summarizing.

In this thesis, the distinction between the summary and the abstract is upheld in the manner of Hutchins (1993) and, to some extent, Jordan (1993) and Maizell et al (1971). These authors all view the summary as being longer, and more informative than an abstract. In addition, we would add that a summary is viewed here as a less concise mode of expression of text content than that seen in an abstract.

After the version (and precis), then, a summary is the most informative representation of text content. It contains the most relevant points from the text, along with some associated information, but leaves some of the information unstated. A summary can vary according to its coverage, the informativeness required, the purpose for which it is being prepared, and the recipients for whom it is planned (Hutchins, 1993). Summaries are generally written in fairly concise prose, (although this can vary with the readership) and are likely to contain compact terms such as the compound nominals advocated as aboutness expressions in this work. It is, however, unlikely that information will be as densely compacted into such terse form as occurs with the longer expressions generated by the COMMIX system, which are not typical of normal prose. The length of a summary can vary according to both the amount of detail included in it, and to some extent the length of the original text.

From the point of view of the reader of a summary (who is a prospective reader of the full associated text), the summary may provide sufficient information for their needs, so they may not subsequently need to refer to the extended text at all.

The summary, then, is a method of reproducing the gist of the original text. It generally lacks some of the information present in the original, but includes the most important points and some associated information, and may in itself be sufficient to satisfy the information requirements of a reader. In terms of the ABox/TBox distinction discussed above, we can say that a summary contains an expression of most of the Tbox referents, along with the more important Abox information associated with them. It is presented as normal prose, and may include some more concise terms than those used in the original text.

2.1.2.4 The Abstract

The discussion presented in this section is aimed at giving a good idea of what is meant by the term 'abstract' in general. Since the abstract is the entry point for the work described here (and is the level of textual input to COMMIX), a more detailed discussion of abstracts and abstracting, with particular reference to those used in this work, appears in Section 2.3 of this chapter.

The purpose of the abstract is similar to that of a summary: i.e. to reduce the amount of information which readers need to attend to in order to be able to find the information they are seeking. Whereas a summary may well retain sufficient information to tell the reader what they want to know, an abstract will normally contain less information than the summary.

The reader of an abstract may get the information they seek from the abstract itself, but is more likely to consult the abstract in order to determine whether or not they need to read the associated text. From this point of view, the functions of an abstract are twofold. Firstly, an abstract cuts down the amount of work and time a reader has to expend when looking for information. If a reader wishes to know what the whole associated text is about, a well-written abstract will provide them with a good idea as to the content, without their needing to read the whole text. On the other hand, if a reader is searching for information about a specific topic, they can use the abstracts of a set of texts to help them identify a relevant subset which they may then wish to study in more detail.

In general, then, the view taken here is of an abstract being a condensed version of a piece of text. It is a short piece of prose, often only a single paragraph, which presents the main points of the associated text. It is not an expression of the full content of a text, and would normally be less informative in terms of detail than a summary. An abstract is, nevertheless, expected to give a good, concise idea of the main message or points. The style of a good abstract is usually more compact, with information being more densely expressed, than in either the original text or a version, precis or summary of the text.

There are differences of opinion relating to what constitutes an abstract. For example, Jordan's (1993) view is that the abstract contains mainly structural information, rather than the reproduction of content. We should note, however, that his comments relate particularly to very formal, technical texts, and as such, may indeed apply to that particular subset of abstracts. Some abstracts are simply composed of excerpts from the original text (Maizell et al, 1971; Endres-Niggemeyer, 1995) which have been cut and pasted together. Although this method of production may suffice in giving a reader an idea of what the associated text is about, they do little to convey the information in a cohesive way, and the text normally reflects the disjointed method of production.

An abstract may be written so as to be either **informative** or **indicative** of the content of the associated text (Hutchins, 1977b)[6]. An *informative* abstract aims at passing on the main points given in the original text, along with some of the additional information, albeit expressed in a more concise way. An informative abstract, then, is aimed at reproducing the gist of the original text, in a more concise way than typically occurs in a summary. Again, this can be expressed in terms of the ABox/TBox distinction discussed above. Thus, an informative abstract may be said to contain an expression of the most important Tbox referents, along with the more important Abox information relating to them. It is presented in the form of normal prose, and is likely to include some more concise terms than those used in either the original text or the summary.

As an example of an informative abstract, consider the following, which is taken from the data set used in the evaluation of the COMMIX System (in Chapter Six),

---

[6] Although some would disagree with this division, e.g., Endres-Niggemeyer in comments made at workshop (1993).

and which is also that used as a worked example in Chapter Four. This abstract is referred to as 'Text 1' of the data set of 20 abstracts (the complete set appearing in Appendix 1):

> *The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling. The main competitors in the market are Sega and Nintendo. Nintendo will spend 15 million pounds sterling on advertising over Oct-Dec 1992.*

This abstract has clearly been written with the aim of passing on the main points of the associated text, and goes some way to providing the reader with the same information they would get from reading the text itself. As such, it is an example of an informative abstract.

On the other hand, an *indicative* abstract is written with the aim of indicating what a text is about, rather than actually imparting any of the specific information given in the original. As an example, an indicative abstract corresponding to the informative one above might read something like the following:

> *The growth of the video games market is discussed, with particular reference to the actual and expected increase in its value over the period from 1991-1994. The main competitors are mentioned, with an example of the advertising budget of one of them.*

This is a functional description of the associated text, which merely tells the reader what kind of information is to be found in the original text. It contributes little towards actually imparting any of that information, and its function as an indicator of the aboutness of the text is clear. In ABox/TBox terms, then, an indicative abstract contains an indication of the Tbox referent types, but gives little, if any, Abox information. It is presented in the form of normal prose, but since it does not include much of the textual 'Abox' information, does not normally require concise expression.

Taking another look at **Figure 2.1**, it becomes clear that the abstract may be viewed as a 'bridge' between modes of reproducing the gist of a text, and modes of indicating the content, in terms of its aboutness. The next section discusses the utility of the indicative abstract for this latter purpose, along with the other modes used to indicate aboutness.

2.1.3 Textual Representations of Aboutness

Any piece of text is about something. We can say what kind of thing the text is about, referring to its main referents, with the aim of indicating its aboutness to prospective readers. This endeavour is distinct from that which aims at reproducing the gist of a text, and indicators are generally the first port of call consulted by readers when searching for information related to their requirements.

The indicators of text content which are commonly used are *key expressions* or *simple indexing terms*. This section discusses these modes of indicating aboutness, which may be viewed as different expressions of 'Tbox' information. We begin, however, with a few words about the appropriateness of using the abstract as an indicator of content.

2.1.3.1 Indicative Abstracts

Although useful to a degree in this role, an abstract written solely for the purposes of content indication carries a lot of redundancy. It is written as full prose, but only carries meta-information: that is, information about the kind of information expressed within the associated text. The indication of content tends to comprise informing the reader what thing, topic or subject the text is about. In this case, this type of information is perhaps better (or at least, just as well) expressed in the form of one or more noun phrases expressed in isolation, rather than being embedded in a paragraph of full prose. Chapter Three discusses in detail the issue of the nominal expression of information, and it will suffice at this stage to point out the relative unsuitability of the abstract for the purposes of content indication, particularly in cases where the reference is primarily to topics, objects or entities. In other words,

where the information related is primarily 'Tbox' information, it seems inappropriate and unnecessary to use full prose as its textual representation[7].

In such cases, it is more economical in terms of space and time, and indeed, more appropriate to use some form of nominal expression. The reader can thereby avoid having to read sections of full prose, which consequently need not occupy space unnecessarily within a database. Content indicators which comprise isolated expressions are also much easier to use as full or partial matches for user queries, rather than terms which are embedded in a paragraph of full, 'open' (i.e., not concise) prose.

2.1.3.2 Key Expressions and Indexing Terms

Indicators of content normally comprise individual, shorter expressions, which are typically known as *key expressions* or *indexing terms*. Both key expressions and simpler indexing terms serve to indicate to prospective readers the type of thing/s that the text is about. As such, they invariably take the form of nominal expressions, and may be said to represent one or more of the Tbox referents occurring in the text. Whilst simple indexing terms are often single words, key expressions tend to consist of key named referents (people, places, etc.) or key topics with some modifying expression. The latter can give concise reference to the central topic/s of the text, but are seldom longer than three-word expressions.

Personal communication with professional abstractors and librarians indicates that *indexing* terms tend to be controlled (i.e., selected from a list of permissible terms, which is designed to avoid ambiguity and duplication). On the other hand, *key terms* are generally more 'free-text', natural language expressions. The distinction, however, does not always hold, and the terms tend to be used interchangeably in the same manner as we see occurring with the terms *summary* and *abstract*, as mentioned above.

Each of these types of expressions are usually selected from the text by virtue of their importance (as judged by human indexors) or based on the frequency of occurrence of items in the text. They would often be expected to occur within the text of a well-written abstract, although personal communication with a professional abstractor

---

[7] These comments apply to situations in which the content is best indicated by means of reference to the main topic/s, rather than in cases where a lot of a certain type of information is related, which may indeed be better indicated by prose.

confirmed that in many cases they must be selected from a 'menu' of acceptable terms.

Some frequency-based methods of identifying indexing terms also utilise the generality that the first sentence of a text tends to be disproportionately representative of the content (Baxendale, 1958), although the success of indexing based on this assumption clearly relies on the texts being well-written in the first place.

The identification of key expressions and indexing terms has been at least semi-automated for some 40 years, and the remainder of the discussion relates to issues involved in their automation. These are discussed in Section 2.2 of this chapter, which discusses some of the different methods used to automate the re-presentation of text content.

2.1.4 The Need for More Informative Indicators

So far in this chapter the discussion has been related to well-established modes of re-presenting the content of a text. We have made the functional distinction between reproducing the gist of a text, and indicating its aboutness, where, by analogy, the expression of aboutness may be seen as the expression of Tbox information.

It is useful to view the different ways of representing the information from a text as a sequence throughout which the representation becomes progressively shorter, more densely packed with information, and therefore more concise in terms of the expressions which it typically contains. In short, then, the amount of information regarding what a text is about decreases in the order shown in **Figure 2.1**. Like many processes which are described as sequential, the stages really represent slices through a continuum, in this case, a continuum of decreasing informativeness and increasing conciseness of representation, with the latter extending across the realm of gist, and entering into the realm of aboutness.

Although a well written informative abstract imparts much of the content of the text it represents, from the point of view of an information seeker, it is time consuming to sift through a large set of abstracts of potentially relevant documents. What is often required is fairly detailed information concerning what the document is about, but not in the form of full prose.

As discussed above, indicative abstracts can be useful indicators of content, but have the disadvantages of not containing any 'Abox information'or concise indicators of content. On the other hand, key expressions and indexing terms are often too general, relatively uninformative, and constitute a variably accurate means of indicating content.

One further potential source of information about the content of a document is its title. The title of a document traditionally has the role of indicating the content of the associated text, and can sometimes be of more use to the information sifter than either an abstract or key terms. Unfortunately, however, titles are not always truly representative of the content. This is particularly true of newspaper articles, where headlines and titles have the primary purpose of attracting prospective readers. The abstract given above provides a good example of the anomaly often found between the title and what the text is about. Although this particular abstract is about the video games industry and the money spent on advertising by associated companies, its title (from the original document which appeared in *The Observer* of October 11th, 1992) is 'When the Chips hit the Fan'. This title is clearly coined so as to have maximum 'reader attraction potential', and is highly uninformative of the actual aboutness of the document it heads. The degree to which a title is representative of the content of the text is, therefore, highly variable and unpredictable, and it is clearly inadvisable to rely on title information, especially when dealing with newspaper texts.

There is a need, then, for a level of representation of what a text is about, which is more concise than the abstract, does not require the reading of lots of text, and gives a good informative indication of what a document is about. Such a representation would, in effect, bridge the gap between the informative abstract and key expressions, and provide highly informative and representative aboutness expressions which relate specifically to the content of the associated document.

It should be clear from Chapter One that this thesis concerns the generation of such aboutness expressions, in the form of highly informative compound nominals, from the abstract relating to a text. The information represented in these expressions is more compact than occurs in the corresponding abstract, and each expression acts as a far more informative indicator of content than is typical of key words or indexing terms.

2.1.4.1 The Missing Indicator of Aboutness: Essence

In this section we introduce the type of aboutness expression which is being advocated here, with particular reference to **Figure 2.1**. As this figure shows, there is a scale of informativeness, which traverses the different modes of re-presentation shown on the diagram, decreasing from left to right. There is a corresponding scale of conciseness of expression, which extends (left to right in the diagram), increasing, into the realms of aboutness, but does not include the most commonly used modes, namely 'key expressions' and 'simple indexing terms'.

The Tbox/ABox analogy is useful here. We have seen that these latter two modes, which appear at the lower end of the scale of informativeness, can be viewed as being modes of expressing (mainly) Tbox information. They are not intended to include assertional information, and do not typically constitute concise terms of reference to complex topics (as discussed in Chapter One).

The intuition underlying this thesis is that there is a gap in the modes of indicating aboutness. This gap could be filled by a type of expression of aboutness which would express the main topics, or referents, of the text, but which would additionally include some of the ABox information which pertains to the referents. In other words, our incentive would be to *make the Tbox information more Aboxey.* In functional terms, expressions of this sort would provide readers with a good indication of what a text is about, at a more informative level than that achieved by consulting currently existing modes which indicate the aboutness.

The main thrust of this work concerns the exploitation of the compound nominal as a highly concise form of nominal reference, which can be used as an expression of just this sort, to fill this information gap that exists in modes of aboutness indication. This gap is at the level of expression of the *essence* of a text, where 'essence' is a subtype of aboutness terms of reference.

We have shown here that there are already well-established modes of referring to the general aboutness of a text, and specify the 'essence gap' as a missing mode. It is important to note here that the more general term 'aboutness expression' is used throughout the thesis to refer to the more specific usage described here: that of 'essence'. The decision to use the more general term was based largely on the clarity

of the term 'aboutness', in the expectation that the term 'essence expression' would not be as clearly representative of the function.

In regard to the stages of representation indicated above, then, these aboutness expressions add another slice through the continuum, which in effect constitutes an additional layer of representation occurring between 'abstract' and 'key expressions'. If we think in terms of content indicators, it becomes clear that this additional stage, in effect, processes a natural language statement of the gist of a text (in the form of an abstract) and produces indicators of the aboutness which are more informative than those currently produced as key terms or indexing expressions.

2.2 Approaches to Automatic Content Re-presentation: Related Work

There has been an enormous amount of work done in recent years, aimed at automating the processes by which textual information is reduced and re-presented in a different form. A discussion of each system which has been built to do such a task is well beyond the scope of this thesis, and we should stress that the lack of inclusion of any particular system in no way implies a belittling of its value. It is important, however, to discuss the main approaches taken to the automatic reduction of text, and to mention the main areas, with some noteworthy examples.

2.2.1 Content Re-presentation as 'Information Extraction'

There is a sense in which the term 'Information Extraction' (IE) should be treated as a general term, largely synonymous with the expression 'automatic content re-presentation'. Used in this general sense, its coverage would include any mechanism by which the content of particular texts is filtered, with the selection of certain subsets of material to be re-presented in some form, depending on the purpose driving the task. Used with such a coverage, the term 'IE' would include all methods of reproducing and indicating content (with the possible exception of the 'version'), as already discussed in relation to Figure 2.1. The different re-presentations would then simply be alternative instances of IE, differing in granularity, and produced according to the end purpose for which they are required.

There is, of course, a far more specific usage of the term 'IE', which has evolved from a background of template-filling endeavours (e.g., work done by Wilks in 1964, reported in Wilks, 1987; Cowie, 1983; Sager, 1981; DeJong, 1982). The general aim of IE systems is to identify and record highly particularised types of information from

large numbers of texts. The information is then 'slotted into' templates, with the result that, for each text, the required information is recorded in a highly ordered and specified manner. Templates are typically preconstructed by hand, but as Wilks points out (in Pazienza, 1997, p. 7), a major research issue is that of developing automatic techniques by which templates can be derived from corpora found to contain significant patterns of template-like material.

When the information required to be extracted from a large number of texts is of the same type, the filling of slots in templates is a highly effective means of performing this task. The requirement for this kind of task tends to be associated with specific domains from the outset, and so the 'criticism' of the methodology being knowledge-hungry and domain-specific is not directly relevant. Any attempt to adapt existing IE systems to new domains will, however, be subject to these disadvantages, a point also mentioned by Wilks as a possible limitation of IE systems (in Pazienza, 1997, p. 7).

The type of information generally required to be extracted by IE systems centres on entities and the changes that occur in relation to them.. The emphasis is therefore largely centred on identifying noun phrases, and recording the information associated with the entities to which they refer. Changes described in relation to these entities (such as promotions, hirings, sackings etc.) are highly relevant to the typical IE task, and to the final recording of representative information. (For a good coverage on IE systems see Pazienza, 1997, and papers therein).

The work described in this thesis relates directly to IE in its former, general sense, but relates neither in its specific objective, nor in its methodology, to the latter, restrictive usage of the term 'IE'. COMMIX has the advantage of being domain-independent, and does not require templates. It concentrates on building compact, informative compound nominal expressions, packing information into expressions, rather than locating existing noun phrases and extracting information from them.

The following discussion of automatic means of re-presenting content does not therefore aim at describing work in the field of (restricted usage) IE. Rather, it concentrates particularly on the main approaches taken in the more relevant fields of automatic summarization, abstracting, and indexing. Note that the first two of these terms are often used synonymously, and are therefore discussed in combination.

2.2.2 Summarising and Abstracting

Early work on automatic summarising and abstracting was based on the identification of topic sentences using a number of surface indicators. These included, for example, frequency of occurrence in the text and the position in the text (e.g., Luhn, 1958); cue words, title information, and heading words (e.g., Edmundson, 1969), and 'self-indicating' sentences which were characterised by their starting with (or containing) certain combinations of words, such as *It is the aim of this work to show ...*, *'the results indicated that ...'* and so on (e.g., Baxendale, 1958). The sentences so identified were virtually cut and pasted together, and the resulting texts were consequently disjointed and incohesive.

Later approaches have utilised some of the principles of Artificial Intelligence in the endeavour. For example, DeJong (1982), in FRUMP, used partially-specified scripts (cf. Schank and Abelson, 1977) in a frame-based knowledge specification. Fum et al, (1985) based their system SUSY in logical representation of text, using production rules to govern the identification and combination of topics. Reimer and Hahn (1988) developed TOPIC, which again had its basis in the representation of semantic knowledge as frames. Although Rau et al (1989), in SCISOR, use a hybrid approach, this is again based on the detailed knowledge specification required of a deep semantic approach. Paice (1990) emphasises the importance of just such a full underlying layer of knowledge specification, which is aimed at facilitating a full understanding of the text. The value of the type of frame-based semantic methods used to represent this knowledge is discussed by Paice (1990), in relation to the overall aim of automatic abstracting. He also gives a good summary (ibid.) of the different methods adopted in the pursuit of both automatic sentence extraction (or identifying the most relevant sentences within a text) and textual cohesion (or avoiding disjointedness).

As Paice (ibid.) points out, using a detailed semantic representation of knowledge facilitates the resolution of anaphoric references, and therefore helps to overcome the problems associated with lack of cohesion, which typically result in a disjointedness in the final text.

Although the specification of detailed semantic knowledge is clearly of benefit in the condensation of text, this method has a severe drawback. The reliance on the specification of detailed knowledge means that each application is highly domain specific. Although there have been attempts at automating the acquisition of knowledge, to fill predefined templates, these attempts have met with limited success (e.g., the CRYSTAL system of Soderland et al, 1995).

Other approaches to identifying the important topics of a text have been based on a shallower statistical analysis, requiring full parsing. Salton et al (1994; 1997), for example, take a probabilistic approach to the identification of the important topics. Some of these methods (e.g., Kupiec et al, 1995; Aone et al, 1997) allow the user to manipulate the features which specify the sentences to be extracted. These latter authors base their approach on analysing a text sentence by sentence, comparing each one to the abstract of the text, and then computing the probability that it should be included in the summary, based on its similarity in relation to the abstract. This approach is still popular, e.g., Teufel and Moens (1997; 1998) and Mani and Bloedorn (1998), but still relies on putting together extracts of the original text (see the comment by Hovy and Lin below).

More recently, studies of the discourse structure of abstracts and summaries (e.g., Sparck Jones, 1993a; Liddy, 1991), as well as procedural models of abstracting and summarising (e.g., Endres-Niggemeyer, 1990a), have been used to guide, if not to direct, the course of automatic abstracting and summarising (e.g., Rino, 1996, on the analysis and automation of scientific abstracts; Aretoulaki, 1996, on the specification of pragmatic features to describe segments of text). However, the situation does not seem to have fundamentally changed, and the overall approach is still to identify the most important topics and track them through the document. Excerpts of the text which refer to these topics are then juxtaposed/concatenated in the form of a 'summary'.

Although we have merely skimmed the surface in terms of describing these approaches, the general point to be made here is that attempts at *intelligent* automated summarization still have a long way to go before they can truly replace human hand-produced summaries. Some of the many interesting systems which have recently been implemented are detailed in Mani and Maybury (1997). In addition, Aretoulaki (1996) contains a good comparison of the different

methodologies, in relation to summarising in particular. However, as Hovy and Lin (1997) point out, there are several current systems which claim to perform text summarization, but "their so-called summaries are actually portions of the text, produced verbatim" (ibid., p. 18).

2.2.3 Indexing and Key terms

We turn now to methods used for automatic indexing, and the identification of key words. Attempts at automating the indexing process began in the late '50s, when Baxendale (1958) described some experimental work on different methods of reducing the number of words used to represent a text. He found particularly that identifying and using phrases consisting of nouns and modifiers constituted the best indexing terms, although using isolated nouns along with their adjectives, and their frequency of distribution in the text also represented the content significantly.

Since then, most automatic indexing and key term generation has been based on a combination of topic identification and frequency of occurrence of individual words or word combinations. Since it was first reported by Baxendale that the first sentence is particularly representative, this belief now plays a significant part in topic identification, and hence in the automatic generation of indexing terms. The indexing terms identified are assigned different weights according to the degree to which they are judged to represent the content. In an IR system, query terms may then be compared to these indexing terms in order to assess the relevance of each document to the query.

The SISTA project (Whitehead 1994) is a semi-automatic indexing system, which generates indexing terms for use by indexers. It takes as input an abstract, and returns a list of prioritised index terms. Based on statistical parsing, it clusters sequences of words, based on their part of speech category. It uses a large corpus of pre-indexed abstracts to train the statistical model. Abstracts are then parsed, and "potentially diagnostic constituents" are identified and compared to the correct (i.e., previously prepared) indexing entries.

The main approaches to automatic indexing are described in Salton (1991). Methods have generally been based on frequency of open-class words and suffix-stripping, combined with weighting factors, and it is the specification of the latter which tends to vary. A method which has become common, but is not new (e.g., Sparck Jones,

1971) is based on the incorporation of a thesaurus, reference to which allows the inclusion of hypernyms (in this case), but also of synonyms. Such methods, however, are still based on identifying the most salient topics, which tend to be single words.

On the other hand, Jacquemin et al (1997) have shown how the expansion of multi-word terms (based on a largely syntactic approach) can be useful for automatic indexing. In COMMIX, however, the opposite ploy is instantiated: namely, that of constructing, rather than deconstructing, word groupings in order to represent what the text is about.

Park et al (1996) discuss the importance of the compound noun in the field of automatic indexing, but concentrate on the analysis of the compounds in terms of their component nouns (by "extracting useful component nouns from compound nouns", p. 518). They concern themselves with the compound nouns which exist already in the text, and not, as we do here, with their construction, which can yield more representative aboutness terms.

2.2.4 Work Showing Similarities to the COMMIX System

We have seen that much of the work in automatic summarization has centred around the identification and extraction of the important sentences from texts. The sentences are then combined in the form of a summary. Whatever the specific methods used to determine which sentences are important, the common trait is that the extraction and combination, or juxtaposition of sentences taken from the original text suffice to form a representative summary.

The work described in this thesis goes below the level of the sentence, and concerns itself with the identification of the most salient topic(s), combining information which pertains to it or them. Thus, the COMMIX system does *not* aim to produce summaries or abstracts of texts. Whilst it is concerned with the reduction of the information in the text, it treats some elements of the text as redundant, and aims to condense the most important topics ('Tbox information') along with the information which pertains to it (related 'Abox information'). The emphasis is on constructing new, concise forms of reference to complex topics (Abox-enhanced Tbox entities), which represent what the text is about.

There are some approaches which bear similarities to the working of the COMMIX system. The similarities are sometimes in terms of the conceptual approach, and

sometimes involve similar methodology, but different overall aims. This section discusses recent similar work in relation to this distinction, beginning with work which adopts a similar perspective to that taken in the development of COMMIX.

2.2.4.1 Conceptual Similarities

In general, any approach which recognises the importance of the concepts expressed in a text (rather than simply the actual words used) is approaching the problem from the same point of view. Some examples are given here, although there are clearly many systems constructed from the conceptual point of view.

Sparck Jones and Tait (1984), although approaching the problem of automatic searching of a database from the perspective of the user's search term, take a strongly conceptual view of the search for information. They recognise the generality that indexing terms are simply taken verbatim from the text, and emphasise the need for alternative representations of the concepts involved. Thus, they advocate the use of variant terms, even though their interest involves the generation of variations of the search term rather than variations of the forms which occur in the text.

More recent work by Evans and Zhai (1996) bears some similarity to the approach suggested in this work. These authors recognise the importance of noun phrases, and in particular the importance of 'subcompounds' from complex noun phrases. However, their work involves indexing on the basis of the *analysis* of such subcompounds, rather than on their construction, and their analysis is based on corpus statistics combined with linguistic analysis. They recognise the fact that, ideally, the structures representing the indexing items should reflect the linguistic relations between terms, resulting in a more accurate representation of the conceptual content of the text. They also recognise the need for shallow processing, since a deep understanding of the text is not normally necessary for the purposes of representing what it is about. They express the opinion that "ideal indexing terms would directly represent the concepts in the document", but add that "concept-based indexing is an elusive goal" (ibid.), since they are difficult to define, represent and extract. As they also point out, nearly all commercial IR systems index purely on words rather than trying to identify more complex structures, and are thereby severely limiting their power to discriminate between items which at a glance seem related. Rather than attempting concept-based indexing, then, these authors focus their efforts on a phrase-based approach, recognising two main types of useful

an entire document (they process the document a paragraph at a time, and then merge the resulting representations into a graph), and that content should not simply be represented as a concatenation of sentences which refer to the most important topics. The method they use, however, is extremely domain-specific, relying on the initial manual specification of concepts (hand-coding the specific concepts occurring in the text set to be processed). Hahn and Reimer provide no evaluation of their system, nor specific examples of input and output, so it is difficult to comment on its performance. However, we can say that even if the output graphs proved to be highly informative of the content of input documents, this methodology remains severely limited in extension, since it requires such intensive knowledge input.

In contrast, the COMMIX system is domain-independent, and exploits the easily available lexical resource WordNet (Miller et al, 1990), which it effectively uses as a knowledge base. It generates compound nominal expressions of the essence (a subtype of aboutness) of text, rather than producing more abstract representations, such as the text graphs of Hahn and Reimer.

2.2.4.2 Methodological Similarities

The exploitation of lexical resources, particularly dictionaries, is not new, and has been going on since Amsler (1980) established a (hand) taxonomic representation for a subset of the *Merriam-Webster Pocket Dictionary*. The history of the use of computational lexical resources, including 'machine-tractable dictionaries' (a phrase coined by Wilks et al, 1987) is clearly documented and discussed in Wilks (1996).

The methodology used in the COMMIX system to establish semantic relatedness between distinct textual items is described in detail in Chapter Four. When this system was initially implemented, the method was fairly new, and it is a sign of its success that it has since then become used by a number of different research teams. The method, known now as *lexical overlap,* (Lesk, 1986) is based on identifying semantic relatedness between distinct textual items by virtue of finding words which are common to the definitions of each of the terms.

Binot and Jensen (1987) implemented a system which attaches prepositional phrases using weighted links, where the links are identified via lexical overlap between items looked up in LDOCE. Jensen (1991) stresses the point that the text itself is the ideal

medium for representing meaning, and that shallow processing is therefore generally sufficient if no deep conceptual interpretation is required.

Much of the work which adopts the method of lexical matching, including the original Lesk paper (1986), is aimed at disambiguating multiple word senses. Luk (1994), for example, searches for semantic connections (based on lexical overlap) between words in the same clause, with the aim of linking different senses of a word to its clausal context. Although such work differs considerably in its overall purpose, the basis for establishing semantic relatedness between distinct items is the same as that used in COMMIX.

A variation of this methodology, which traces lexical overlap through chains of items, has become known as *lexical chaining* (Hirst, 1987). In relation to the methodology employed in COMMIX, lexical chaining corresponds to the extension of the search for relatedness to multiple levels (see Chapter Four). This notion has its basis in lexical overlap, and is being used increasingly in the identification of semantically related textual items. Barzilay and Elhadad (1997) give a good summary of the evolution of this methodology, which began with a theoretical computational model for defining the cohesion relations between items occurring in *Roget's Thesaurus* (Morris and Hirst, 1991). Barzilay and Elhadad use this method for the purpose of identifying and extracting significant sentences from text. Although they analyse text segments below the level of the sentence, they then proceed with the standard approach of summarization, 'putting together' the complete sentences which contain the related referents. Although this method can give cohesive end results, which present a reader with the most important points, the summaries produced do not increase the conciseness of expression in any way, and readers therefore get a patchy idea of what the text is about.

Although he does not use WordNet, Ginsberg's (1993) automatic indexing system does have its semantic analysis based on tracing links between nodes in a semantic network in order to find common hypernyms, thereby asserting that the hyponyms are semantically related. Once again, though, there is no mention of the possibility of using this relatedness as a justification for the cross-application of modifiers of one term to another related term, which features strongly in the COMMIX system.

One final approach to be mentioned here is that developed by Resnik (1993), which is based on the notion of semantic distance (semantic spanning), and uses WordNet as

the means of assessment of such distance. This approach is also used by, for example, Sussna (1995) in relation to Information Retrieval in particular. It is similar to the approach adopted in COMMIX, to the extent that a specified goal is to "push the semantic distance approach ... " (ibid.). The work differs, however, in that Sussna does this without using term frequency, basing the whole of his approach on the semantic distance between *query* and *document* terms as they stand.

This section has discussed a number of approaches to automatic summarization and indexing which bear similarities to the COMMIX system. Boguraev and Kennedy (1997) stress the need for indexing terms to be elaborate and informative expressions, rather than simple terms. Evans & Zhai (1996) recognise the usefulness of at least small nominal compounds for representing content, whilst Wacholder (1998) stresses the importance of heads of noun phrases, and of clustering heads in order to obtain salience ratings.

To recap, then, COMMIX deals with the automatic processing of short texts (abstracts), from any domain, from and for which it generates compound nominal expressions of the essence (a subtype of aboutness) of the input text. It utilises the methodology of establishing semantic relatedness between distinct lexical items by means of lexical overlap. It identifies salient items of the text, based both on frequency of occurrence and on existing pre-modification in the text. It collects together the modifying information pertaining to related salient items, and constructs compound nominals to express the accumulation of modifying information along with the most salient of the related head nouns. At the time of writing, this combination of methodologies constitutes a novel approach to the problem of indicating the essence and aboutness of text.

The next section deals with matters relating to the data used as input to the COMMIX system. It discusses abstracting in general, and gives some detail of the constraints on the professional abstractors who produces the abstracts used in this work.

2.3 The Starting Point for COMMIX

The 'entry point' in the sequence of representations of the information in a text, as far as this work is concerned, is the abstract. Jordan (1991) discusses at length the linguistic genre of the abstract. He does, however, relate his comments about their

structure largely to the scientific domain. The view taken of an abstract in this thesis is not as restrictive as Jordan's, and applies across any domain, although we recognise that some (e.g., biographical) texts are more difficult to represent in the short and concise form of an abstract. We recognise that the guiding aim of anyone who writes an abstract is to produce a statement of what they see to be the main message conveyed by a text, and thus provide the potential reader with a good idea of the sort of information contained in the text.

2.3.1 Variation within Abstracts

There is enormous variation in the degree to which an abstract represents or indicates what its corresponding document is about. An abstract may be written by the author of the original text or by an independent reader, who may be a professional abstractor. It is beyond the scope of this work to provide a detailed study of approaches to, and styles of abstracting. However, informal examination of abstracts from several sources reveals that it is common for author-abstractors to simply cut and paste extracts from the original document together and the resulting abstract is then likely to give 'snapshots' across the whole picture rather than provide a good, concise idea of what the text is about. Abstracts (as well as other texts!) which are produced in this way tend to have a jerky, disjointed, incohesive style, and do not exhibit the same degree of succinct expression as occurs in purpose-written abstracts, particularly those written by professional abstractors.

2.3.2 Advantages of using Professionally-Written Abstracts

What does a professional abstractor do that cannot be done by an author-abstractor?

2.3.2.1 The Author's Perspective

When an author writes an abstract for a piece of text they have produced themselves they have a clear idea of the information they intended to convey in the text. However, they are likely to view what the text is about from a biased viewpoint, which will be influenced by their intentions in writing the original text. In other words, the author may well have had in mind a primary message they wished to convey, but the text is likely to contain information about other things as well. An author-abstractor, then, cannot avoid writing an abstract from a subjective, and often restricted, point of view.

2.3.2.2 The Reader's Perspective

Different readers will read a text from different perspectives, seeking different information from it, or putting different slants on the information contained within it. So, from the point of view of different readers, a given text may be viewed as being about a number of different things. An abstract which purports to represent what its associated document is about, but which has been written from a subjective point of view, will be less representative than a well-written abstract, written by a professional abstractor from an objective point of view.

It could be argued, then, that a text is better represented by an abstract which has been written by someone who can view it objectively. Although it could be said that each individual reader cannot avoid putting their own personal bias onto anything they read, a professional abstractor is more likely to give an unbiased view of what a given document is about.

It is well recognised that an expert in a particular field is not necessarily good at passing on their knowledge to the lay person. There are clearly exceptions, but in the general case an abstract produced by a professional abstractor (who is not an expert in the field) is likely to be more informative as far as the lay reader is concerned than an author-written one. Professional abstractors specialise in formulating and utilising concise means of expression, although there may be constraints under which their work is carried out.

2.4 The Database used in COMMIX

There are many databases of abstracts which could have been used as a source of material for the input to the COMMIX system. The majority comprise collections of abstracts taken from particular journals, where each abstract has simply been stripped off the front of its associated text. For the purposes of this work, however, it was considered more appropriate to use abstracts which have been purpose-written by professional abstractors. This gives a degree of consistency between the different abstracts, and has the advantage of having an associated set of specific 'guidelines' to which the abstractors were strongly encouraged to adhere. These constraints are discussed in the following section.

Many databases of abstracts have a rather specific coverage: they tend to be restricted to a particular science, or group of sciences, to economic texts, and so on. In terms of the current work, however, in which the domain-independence is emphasised, it is preferable to test COMMIX on a set of abstracts with a broad range of coverage. The actual database selected for use[9] has a broad coverage in terms of subject matter. In fact, it covers the top one hundred periodicals, judged by frequency of readership by undergraduates, and thus includes a wide variety of specialist and non-specialist journals, in addition to a wide range of daily newspapers. It therefore provides an ideal source of abstracts relating to a wide range of general topics.

2.4.1 Subset of the Database used to Test COMMIX

The methodology developed in this work is aimed at having a broad applicability. Throughout the work an underlying aim has been to avoid falling into the trap of creating a system which will work only in a limited field, and which requires 'tailoring' in order to be applied to another specialist field. Rather, COMMIX has been constructed to allow it to apply generally, across a broad range of subject matters, and copes with any abstract whose words appear in the dictionary used. The generality and cross section of subject matters we find in newspapers makes the use of abstracts relating to newspaper articles a good choice in this work. It should be emphasised, however, that any abstract from the database could be used as input, as indeed could any abstract from any source, although highly specific terminology associated with specialist fields may not appear in the lexical resource. If this were the case, we could still expect some output expression (based on direct matching of items in the text itself), but this output would be more limited.

2.5 Constraints on Abstractors

The database company whose abstracts are used in this work has a number of guidelines to which their abstractors are strongly encouraged to adhere. These guidelines were obtained during a site visit, at which point the data to be used was collected, and an in-depth (informal) discussion took place with the co-ordinator of the group of abstractors. In addition, there was one follow-up meeting, and some telephone conversations with one of the abstractors practising as part of the group.

[9] General Academic Index

The following section describes those guidelines which are relevant to this work, and draws on the aforementioned discussions.

2.5.1 Set Guidelines for Abstracting

Abstractors are encouraged to make their abstracts informative wherever possible, rather than simply indicative of the type of content to be found in the text. The guidelines state that they should use clear and concise natural language prose which is their own, rather than lifting sections of text from the original document. There are three methods of abstract writing, often used by novice (or hurried) writers, which are specified as inappropriate and therefore disapproved. These are: the rearrangement of sentences from the beginning of the text; cutting three or four ideas from the main text and pasting them together; and actually writing the abstract 'on the wing', whilst reading the text for the first time (and therefore before having obtained an idea of what the whole text is about). It should be clear that abstracts written according to any of these methods will be neither concise nor reliably representative of the content of the text. To have them specified as outlawed methods can therefore only help in the overall aim of obtaining truly representative abstracts. This can only be of benefit to this work.

Abstractors are also required to provide key words and synonyms for key words, with the aim of enhancing the effectiveness of searching on isolated terms. The company has a 'controlled vocabulary[10]' which should be used to provide synonyms wherever possible, although abstractors may use terms which do not appear on this list if necessary.

The company specifies three types of abstract: long (100 -150 words); medium (50-75 words); and short (about 30 words). In each case the guidelines specify that the first sentence should be an informative one, summarising the main point of the article. The remainder of a long abstract should give details on the secondary points, and give any supporting data. Medium and short length abstracts should have this secondary information more concisely presented, with only one or two sentences available for secondary points of short abstracts, which may therefore need to be indicative rather than informative.

---

[10] Note that this is an example of the conflicting definitions of key terms and indexing terms. The key terms here are controlled, rather than being a freer representation of content.We therefore treat these 'key terms' as at least partially synonymous with 'indexing terms'.

There are specific guidelines relating to grammar and punctuation, although these need not be discussed here, since they simply aim at pinpointing what are commonly seen as errors. More relevant to this work are the guidelines on writing style. These specify that abstractors should aim for simplicity and clarity, and they pinpoint things to avoid, namely, long and complicated sentences, short and jerky sentences, unfinished sentences and 'wordiness'. Abstracts must be written to enable readers to understand what the original text is about without needing to refer to the title or any indexing terms.

In addition to being aimed at the whole abstract, there are also style guidelines aimed at the sentence level. Thus, *subject-verb-object* sentences are encouraged (because they are easy to read) and the use of dependent clauses is particularly discouraged, especially at the beginning of a sentence, again in the interests of fast and easy reading.

Whenever a text is about some non-simple or specialist topic, reference to the main point can require either compact expressions of the type advocated in this work, or specialised jargon terms. Although the use of jargon is discouraged, it is recognised as being necessary on occasions, to facilitate succinct expression of the principal topic. There is, however, a further restriction which inhibits the direct reproduction of long phrases (such as compound nouns), which states that strings of three or more consecutive words lifted out of the original document would constitute plagiarism and should not be copied directly. Rather, abstractors are required to use a paraphrase to express the same thing, and must suffer the consequences both of having to use additional words (in the context of having an upper word limit), and of losing conciseness of expression.

Although the specification of the above guidelines may seem superfluous here, this does give us a clear statement of the required features of an abstract. They also go some way to guaranteeing the quality of the abstracts which are used as input to the COMMIX system. Therefore, by using informative abstracts produced by abstractors adhering to these principles we can ensure that the input to COMMIX is well-written, cohesive, normal (although concise) prose and highly representative of the content of the associated texts. This assumption would not necessarily hold for other sets of abstracts, taken from databases which have no such restrictions or guidelines, although there is no practical reason why such input should not be used.

2.5.2 Advantages of the Guidelines

As stated above, there have been studies (e.g. Liddy, 1991; Endres-Niggemeyer, 1993) which have sought to formalise the structure of abstracts, with automation as a major aim. It seems clear that, in a similar vein, the specification of the working constraints underlying their production can only be advantageous as far as this work is concerned, since the prescribed structure and guidelines can be assumed to hold for all the abstracts used as input.

Perhaps the most relevant example in this respect is the requirement that the first sentence should be particularly representative of the content. This is especially useful in regard to the overall aim of the work, namely, to identify and express the aboutness of an abstract, since we can then operate on the assumption that the first sentence will not only be concise, but, more importantly, will be disproportionately representative of the aboutness.

In addition, the guidelines set standards and to some extent a methodology for all abstractors to adhere to, which clearly encourages good writing, and results in an increased likelihood of the abstracts being truly representative of the content of their texts.

2.5.3 Disadvantages of the guidelines

Note that the comments here relate to the specification contained within the guidelines themselves. The main disadvantage of the constraints is that they are perhaps over-restrictive, a point which was confirmed by informal discussion with one of the abstractors employed by the company. One particularly constraining guideline was that which restricts the length of expressions which can be lifted directly from the original. According to this restriction, abstractors must not use expressions comprising strings of more than three words, where such strings exist in the original document. Thus, for example, if the original text uses the compound nominal expression *computer games industry growth*, the abstractor would be constrained from quoting the expression directly in the abstract, even though the term is highly representative of the information content and should feature strongly in the abstract of the text. It is of interest to note that this constraint is applied not for any linguistic reason, but rather as protection against accusations of plagiarism. This

restriction can leave abstractors feeling unable to adequately express the key issues, and may result in longer and/or disjointed text.

Other disadvantages are secondary, and would include the strict limits placed on the length, and the strict adherence required to the predetermined indexing or key terms, which abstractors must provide in addition to the text of the abstract itself.

2.6 Summary (not abstract!)

This Chapter has discussed matters relating to the more condensed representation of the information contained in a text. It has described in particular the *summary* and the *abstract*, as well as *key expressions* and *indexing terms*.

A particular contribution has been the presentation of the functional distinction which can be made between the *reproduction* of the *gist* of a text, and the *indication* of its *aboutness*. In this context, a missing mode of re-presentation is described, this being a mode of indicating aboutness, and this is referred to as being at the level of expression of the *essence* of the text.

Section 2.3 has presented an overview of approaches taken to the automation of different modes of re-presenting textual content, with the emphasis on general approaches to summarization, abstracting and indexing. It has given an idea of how similar work relates to the COMMIX system, mentioning particularly work which is concurrent with the submission of this thesis.

The field of abstracting has been described in more detail, with particular reference to the characteristics of, and the reasons for selecting, the database of abstracts used as the starting point for this work.

Although this chapter has mentioned the importance both of adopting a conceptual approach to indexing, and of the compound nominal as a good means of expressing aboutness, the main discussion of these topics follows in Chapter Three.

Chapter Three

3. Compound Nominals

The main purpose of this chapter is to discuss the characteristics of compound nominals which make them particularly useful for the expression of aboutness. The discussion centres around two essential features of compound nominals, i.e., *nominality* (by which term is meant 'being nominal in character') and *compactness*, and examines aspects of linguistics which relate to these features. Section 1.2 of Chapter One laid out these principal points, and their elaboration is the main purpose of the present chapter.

3.1 General Approaches

There is a substantial history of studies concerning linguistic constructs which are similar (if not identical) to what are termed in this work 'compound nominals'. The majority of the studies have concentrated on the problems associated with their analysis (e.g., Arens et al, 1987; Finin, 1980; Leonard, 1984; Resnik and Hearst, 1993; McDonald, 1982; Sparck Jones, 1983; Vanderwende, 1993), rather than looking to exploit their characteristics in the manner described in this work. As such, they are not directly relevant here, although they do exemplify some of the interesting features of compound nominals, which are the primary concern of this chapter.

We cannot hope to even summarise all the approaches taken to the *analysis* of compound/complex nouns/nominals, but can say that the two main approaches are based on statistical analysis of corpora on the one hand (see, e.g., Lauer, 1996; Charniak, 1993) and, on the other, feature specifications, which require detailed, knowledge-hungry semantic specification (e.g., Gay & Croft, 1990).

One of the problems with approaches like this latter one is that they rely on the basic assumption of there being a set of specifiable categories, according to which all nominal compounds may (and should) be classified. A closed set of relations (semantic roles) need to be prespecified for each type of category, and nominal compounds are required to be formulated according to the specifications. This kind of prespecification of features, relations, roles and so on, which is inevitable in this type of approach, has been criticised on a number of occasions (notably by Downing,

1977). It has been one of the incentives behind the work described in this thesis to avoid falling in this mould.

There have, however, been a number of studies which recognise the potential of 'compound nouns' (e.g., Evans and Zhai, 1991; Isabelle, 1984; Marsh, 1984) and 'nominal compounds' (notably Sparck Jones & Tait; 1984), from the point of view of their construction, and it is this emphasis that is most relevant to this thesis.

3.1.1 A Broader Definition of 'Compound Nominal'

It should be stressed that the discussion of the compound nominal, its characteristics, uses, and its exploitation in this work, relate specifically to English, and do not necessarily apply to other languages.

The notion of compound nominal is best viewed from the point of view of its distinctive features, which were briefly described in Chapter One, and are further elaborated in the current chapter. We also use the notions of **head** and **modifier,** which become particularly useful.

The phenomenon under discussion here is not simply characterised as a string of nouns, the last one being the head (Lees, 1970), or even (in its fuller sense) as a head noun preceded by a list of nouns and/or adjectives. As a general phenomenon, a compound nominal may be said to comprise a head, which is nominal in character, preceded by any type of modifying expression.

Also relevant to a general definition are the two essential features of compound nominals: nominality (or being nominal in character) and compactness. As discussed below, the feature of compactness means that there is always some information implicit in their form (even if only a preposition, as in the example 'chair leg'. In this case the preposition 'of', representing the semantic relation 'PART-OF' has become implicit). Purely speaking, then, a full definition of compound nominals should include all examples of ADJECTIVE - NOUN combinations (such as 'red chair'), in which the semantic relation 'has colour' is not explicitly mentioned, but remains implicit. Perhaps more obvious (and famous) an example would be the case of something like 'orange saw', the ambiguity of which is discussed at length by many authors (e.g., Levi, 1978). A general definition of this sort is as follows:

> *a compound nominal expression is a combination of a head, which is nominal in character, preceded by at least one modifying expression, with the result that some*

information contained within the phrase becomes implicit. The head may itself comprise more than one word, and the modifying expression may be one or more further compound nominals, or a stand-alone phrase.

There should be no requirement for the modifying expression to be a syntactically well formed expression in its own right (although it can be), since the process of compaction, which renders information implicit can result in the successive removal of particular classes of words - typically prepositions and articles. For example, whereas the example '*A you scratch my back and I'll scratch yours approach*' shows that the modifier can be a stand-alone, syntactically well formed phrase, the example '*eliminating-vapour contamination dry running pump development*' (which is one of the aboutness expressions generated for Text 4) indicates that it need not be.

Although the general description and definition of compound nominals has a broad scope, the type of expression used in the generation of aboutness expressions in this work is restricted to a subset of the whole. In this work, then, we generate as compound nominal expressions any noun or compound nominal, which is the main topic being referred to, as the head, preceded by a modifying expression, *M*, where *M* comprises any string of words of the class noun (including cardinals and nominalised verbs), proper name, adverb or adjective.

3.1.1.1 The Head of an Expression

The notion of *head* is generally referred to as a syntactic phenomenon, which is normally used in the context of a syntactic description of a sentence or phrase. It is used here in its general sense, to refer to the main element of a phrase, which is distributionally equivalent to the complete phrase.

In this sense, we may think of the head as being a linguistic statement of the general type of thing being discussed, and thus as a general representation of the type of topic, which may be modified by the accompanying text in a variety of ways. We should note that although the notion of head does refer to other syntactic parts of speech, its use in relation to this work is predominantly in regard to nouns (or information which can be converted into nominal form, as discussed in Section 3.2).

The notion of head is used in this work, then, to refer to the main topic of an expression (a semantic notion rather than a syntactic one), by which is meant the

type of entity represented by an expression. It need not necessarily be represented only by a single lexical item, but can consist of more than one word.

3.1.1.2 Modification

The sense in which the term *modification* is used in this work applies to information occurring before and/or after the head. In more formal terms, we allow for both pre- and post- modification, rather than distinguishing between modification and qualification on the basis of position in relation to the head (i.e. occurring before and after the head respectively), as occurs in Hallidayan grammar.

In this thesis, we utilise two related senses of the term *modification*. In one sense, we use it to refer to the placement of modifiers (or 'modifying information') in the aboutness expressions constructed by COMMIX as output. In the second sense, it refers to any information which enriches our understanding of the particulars relating to the general topic. In terms of the Abox/TBox distinction discussed in Chapter Two, it refers to Abox material. It should be clear by now that COMMIX aims to identify all such pertinent information, packaging it together into an expression which adequately combines and collectively represents the essential information expressed in the text about the topic. Although the common understanding of modification may not extend this far, it could be argued that this is the essence of what modification is all about.

The COMMIX approach, then, is that any information pertaining to a 'topic' which occurs somewhere in the text can potentially be included alongside the head which refers to the topic to which it refers, together with all other information which modifies the same topic. When identifying modifiers of a particular topic which has been judged as being a type of thing the text is (at least partly) about, we do not restrict the notion of modification to being pre-head. Rather, the approach is to identify all the information which pertains to, or relates to, or already modifies that topic, and create an overall modifying expression which is applied to the main head.

The result is an expression of which the topic is the head, with the remainder being the modifying expression, which comprises any *legal* combination of all related modifiers. The specification of what constitutes a 'legal' combination is one of the experimental areas of investigation for which the system could subsequently be used (see Chapter Seven, Future Directions).

3.2 Nominality and Compactness: Related Linguistic Issues

Chapter One introduced the notion of the compound nominal as a highly representative means of expression of complex entities, attributing their appropriateness to their two principle defining features. Firstly, they are by definition nominal in character and function; secondly, they have enormous potential for the compaction of information into a highly concise form of expression.

So what is the relevance of these two general features?

3.2.1 Nominality: Nominalising and Lexicalising Given Information

The essence of nominality, or being nominal in character, lies in the fact that what is being represented is a thing, or an entity[11], at least from the conceptual point of view.

A text is about something, or some things, and those things, having the conceptual status of entities, are representable by some type of nominal expression. From the point of view of the writer of a text, treating a complex chunk of information as an entity allows something to be said about that chunk. For example, if we wanted to communicate the following state of affairs: *There is an industry which has grown up around the use of video games, and that industry now dominates the toy market* (which state of affairs will be referred to as *SA1*), and to go on to say something about that state of affairs, it would be necessary to present (or at least conceive of) *SA1* as an entity.

To illustrate this point we can consider how a writer may continue on to say something else about *SA1*. Perhaps the most common means of saying something new about this state of affairs would be to use the determiner 'this' in an anaphoric role in the subsequent sentence as in, for example, *This has had a detrimental effect on the health of our children.* Here, the determiner 'this', which begins the second sentence, refers to the whole information content of the first sentence (which represents *SA1*).

Such anaphoric usage of 'this' carries an unspecified but implicit associated noun, such as 'situation' or indeed, 'state of affairs'. Anaphoric 'this' is thus the determiner at the beginning of the implied noun phrase *this situation* or *this state of affairs*. Indeed, the use of the determiner 'this' is often a sign of a writer treating some previously

[11]As pointed out in Chapter One, this reflects the traditional notion of the noun as a representation of things, places and names, now often seen as problematic in its evasiveness.

mentioned chunk of information as a thing, or an entity, about which they can then go on to say something else.

The simple use of a determiner is sufficient if an author wishes to refer to the unnamed 'thing' just once, immediately after its initial introduction. If, however, they wish to make repeated reference to the same thing, it needs to be named in order to facilitate further discussion about it whilst avoiding the necessity of repeating long-winded sections of text. The 'thing', or entity, be it an event, a situation, state of affairs, etc., needs, in effect, to be accorded a label.

3.2.1.1 Nominalisation

When it comes to the linguistic representation of entities, such as the 'state of affairs' given in the above example, the writer needs to express the information as some sort of a nominal expression. The term applied to this process of expressing what has been non-nominal information in nominal form, is *nominalisation*. This is another term whose usage varies widely across the linguistic community, both historically and currently, and therefore its meaning in this work needs to be specified. The term is often used in a restricted way, to refer solely to the conversion of a non-nominal form of a word into its nominal counterpart, such as *dominate -> domination, grow -> growth, happy -> happiness,* and so on. In fact, this word-class conversion into nominal form may be considered as a subprocess of a more general process of nominalisation which does not centre around the word as the basic unit to be nominalised.

The manner of usage of the term 'nominalisation' in this work is broader than this restricted, word-centred usage, and is more in line with the approach adopted by, for example, Comrie & Thompson (1985: 349), who state: "The term 'nominalization' means in essence 'turning something into a noun' ". That 'something' can be equally a chunk of information relating to an event, a state of affairs, or an action. A similarly 'open' usage is adopted by Cumming (1991: 43), who states that: "a morphological relationship with a verb or adjective is not a necessary condition for an NP to be considered a nominalization."

To summarise its usage here, then, we use the term *nominalisation* to refer to the general process by which previously non-nominal information, which may be represented by single words or larger chunks, become represented in nominal form, We may then specify the subclass, such as 'verb nominalisation' (i.e. the word-class

conversion from verb to noun), 'phrasal nominalisation' and so on, as required for clarification.

3.2.1.2 Types of Nominalisations

So what kinds of nominalisations might a writer use? Reference has been made in the above section to 'verb nominalisation' and 'phrasal nominalisation', each of which is particularly salient to this thesis. These two subclasses of the general process will be elaborated in this section.

Verb nominalisation

In cases involving the conversion of single non-nominal lexical items, it may be that there is already a noun form of the word established in the lexicon, which can be used to replace the non-noun. For example, the expression *x dominates y* can be represented as a 'state of affairs' entity by nominalising the verb and using the passive voice, to give:

(a) *the domination of y by x;*

(b) *the x domination of y*; or, most succinctly,

(c) *the x y domination.*

Similarly, given the initial expression:*the video_games_industry dominates the toy_market,* we would correspondingly be able to express this as:

(a) *the domination of the toy market by the video_games_industry;*

(b) *the video_games_industry domination of the toy market;*

(c) *the video_games_industry toy_market domination.*

Where there is no corresponding noun form of the non-noun already occurring in the lexicon, the introduction of a new term, defined by the author may suffice. If such a newly coined term is adopted by many of the readership it may become fully lexicalised: i.e., it may become sufficiently well established to be considered a member of the lexicon of that language.

In the COMMIX system described in this work, we are concerned with expressing the main content words in nominal form. Since these are primarily nouns and verbs, the main type of single-lexeme nominalisation is therefore that of verbs.

Providing the nominal form of a verb automatically is no trivial matter. Even if the noun forms of verbs were listed in their dictionary glosses (which is sometimes, but not always, the case), there would be no simple way of distinguishing between multiple noun forms of a particular verb and selecting which to use on a particular occasion. We would find, for example, that *dominate* has corresponding noun forms *dominance* or *domination*; *receive* has *receipt* or *reception*. In such instances it would be desirable for the system either to obtain the noun form of a verb from the verb's gloss (which information is not always included), or to specify a series of verb -> verbal noun translation rules, such as *(transitive) -ate -> (abstract) -ation*, and to select the appropriate form (where multiple forms exist). However, such a methodology would require much more specific lexical knowledge to be made available to the system (to cope particularly with the idiosyncrasies of word formation, and the identification of a noun's semantic role in the text, the specification of both of which lies beyond the scope of this thesis). Such a facility could be added as an enhancement of the basic system at a later stage.

In view of these difficulties, which are a distraction from the principal aim of this work, the COMMIX system relies on a relatively simplistic, but adequate methodology for providing these verbal nouns (i.e., the noun forms of verbs). This process (which might be referred to as semi-nominalisation) involves expressing the verb in the form of an abstract noun by adding *-ing* to the verb stem, giving the verb in its gerund form. Thus, for example, the nominalised form of *grow* which is used by the COMMIX system is *growing*, as in, for example, *the growing of the industry*. This is clearly not an ideal solution to the problem of providing nominalised forms of verbs, but it does render the verbal items at least with some nominal character, and therefore constitutes a crude but workable solution to the problem.

Phrasal Nominalisation

In cases where the information to be nominalised is more than a single word, it seems clear that there are three options open to the writer. The first possibility would involve the specification of the appropriate nominal form of whatever general state of affairs, or situation is to be enlarged upon. In other words, the writer may select only the head of the nominal expression for use at the beginning of a separate 'new information' sentence. In this case, the author selects the type of thing being referred to, without mentioning the specifics, relying on the assumption that the reader

remembers those details. In the example *SA1* above, the writer may select *domination* as the nominal form corresponding to the verb *dominate*, using it to make more specific the anaphoric use of the determiner 'this'. Thus, the 'new information' sentence would become: *This domination has had a detrimental effect on the health of our children.*

If this option is chosen, however, *this domination* may not be informative enough to convey the required meaning, particularly if the term is to be used at a later point in the text, at a distance from the initial introduction of the thing being referred to. As a second solution, the writer may utilise an open nominal expression, such as *the domination of the toy market by the video games industry* and continue to add the new information to the same sentence, giving, in our example, *the domination of the toy market by the video games industry has had a detrimental effect on the health of our children.* Using this type of open expression is, however, a wordy option, which leaves less scope for adding new information, since the resulting sentence becomes too long and consequently difficult to follow.

The third option open to the writer is to use a denser, more compact compound nominal expression of the type advocated in this thesis. This facilitates the presentation of the already specified information more succinctly, which may be done for reasons of limited space, style, clarity or, indeed, necessity (see section 3.2.1.3). By doing this, the author acknowledges the entity-status (requiring nominal representation) of the subject matter and in effect accords it a label which is concise and yet still highly representative of its nature.

This option also enables the writer to be highly specific in their reference to the entity. They can add more of the relevant specific information, but without either being excessively wordy, or relying on the reader to have memorised the specifics.

We can use the subject matter of Text 1 (of the data set shown in Appendix 1), to illustrate how the author of the text might further discussion about the information which appears in the abstract. In this example, the thing under discussion may be viewed as being a combination of a type of *industry* along with the fact that it now dominates the toy market, with the verb itself being viewed and represented as an entity (which could be termed an event-entity). Thus, the entity or thing which is the topic of the discussion becomes the *domination* of the toy market by a particular type of *industry*; in this case, the *video games industry*, the whole of which in its most

compact form (as a compound nominal) becomes *video games industry toy market domination.* In furthering the discussion of this thing (in order to say something else about it), the fully compacted compound nominal may then be used as the subject of a sentence which includes the 'new' information, resulting in, for example, the single sentence: *The video games industry toy market domination has had a detrimental effect on the health of our children.*

3.2.1.3 Necessary Nominalisation

Halliday (1988) sought to explain the type of written language used in physical science, although his comments do apply generally to expository text. In this paper he reported his observation that in a piece of scientific text, the exposition often *requires* that complex bits of information be accorded nominal status, describing nominalisation as "an essential resource for constructing scientific discourse" (ibid., p. 169), and acknowledging that nominalisation has the effect of "packaging a complex phenomenon into a single semiotic entity" (ibid., p. 168). In other words, in cases where the exposition requires that chunks of information be referred to again, it can actually be *necessary* to coin a nominal phrase to represent the chunk, which may be a reported observation, in order to be able to continue on to say something else about it. In Halliday's words, it is akin to saying "you remember what I said just now? - well, we're going to move on from there" (ibid., p. 168). As he points out, it is essential to be able to do this, since 'what I said just now' may be "the summation of a fairly complex argument" (ibid.).

3.2.1.4 Theme and Rheme, Given and New

Reference has been made at several points above to *new* information, and it is pertinent here to make some comments about the information content of a text in terms of what is already known to the reader, and what is presented as new information. Of the many terms which have been used to refer to different aspects of the information content of a text, this discussion will focus (sic.) on the notions of Theme versus Rheme, and Given versus New.

The term *theme* is used here in the manner of the Prague School of linguistics (see, e.g., Lyons, 1977), according to which it refers to that part of a sentence which is already known to the reader, and which is thus the least informative in terms of providing new information. The theme constitutes the point of departure of the message, which in English occurs at the beginning of a clause. In contrast, the term

rheme refers to that part of the sentence that carries information which is in some way new to the reader, or which cannot be predicted from what has already been said.

Section 3.2 thus far has discussed the different types of nominal expression an author may use in expressing a complex semiotic entity. In terms of the theme of a sentence, then, we can say that in instances where the author uses anaphoric means (such as determiners and pronouns) of referring to the entity, that theme is bound to the previous text. In such cases the theme may also be said to be assumed as *given*, or already made known or established by the preceding text. Here, *given* refers to the information which has already been provided within the previous linguistic context. In contrast, the term *new* refers to just that: to information which has not previously been provided by the text.

The Given, then, represents what has been communicated to the reader already, and can be presented in a fairly succinct way, without risk of 'losing' the reader. In contrast, New information is usually more open and "often needs to be stated more fully than the Given (that is, with a longer, 'heavier', structure)" (Quirk et al, 1985, p.1361). Halliday (1988, p.163) provides some examples of this difference in degree of 'openness of expression', notably in his reports of the (compound nominal) terms *crack growth* and the subsequent *crack growth rate* as seen in a Scientific American article which he studied. He makes the point that it is only after the notions of the cracking of glass, the growth of such cracks, and the rate of their growth have been introduced at some point as New that the concise terms are found. At this stage, they appear at the beginning of clauses, and are used as background elements, about which the New can then be stated.

The major point of Halliday's paper (ibid.) is to make the claim that, at least in the genre of scientific writing, nominalisation is required in order to facilitate the continuation of an exposition. In addition to packaging complex information into a single semiotic entity, nominalisation has the effect of establishing as presupposed the information which has already been presented, and using it as a starting point, thus enabling the new information about it to be clearly expressed. The expression of something as a noun phrase thus has the function (within the discourse) of presenting it as Given within the context of the text. As Halliday points out (1988, p.168), it is usual in English for the Theme of a sentence to also be Given. In these cases the effect (from the rhetorical point of view) is known as *backgrounding*: i.e., the

backgrounded information is then assumed to hold true, becoming presupposed, and beyond question. In the example sentence discussed in Section 3.2.1.2 above, the information that there is an industry associated with video games, and that there is a toy market which is dominated by that industry, is presented as Given, or assumed to hold. As is typical in the case of presupposed, or Given information, the scope for questioning its validity is limited, since backgrounding bestows on it a high degree of credibility, and the strong implication that it is accepted and indisputable knowledge.

This bestowal of Given status on nominal information is interesting from a pragmatic perspective. It is not, however, problematic from the point of view of the COMMIX system. On the contrary, it can be seen as advantageous if considered from the perspective of the user. Someone using an IR facility wants to be able to ask about some thing, some entity whose existence (at least conceptually) is not in question, in order to find out more about it, even if the information they glean is to the effect that the thing does not exist! In effect, they are saying: 'I want to know more about this particular thing - what documents are there that will be relevant?'

If users have at their disposal a system which can process complex queries, presented in the form of compound nominals (or even create concise complex query terms from more openly phrased user input), then the scope for that system to provide the relevant information would be enhanced. This type of highly specific matching of complex queries, however, relies crucially on there being adequate representations of what the different texts are about, at a more informative level than is currently associated with indexing terms and key expressions. As discussed in detail in Chapter Two, it is precisely at this gap in the representations of aboutness of texts that the work described in this thesis is aimed.

3.2.1.5 Lexicalisation

There has been much discussion so far in Section 3.2 regarding the forces driving the representation of chunks of information in nominal form. The extension of this process logically ends up in the creation of new lexical items, be they composed of a single new item, or the combination of items which already exist. Fairly recent examples of items which have become lexicalised (by at least some members of the language community associated with the use of English) include *compact disc, video game* and many more. This process, by which new terms become established in the

lexicon, is a well-established field in the context of linguistics (see, for example, Lyons (1968, Ch. 9; 1977, Chs. 8 & 9); or Sager (1990) in relation to terminology in specific disciplines).

The process is mentioned at this stage, since it constitutes a link between the two essential features of compound nominals. The process of nominalisation, discussed above, may culminate in the formation of a fully lexicalised item, and lexical items formed in this way will carry an amount of associated implicit information, which results from the compaction of information associated with the nominalisation and lexicalisation processes.

As far as COMMIX is concerned, if items are not listed in the dictionary, they do not count as lexicalised items. The system does, however, check the dictionary for already-existing noun strings which, if found, are identified as lexical items in their own right. Thus, for example, it identifies *video games* and *compact disc* as each being independent lexical items, even though the input text shows no indication of this fact.

3.2.2 Compactness: Compositionality, Semantic Relations and Ambiguity

The second of the two principal defining features of compound nominals, *compactness*, also makes them particularly suitable for the expression of aboutness. We have seen above an example of how much information can be packaged into a highly compact compound nominal expression. The main advantage of this information packaging (that has been discussed so far) concerns the effective production of a single semiotic entity, which allows quite complex issues, states of affairs, situations, and so on to become the subject of further elaboration. In terms of the formation of compound nominal expressions, then, it should now be clear that the compaction of information may be optional or necessary, and that it may be conceptually or editorially motivated, or related to a particular style (such as expository text).

Although Halliday did not discuss compaction specifically, his comments regarding the necessity of nominalisation on some occasions seem to hold also for compact nominal expressions as well as nominalisations in general. In other words, some nominalised expressions may be too long to be adequately expressed as open nominals, and compact expression may be necessary in order to present the

nominalised information in a form in which it can be treated as Given, in order to facilitate the statement of the New.

There are a number of consequences related to this compaction of information into such expressions, which do not overtly state all the information they represent. These issues are often seen as problematic in the field of computational (and, indeed, theoretical) linguistics, particularly when approached from the point of view of their interpretation or analysis, rather than their generation. The remainder of this section discusses matters relating to the compaction of information, presenting the issues (as they are typically seen) and explaining why they need not be problematic from the point of view of generating compact compound nominal expressions to represent the aboutness of a text. It should be stressed that each of these related areas constitutes a vast subject in its own right, and the discussions here will be restricted to an indication of why the typical problems may not apply in the context of this work.

So what are these problems which are typically associated with the interpretation of such compact expressions?

As was mentioned in Chapter One, the main general issue which is seen as problematic from the point of view of interpreting highly compacted expressions is that the reader/listener is confronted with an expression in which the information content is not all overtly mentioned. In other words, some of the information content has become implicit in the concise form.

To illustrate this, we will consider the highly compacted compound nominal expression used already in this chapter as our example. We might consider an initial amount of information, expressed by the following 'open' phrase: *domination of the market relating to toys, by the industry associated with video games.* (This is assuming that *video games* is treated as an independent lexical item.[12]) A highly compacted compound nominal expression of this information would be *video games industry toy market domination.*

It is clear that there are three main types of differences between the open and compacted expressions of what is essentially the same information. Firstly, there are some words missing from the compacted form; secondly, some of the words that do

[12]As mentioned in Chapter One, this need not necessarily be the case for all speakers of the language, so in cases where it is not lexically established, the further expansion of *video games* would be necessary.

remain in the compacted form occur in a different order from that in which they occur in the open expression; thirdly, the boundaries between constituents of the compacted form are not clear (e.g. one of the constituents could be *video games industry toy*, in which reference is to a type of toy which dominates some unspecified market, rather than a type of industry which dominates a specific market).

There are a number of repercussions relating to these three differences, which may be generalised as concerning three main areas: firstly, the disparity between the meaning of the whole expression versus the meanings of its component parts (see section 3.2.2.1 below); secondly, regarding the unspecified relationships that hold between component parts of a given expression (see section 3.2.2.2 below); and thirdly, the high degree of ambiguity which results from compacting large amounts of information into denser expressions (see section 3.2.2.3 below). The following three subsections discuss these three issues (which are extensive fields of enquiry in their own right) with regard to why they do not need to be particularly problematic here.

3.2.2.1 Compositionality

When any amount of information is condensed into a smaller number of tokens, it is inevitable that some of the original information will become implicit in the compacted expression, unless alternative tokens (or lexemes) are used, which encompass the composite meaning within their normal definitions.

The general notion of compositionality concerns itself with the degree to which some whole may be analysed, described and even defined by the sum of its parts. Taking the example of '*chair leg*' or '*video games*', we would say that an analysis of the compositionality of these items relates to the extent to which the implicit information (discussed above) has become part of the meaning of the whole. To coin a phrase, the whole (meaning) has become greater than the sum of (or combination of) its parts.

The most interesting aspect of compositionality as regards this work, is that we can refer not only to the compositionality of items which have already been tokenised, but also to the extent to which the component parts of a text make up the whole of what that text is about. This point is made by Hutchins (1977a, p.26) in relation to the aboutness of a text. Reference is made here to work by Fairthorne (1969), who distinguished between the *extensional* aboutness of a text and its *intensional*

aboutness. The former refers to the topics of the component parts of the text, whilst the latter concerns the topic of the text as a whole.

If we use this notion of the intensional aboutness of a text, it becomes clear that the overall topic of a text is more than a combination of the topics of its component parts. It seems, then, that the notion of compositionality can be perfectly well extended to encompass the text as a whole, and that representations of the essence of a text (as discussed in Chapter Two) need not be expected to be entirely compositional.

Interpreted in relation to the aboutness expressions generated in this work, we can say that these expressions of the essence of the text, which are formed by the combination of some of the component parts (or extensional topics) of the text, can reasonably be expected to take on some meaning which is beyond a purely compositional analysis. However, from the perspective of the *generation* of aboutness expressions from text, this non-compositionality is not seen as problematic. On the contrary, it may be viewed as advantageous, in that it facititates the *implicification* (or the *rendering as implicit*) of textual information, upon which the concise compound nominal expressions depend.

Partial or total representativeness?

Related to the notion of intensional versus extensional aboutness (as discussed above) is the question as to whether the aboutness expressions generated constitute partial or total representations of what the text is about.

It should be clear that, from the perspective of a user of a document retrieval system, the facility to express complex entities will have the effect of increasing the specificity of the query terms they can use, thereby increasing the likelihood that the documents retrieved will be relevant to the query. As pointed out by Hutchins (1977, p.32-33), a full expression of the aboutness of a document (as perceived by indexers) is unlikely to match precisely to a query term expressed by a user of a document retrieval system: and it would be of little use if it did. The user of such a system is, in effect, expressing the theme of their interest; is giving the starting point from which they wish to go forward (to find out more). For a document to be of use to them it must contain new information about the theme. So, if the aim in producing aboutness expressions were to be totally representative of the content of the text, and if the most specific aboutness expression generated for a document were to match exactly with a user's query term, the document would be uninformative for that user. Therefore, in

the production of the COMMIX system it has not been an aim to necessarily formulate compound nominal expressions which represent the content in its entirety. Although some of the expressions generated by the system might be viewed as approaching a representation of an overall intensional topic, others must be viewed as partial representations, or, in Fairthorne's terms, as expressions of the extensional aboutness.

In these terms, then, the aboutness expressions generated by the COMMIX system have to be extensional, at least to some degree. This would give users of a full IR system the opportunity to match a variety of query terms against a number of aboutness expressions, which are seen as extensionally representative of the content of the text. The ultimate aim would be to generate, in addition to a number of extensional aboutness terms, an expression to represent its intensional aboutness, (although it is arguable whether such an aim could even be achieved).

3.2.2.2 Semantic Relations

The second issue to be discussed here is that relating to semantic relations. There is a degree to which this issue may be described as a subclass of the issue of compositionality. The effect of leaving out specific mention of, for example, the prepositions which, in the open form, specify the semantic relations between items, leads to some ambiguity as to the nature of the semantic relations that hold between the components of a compound expression.

There has been considerable study of the issues associated with the semantic relations that hold between constituent items of compact nominal expressions. Perhaps the most extensive studies were carried out by Levi (1978) and Lees (1970). Each of these authors discusses at great length the implications of specifying the semantic relations, but these views are both related to their *analysis*: in other words, of rediscovering what the semantic relations were, after they have gone (or at least after they have become implicit).

The rendering as implicit of the specific semantic relations which hold between items of a compound need not be problematic for the user of a system like COMMIX. On the contrary, as we mention above, this can be used as a positive and exploitable feature, allowing us to be non-specific in the type of relationship between composite elements.

It is likely, however, that a user of an IR system will have in mind some particular relationship between the components of a compound term they coin. It is here that the main disadvantage with the flexibility and non-specificity associated with this work lies. If we consider the extension of a COMMIX-based IR system to include a query processing element (as suggested in Chapter Seven: Future Directions), then it is possible that a user with a specific query may end up with a mismatch between what they want and what the system gives them.

There is thus some degree of unavoidable potential for inaccuracy (based on ambiguities of different sorts) inherent in this approach, but this need not be a particularly serious problem. The number of documents which would be retrieved for each compound nominal query term used would clearly depend on the length and complexity of the query term, and the recall and precision of the integrated system would be expected to be high. As explained in detail in Chapter Six, however, these measures of accuracy cannot be used for the COMMIX system operating at its current level, which concentrates on the isolated stage of producing the aboutness expression from the texts.

3.2.2.3 Ambiguity

This is the third, and final extensive field of study which will be mentioned in only minute detail here. Particular mention has been made above to the ambiguity which results from the compaction of larger amounts of information into denser expressions. We discuss here three types of ambiguity, noting the extent to which each might cause us problems in this work.

Word Class

The process which generates the aboutness expressions, which is discussed in detail in Chapter Four, involves looking up each open class word in a dictionary. In cases where there is ambiguity of word class, then, there are potential problems from the point of view of this system. One such occurrence is with the lexical item *won*, which occurs in the text as the past tense of the verb *win*. As we explain in Chapter Four, the default assumtion of word class 'noun' takes precedence over all others. This assumption, however, results in this item being labelled as a noun, in which sense it refers to the monetary unit (of North and South Korea). The occurrence of such errors need not necessarily be problematic for the system, and only has a lasting

detrimental effect either if linkage (which would have occurred had it been correctly labelled) then fails to occur, or if linkage occurs erroneously[13].

Although such occurrences of ambiguity of word class do occur, and may have a detrimental effect on the output from the system, the precise genre of abstracts, particularly those prepared by professional abstractors, means that such potential ambiguities are normally avoided by those producing the texts which the system takes as its input.

Word Sense

When a particular word is looked up in the dictionary, during the course of the processing of an individual text, it may be that there are multiple senses in which it could apply. As Chapter Four explains, however, the aim in looking up the glosses of words is to access the gloss words themselves, and search for overlap between the set of defining words for word *A* and that for word *B*. If the items *A* and *B* are related, it is expected that there will be larger overlap than if they are unrelated. Therefore, the set of gloss (definition) words for word *A* will include those relating to all senses of the word. So if this set contains words which do not overlap with those in the set for Word *B*, then it is of no consequence that the unmatched items relate to a sense other than that which is used in the text. It should be clear that although this could lead to overlap with the gloss word set from Word *B*, the likelihood is, in such short texts as those abstracts used, that it will not happen to such an extent as to interfere with the desired operation of the system.

Structure

Ambiguity of structure occurs when the boundaries between composite items within a larger structure are not clear. Beardon & Turner (1991) give some good examples, such as we see in the phrase *'...this approach to building user interfaces combines the...'*, in which any of the items *'building user'*, *'building user interfaces'*, *building user interfaces combines'*, *'user interfaces'*, *'user interfaces combines'*, or *'interfaces combines'* might be potential complex nominals (their term). Again, from the point of view of interpretation, such ambiguity can be problematic, and has consequently received much attention (e.g., Grishman, 1977; Marcus, 1980). From the point of view of generation, this type of ambiguity may lead to some mismatch between a user's

[13] The reader may need to refer to Chapter Four, which discusses the issue of linkage between semantically related items.

intended meaning and that associated with any matching items, but the advantages to be gained from adopting the approach suggested in this work would be expected to outweigh the disadvantages associated with this kind of problem.

Semantic Relation

The potential ambiguity which may result from semantic relations having become implicit in compound nominal expressions are referred to above. Again, the ambiguity which results from compacting information in the manner described above need only be problematic in situations where a user query might refer to one of the interpretations, whereas a document retrieved by means of its aboutness expressions matching the query may actually refer to a different interpretation. As mentioned above, the severity of such problems in a fully integrated IR environment would need to be assessed in the light of the advantages associated with the technique.

3.3 Nominal and Compact: a Useful Combination

The discussion of the features associated with compound nominals which has guided the course of this chapter has shown the compound nominal to be a highly useful construct in the expression of information in a compact way. We see remarkably high frequencies of these constructs in (English) newspapers, particularly in cases where space is clearly limited. Beardon (discussion) has found that their frequency of use, at least according to their occurrence in *The Times*, seems to be increasing. We do find that they crop up all over the place (as also noted by Lauer, 1996). Anyone who doubts the extent to which they are in current usage need only look at a daily newspaper (particularly the headings). A look at the *subject* lines which accompany email messages provides endless examples. One such recent example, which occurred as the subject heading accompanying a local email announcement, is *double card file grey metal free*, whose associated text advertised the free availability of *an old style grey metal card file.*

3.4 Summary

This chapter has discussed the features associated with compound nominals. It has concentrated on the major linguistic issues which are related to the features of nominality and compactness, and has shown the compound nominal to be an extremely useful means of compacting information, whether out of necessity or choice. The chapter has discussed a range of aspects related to these features which, although traditionally seen as posing problems for the analytical linguist, need not present as problematic in the context of this work.

Chapter Two has indicated a gap to be filled in the modes available for the indication of aboutness of a text. Chapter Three has discussed the expression of complex topics, and has shown the compound nominal to be an extremely useful construct to be used in their expression.

Chapter Four, which follows, comprises a detailed description of the implementation of the COMMIX system, which processes the text of input abstracts, and automatically generates compound nominal expressions to represent the aboutness of the texts. COMMIX is described both in terms of the general approach taken to the generation of aboutness expressions, and by reference to a detailed worked example.

Chapter Four

4. The COMMIX System: Principles and Implementation

This chapter describes the general approach adopted in the development and implementation of COMMIX, along with details of the technical environment and the resources used.

A number of assumptions have been made during the course of the development of COMMIX. These are divided into two sets: *global* and *local* assumptions. This distinction facilitates the differentiation of the former assumptions which apply in general, to the overall methodology, from local ones which are particular to this implementation. The global assumptions are discussed first, in Section 4.4, and the local assumptions are considered at the beginning of Section 4.5, which goes on to describe the methodology employed in the implementation of the system, using a worked example for clarity.

The validity of these assumptions is discussed at the appropriate stages, and the successes and shortcomings of different stages of the implementation are addressed in relation to these assumptions. It should be noted that throughout the development of COMMIX it has been an aim of the implementation to provide scope for future extension, both as exemplified in Chapter Five, and as dicussed at the end of Chapter Seven. The assumptions underlying the work have been stated with such possible future extensions in mind.

4.1 General Principles

One of the aims of the work presented here is to avoid the knowledge specification and representation problems associated with the use of template-driven methodologies, as discussed in Chapter Two. The position taken here is that the enormous amount of semantic and pragmatic information implicitly present within the definitions of a normal dictionary of contemporary language (in this case English), can be exploited in the processing of an existing piece of text to generate a linguistic representation of the salient concepts which are developed within the course of the text. An on-line dictionary thus constitutes, in effect, a rich knowledge

base which is task- and domain-independent. In exploiting a standard dictionary as our knowledge base, we therefore avoid having to restrict the domain to one which corresponds to a particular external domain knowledge base, and thus avoid the problem of having to specify and represent the world knowledge required by mainstream semantic methodologies.

As Chapter Two points out, this method has become known as *lexical overlap*, and is being used with greater frequency as the benefits of this type of shallow semantic method become clear. Some of the work which employs this method has been discussed in Chapter Two, and will not be reiterated here.

The general approach adopted here, then, is to exploit commonalities in the dictionary definitions of distinct terms to build up a network which links key nouns, their modifiers, and associated verbal information. Items which are found to be salient within the context of the text are investigated for their relatedness to other items in the text. Where such relatedness is identified, all the information pertaining to each member in the linked pair is identified and pooled, for expression as part of a larger modifying expression. The sections below elaborate and clarify this process, with particular issues discussed in terms of their status as either global or local assumptions.

4.1.1 Expectations in Relation to Hypothesis

Chapter One included a statement of the hypothesis relating to the expected performance of the COMMIX system, which was expressed as follows:

shallow semantic processing based on the method of lexical overlap, applied to the definitions of distinct terms occurring in an existing text, constitutes an effective means of generating novel compound nominal expressions to represent the aboutness of the text.

The general approach taken throughout this work has been to explore the extent to which this hypothesis is supported by the actual performance of COMMIX, using the shortcomings of the hypothesis to suggest improvements to the system. Some of these improvements will be implementable using information within the dictionary listings (or facilities offered by WordNet), but additional knowledge may be required. The final section of this chapter discusses the improvements which are included in the final 'basic' level of operation of COMMIX, whereas longer-term improvements are discussed in Chapter Seven.

4.2 Technical Environment

The COMMIX system has been developed under Unix, using POP11 in POPLOG as the implementation language. It was the original intention to utilise the on-line version of a standard dictionary of contemporary language, such as LDOCE as the lexical resource. However, for a number of practical and economic reasons, the resource which has been used is WordNet (Miller, 1990). This does limit the richness of individual glosses obtained, since WordNet's *synsets*-(synonym sets) are not full definitions, and its use thus reduces the amount of word overlap ('linkage') identified. It does, however, allow for the possible future utilisation of the additional facilities it provides.

4.2.1 The WordNet Database

WordNet is an online lexical reference database, in which word meanings for nouns, verbs and adjectives are represented by synonym sets, i.e., lists of synonymous word forms that are interchangeable in some contexts. The database recognises both lexical relations (which hold between word forms) and semantic relations (which hold between word meanings).

The reader should bear in mind that the original intention was to utilise an on-line version of a standard dictionary, which would give standard definitions of terms. Since this was not possible, WordNet has been the resource used, although it has been utilised in a manner which approximates the simple looking up of definitions in a standard dictionary.

4.3 Terminology

Throughout the discussion of the implementation of COMMIX, in Section 4.5, the following terminology is used:-

The term *lookup process* refers to the looking up of a term in WordNet, in order to obtain some information about it, such as its defining gloss.

The term *labelling* is used to refer to the process of looking up which results in the labelling of a term according to its syntactic word class.

The term *lookup level* refers to the number of lookup processes which have been required to yield a particular term. Thus all items occurring in the original abstract are designated L0 terms, since they simply appear in the original and are not

obtained via any looking up. When an L0 term is looked up in the dictionary, the items occurring in its defining gloss are designated L1 terms. Similarly, when these L1 items are looked up in turn, the items occurring in their defining glosses are designated L2 terms.

The term *linkage* refers to the presence of a direct match of terms, either in the original text of the abstract, or in the defining glosses, as described above. Linkage is described according to the levels of lookup involved: thus L0-L0 linkage refers to a direct match between two terms A and B, both of which occur in the original text; L0-L1 linkage occurs between terms A and B when these terms do not themselves match, but term A occurs in the defining gloss of term B (or vice versa). L1-L1 linkage occurs when two terms A and B do not match, but there is an item C, which is common to the defining glosses of both A and B. L1-L2 linkage occurs when two terms A and B in the original abstract do not match, and there is no term which is common to their defining glosses, but there is a term C which occurs in the (L1) defining gloss of A, which also occurs in an L2 lookup gloss of one of the L1 terms occurring in the defining gloss of B. Similarly, L2-L2 linkage occurs when there are two lookup stages involved for each of terms A and B in order to produce the match.

This linkage is seen to be representative of semantic relatedness between the two terms. It represents differing degrees of relatedness according to the level of lookup involved. It seems intuitively clear that items which match directly (with no lookup at all) may be assumed to be more closely semantically linked than items matched by virtue of common terms occurring in their definitions. Similarly, the greater the lookup level involved in identifying a link, the less can be assumed to be the semantic relatedness between the original items, since each successive level of lookup introduces a further step of removal from the terms used in the original text.

Given that WordNet was used instead of a standard on-line dictionary, we approximate the L1 and L2 lookup distinction in WordNet by utilising the => indicator. This indicator shows where, in effect, the next level up the semantic hierarchy is being represented. Note that the use of WordNet rather than, e.g., LDOCE (Procter, 1978) has two effects on this linkage. Firstly, the capacity for linkage is reduced, since the WordNet synsets lack the richness of the full definitions of a standard dictionary. Secondly, the looking up of successive linkage levels (which constitutes the equivalent of the method of *lexical chaining* (Morris and Hirst, 1991)) is

greatly facilitated. As mentioned above, however, the richness of definitions suffers as a result.

4.4 Assumptions

There are two sets of assumptions that have been adopted in this work. Firstly, some basic, constant, *global* assumptions which underly the whole approach, and secondly, a set of transient assumptions which are associated with the initial implementation. The global assumptions, which apply to the system as a whole, apply at all stages of implementation. Assumptions which apply specifically to a particular stage of implementation are referred to as *local* assumptions. Whilst the global assumptions are held to be valid, these local assumptions are not necessarily believed to hold in a generalised sense, but their formal statement as assumptions taken at each stage forces their explicit expression. This not only provides a forum for the discussion of relevant aspects of the implementation, but also helps overcome the tendency for such assumptions to fall into the realm of the implicit and end up being taken for granted. The local assumptions pertaining to each stage of the implementation are discussed at the beginning of the relevant sections, and the validity of each is discussed in detail in Section 4.5.3.

4.4.1 Global Assumptions

The following assumptions are general ones which are held to apply regardless of the particulars of the implementation, which would therefore also be expected to underly any future extensions to this work. Some of the notions expressed in them have been discussed in detail in Chapter Three. It should be clear that assumptions iii) and iv) relate specifically to the hypothesis which drives this work, and which is presented in Chapter One (and, for convenience, repeated above), and assessed in Chapter Seven.

Assumption (i)

In some pieces of extended text (typified by abstracts), there are conceptual 'entities' which the text is about, which are developed throughout the course of that text, but which are not necessarily accorded linguistic *labels* overtly within the text.

Assumption (ii)

Compound nominals offer a linguistic means of compacting large amounts of information into premodified nominal form, so that complex conceptual entities may be expressed. For any phrase which refers either explicitly or implicitly to some such complex referent (conceptual entity), it is possible to construct a compound nominal expression as its linguistic representation. As discussed in Chapter Three, the compaction of information may be optional or necessary, and may be conceptually or editorially motivated, or related to a particular style (such as expository text).

Assumption (iii)

Compound nominal terms representing aboutness concepts can be automatically generated by identifying the most salient nouns in the original abstract and applying to those nouns (or their hypernyms, or synonyms) some or all of the modifiers which apply to semantically related nouns, (including noun forms of associated verbs).

Assumption (iv)

Semantic relatedness can be identified by virtue either of direct matching of words[14], or of a match between a word appearing in the text and one which appears in the definition of another word. Relatedness between two items is also held to be signalled by direct matching of words appearing in the definitions of each item.

Assumption (v)

The salience of a term (in regard to the aboutness of the overall text) is directly related to its frequency of occurrence within a piece of text. To some extent the salience of an item is also reflected in its position in the text. This is specifically true for abstracts occurring in the database used here, since one of the specific constraints

[14]Ambiguity of word sense tends not to be as problematic as one might think, at least in the genre of human-produced abstracts (since these are short, concise texts). Abstractors rarely create an ambiguity of this sort, and would normally use a synonym or paraphrase instead of a lexically ambiguous term.

on the abstractors is to make the first sentence as indicative as possible of the overall content of the article.

In this work, which is directly concerned with the genre of abstracts, it is also assumed that any item which occurs in the text in modified form is salient. Therefore, in this work we assume that the modification of an item reflects there being something notable about the item, in the context in which it occurs. This assumption is utilised here to enrich our indicators of salience, and we say, therefore, that modification of an item in the text is also an indication of the particular salience of that item.

Assumption (vi)

Any verbal information (except that associated with modal or auxilliary verbs) can be expressed and utilised in nominal form in the formulation of compound nominal terms which express the aboutness of a piece of text.

4.4.2 Establishing Semantic Relatedness by Applying Global Assumptions

The assumptions specified above, in conjunction with the general approach, can be expressed as a set of axioms which may be applied to any abstract. Thus, given an abstract in which nouns N1 and N2 appear, with respective modifiers M1 and M1, the application of the above assumptions leads to the following assertions:

a) If N1 = N2 (= N), then given a piece of text in which M1N1 and M2N2 both feature, a compound nominal term M1M2N can be constructed and deemed at least partly representative of the aboutness of that text. In this case the relation N1:N2 is the highest possible (because they are the same noun[15]), and may be referred to as R1.

b) If N1 and N2 are different nouns, but N2 appears in the definition of N1, then they can be assumed to be semantically related (to a weaker degree than in a) and on that basis, there is justification for applying at least some of the modifiers of N2 to N1 in the construction of compound nominal aboutness expressions.

[15]See previous footnote (number 14) relating to ambiguity.

c) If N1 and N2 are different nouns, but there is a word, W, which is common to their definitions, then they can be assumed to be semantically related, although to a weaker degree than in a) and b). In this case there is justification for applying the modifiers of both nouns to each noun independently.

The degree of relatedness between N1 and N2 varies according to the number of words which are common to their definitions: the more terms which the definitions of two terms have in common, the greater is assumed to be their semantic relatedness.

4.4.3 Lexical Overlap Method as Used for Word Sense Disambiguation

The method of establishing semantic relatedness by word overlap in dictionary definitions has been used elsewhere, particularly in the field of word sense disambiguation. This is a fast-developing field, classically exhibiting a range of methodologies and evaluation criteria (as discussed in Resnik and Yarowsky, 1997). Although a full and representative discussion of the entire field is beyond the scope of this thesis, it is relevant to mention of the work which is based on matching items using preexisting lexical resources.

There are a number of authors in this field whose work utilises the assumption that concepts which are related to one another will be represented in a dictionary or thesaurus by the same words[16]. Some early examples of this approach are mentioned by Lesk (1986), such as Masterman (1957), who suggested looking up the thesaurus entries for different senses of a word, and counting the words which overlapped with those for adjacent words in the text. Lesk (1986, p.3) also mentions unpublished work done by Lawrence Urdang in the 1960s, who used a dictionary rather than a thesaurus for the same purpose. Lesk (ibid.) describes his own implementation of dictionary disambiguation, using a machine-readable dictionary, but based on the same principle: that word senses which are related to each other will have some of the same words appearing in their definitions.

More recent lexical disambiguation work has utilised various preexisting lexical resources. Yarowsky (1992) used categories of Roget's thesaurus to disambiguate polysemous words, based on statistical analysis of word occurrences from different

[16] Some of this work was mentioned in Chapter Two.

categories. Wilks et al (1993) used LDOCE, and a number of authors have used WordNet (e.g., Sussna, 1993; Richardson et al., 1994; Resnik, 1995). Luk (1995) specified the status of the controlled vocabulary of LDOCE as corresponding to 'defining concepts', and based his disambiguation heuristic on the cooccurrence of these concepts, rather than specific words. Similarly, Agirre and Rigau (1996) concentrate on resolving noun ambiguity by selecting combinations of senses (of contiguous nouns) according to the combinations that maximise the 'conceptual density' (Agirre and Rau, 1995) of the conceptual networks constructed. Although these latter works are based on the extension of word matching to 'concepts' (i.e., words higher up in a semantic hierarchy), the basic intuition of utilising the definitions in a preexisting lexical resource to provide the matching still applies.

Recently, Rigau et al (1997) combine a variety of methods in their disambiguation system, including lexical overlap. Their "Heuristic 4: Word Matching" is used to count the total number of content words shared between a given hyponym and a candidate hypernym, and is once again based on this same principle. In their own words: "*This heuristic trusts that related concepts will be expressed using the same content words.*" (ibid.)

There are clearly differences between the word sense disambiguation work and the work described in this paper, in terms of their respective aims. Whereas the former utilises adjacent words and often incorporates co-occurrence, the COMMIX system identifies relatedness between words which do not necessarily appear together, even in the same sentence[17]. This is a crucial difference between the disambiguation work and the approach as used in COMMIX, which does not restrict matching to the sentence level. However, the basic assumption, that semantic relatedness between distinct items can be indicated (to some extent) by an overlap in the words appearing in their glosses, is the same.

[17] It is important to remember that the COMMIX system and its underlying assumptions relate to the realm of the abstract.

4.4.4 A Small-scale Empirical Investigation of Lexical Overlap

In order to give an indication of the effectiveness of the word overlap method of establishing semantic relatedness, a small-scale word-pair analysis of overlapping words has been carried out, in relation to Text 1 of the data set used in the development of COMMIX.

For this small-scale investigation into the overlap of gloss words from distinct words, Text 1 of the data set, which appears in Appendix 1, was used. The overlap investigation was carried out on two sets of nouns: the first set comprised only those nouns which count as salient in the methodology employed by COMMIX in its basic form (this notion being based solely on its having a frequency within the text of greater than one). The second set includes other nouns which are also considered salient, as indicated by the fact that they are already modified in the input text, even though they only occur once.

Method

The method used to investigate the word pair overlaps was as follows:

For each set of salient nouns (the assessment of which is explained in detail in section 4.5.2), all possible word-pair combinations were identified. For both members of each word pair, the individual (WordNet) entries were looked up, and the number of words common to both entries were recorded.

Results

According to the methodology described in this chapter, for the system in its most basic form, there are 7 salient nouns: games; toy; market; worth; million; pounds_sterling; sales. Allowing for the commutativity of pairing, this gives 21 possible word pair combinations.

For the extended notion of salience, there are 3 additional nouns: industry; entertainment; software. In combination with the nouns occurring in the first set, this yields a further 24 possible word pairs.

Tables 4.1 and 4.2 show the results of the search for overlap between entries for each of the word pair combinations, where n is the number of open class words which the entries have in common.

Word Pair	n
games - toy	2
games - market	8
games - worth	0
games - million	1
games - pounds_sterling	0
games - sales	3
toy - market	0
toy - worth	1
toy - million	1
toy - pounds_sterling	0
toy - sales	4
market - worth	1
market - million	0
market - pounds_sterling	0
market - sales	3
worth - million	4
worth - pounds_sterling	0
worth - sales	2
million - pounds_sterling	0
million - sales	2
pounds_sterling - sales	0

Table 4.1. Number of Words Common to Definitions of Nouns occuring more than once in Text 1

Table 4.1, then, shows the frequencies of overlap using the frequency-only notion of salience.

Table 4.2 shows the overlap frequencies including other nouns which are also considered salient, as indicated by the fact that they are already modified in the input text, even though they may only occur once.

Word Pair	n
industry - games	2
industry - toy	1
industry - market	11
industry - worth	0
industry - million	0
industry - pounds_sterling	0
industry - sales	4
entertainment - games	7
entertainment - toy	0
entertainment - market	1
entertainment - worth	0
entertainment - million	0
entertainment - £_sterling	0
entertainment - sales	1
software - games	1
software - toy	1
software - market	1
software - worth	1
software - million	1
software - £_sterling	0
software - sales	1
industry - entertainment	1
industry - software	0
entertainment - software	0

Table 4.2. Number of Common Words for Additional Word Pair Overlaps (Arising from Including Modified Nouns)

Indication

In such a small scale investigation it would be over-presumptive to try to form any firm conclusions, but the results do give an indication of the reliability of the assumption that overlap can be indicative of semantic relatedness.

In the sample of word pairs examined, there seem to be 3 main groupings: those with 0, 1 or 2 words in common; those with 7 - 11 words; and those somewhere in between. There are 3 examples which fall into the top group: namely *market - industry; entertainment - games;* and *games - market.* Subjectively judged, we can say that the first two of these pairs are strongly related semantically, and that there are degrees of similarity in the third pair.

Occurring in the middle range group (those with between 3 and 6 overlapping words), there are 5 word pairs: *industry - sales; games - sales; toy - sales; market - sales;* and *worth - million.* Whilst we might expect to see more overlap between, for example, *market* and *sales* (with an overlap of 3) than between *games* and *market* (overlap of 8), it is clear that the words in this mid-range are at least partially related.

The remainder of the word pairs (37 of them) exhibit low numbers of common words (see the tables above for all the examples). In general, these pairings involve words which are, indeed, not related, or are only slightly related to one another.

In some of the examples, word pairs which are related (as judged by human experience) can end up with a low overlap, or even no overlap at all. This is unavoidable to some extent, but the problem is exacerbated by the cultural difference between American and British usage of English. Of particular importance in terms of our example abstract is the failure of WordNet to use the word *money* in its definition of *pounds_sterling*. This clearly affects all pairings in which *pounds_sterling* is involved, and we see, for example, no overlap between *pounds_sterling* and *sales,* which most British people would judge as being fairly highly related. It would be expected that this problem could be overcome by using a more standard dictionary.

With this small sample size there would be little point in applying any test for statistical significance to the results, since a much larger sample would be required for meaningful results. Such an extended study is beyond the scope of the current work, but it is possible to get some indication of tendencies. This indication is that

highly related words do have a larger number of words in common to their definitions, and unrelated words do have lower numbers, with partially related words having rather more variable results.

It comes as no surprise that there are pairs of words that are unrelated, yet contain common words in their definitions. What is indicated by this small investigation is that when words are related (such as *market* and *industry*), the number of common words is much greater. Whereas unrelated words may have an overlap of 1 or 2, or even a few more words, the indication is that they do not have more than about 5 words in common.

The initial indication, then, is that the assumtion of semantic relatedness being proportional to overlap of definition words is justified, at least to some extent. Athough this has been taken as a specific assumption in this work, in the manner of many authors working on word sense disambiguation (as discused above), it is still interesting to note this indication, which could perhaps serve as a pilot study for a more detailed analysis (which is, however, beyond the original scope of this work).

4.5 COMMIX Implementation

The main aim of this implementation has been the successful processing of an input abstract to generate some useful output in the form of compound nominal terms which represent the aboutness of the abstract. The extent to which the generated expressions fulfil this aim is the subject of Chapter Six.

The input to the system is an unadulterated[18] abstract, written by an abstractor for a piece of text appearing in a journal or a newspaper. The output comprises compound nominal expressions which represent the aboutness of the input text. The program was developed according to the general principles and assumptions described in preceding sections of this chapter. In addition to the global, underlying assumptions detailed in Section 4.4, the following local assumptions apply to this implementation. These provide the basis of some improvements made to this initial implementation, which is discussed in Section 4.6.

[18] The only exceptions to this are apostrophes and semicolons, which are included in POP11 syntax, and corrupt the input stream. These are therefore the only things to be removed from the input (by hand).

4.5.1 Local Assumptions Relating to Implementation

There follows a list of the local assumptions which were taken in the development of this system. These are not all necessarily held to be true, and some of them are clear over-simplifications, but their statement as formal assumptions provides a forum for discussion of the shortcomings, and aids a clearer initial discussion of the performance (although Chapter Six comprises a full evaluation of the performance of the system over the complete text set). This initial look at the performance appears in Section 4.6, following the detailed description of this implementation in relation to the example text.

Asumption 1

For the purposes of this initial implementation an overly simplistic definition of compound nominals has been adopted. Thus, any noun terminated stream of consecutive items which are labelled as 'noun', 'number', 'name', 'verb', 'adjective' or 'adverb' is initially identified as a compound nominal. The head of the expression is simply taken to be the last item, which is a noun. (For a much fuller discussion, see Chapter Three).

Assumption 2

Some representative compound nominal expressions can be generated by basing the assessment of salience of items solely on those items which occur in modified expressions: in other words, if a noun occurs in an unmodified form, this occurrence does not count towards the assessment of salience of that item.

Note that this assumption is not necessarily held to be true, although it is tested here to see the extent to which this restrictive notion of salience produces useful output.

Assumption 3

It is not necessary to resolve the reference of pronouns and proper names in order to produce some meaningful linkage and consequently, some useful output.
Therefore:-

a) pronouns are deleted as closed class items;

b) proper names are simply labelled as such. An item is assumed to be a proper name if it begins with an upper case letter and occurs within a sentence (i.e., non-sentence initially).

Assumption 4

Sentence-initial capital letters are assumed to be upper case merely by virtue of their position in the sentence, and therefore are replaced by lower case letters.

Assumption 5

Restricting the level of lookup to 1 is sufficient to indicate some linkage between semantically related items.

Assumption 6

The modifiers within a compound nominal expression (some of which may be compound nominals themselves) may be placed in any order.

Assumption 7

At the labelling stage, if there is an ambiguity of word class involving the option *noun*, then this is taken as the default assumption.

Assumption 8

A single output expression is sufficient to give some indication of the aboutness of the abstract.

Assumption 9

Where compound nominals (or at least two consecutive nouns) exist in the text, then it is sufficient to look up the individual words in order to extract the semantic information necessary to establish semantic relatedness between items. For example, in the stream *video games industry* it is sufficient to consider the lexemes *video, games* and *industry* as distinct from and independent of one another. Similarly, in the case of the stream *compact disc market* the independent treatment of the words *compact, disc* and *market* will suffice in establishing relevant semantic linkage.

The COMMIX system was initially developed in accordance with the local assumptions specified above, and its performance at this level is discussed in Section 4.5.3 below. The full implementation of the system, which is evaluated in Chapter Six, and discussed in Chapter Seven, is slightly more elaborate, in that it takes into account the findings related to these discussions. For the purposes of describing the methodology, however, it is far clearer to follow a worked example using the more limited version, which has one output expression, and follows just one level of

linkage. The basic methodology is the same for the fully developed implementation, but would be unnecessarily complicated to describe in full.

4.5.2 Description of the Implementation: Stage 1

The 'Stage 1' here refers to the final comments appearing above. This section follows through a worked example, and describes the methodology used in COMMIX. The example used is Text 1 from the data set, which, for convenience, is shown again here[19]:

> *The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling. The main competitors in the market are Sega and Nintendo. Nintendo will spend 15 million pounds sterling on advertising over Oct-Dec 1992.*

The processing of an abstract falls into the following stages:-

1) Sorting and labelling of the original abstract

This involves the deletion of all closed class words and the labelling of all open class words according to their syntactic word class (as listed in the WordNet [20] database). The morphology-handling facility within WordNet has the advantage that different morphological forms of the same word are recognised as the base lemmas, which greatly facilitates the labelling and subsequent matching of items. If an item is not listed in the database, it is retained and labelled '*n_I*'. Full labelling of the abstract involves numbering items according to both sentence number and word number, and leaving a trace of deleted items, indicating the sort of item that has been deleted. Thus, for the example abstract, the fully labelled version resulting from this sorting and labelling stage comprises constituents of the form:

---

[19]The original text to which this abstract refers appeared in *The Observer* of Sunday 11th October 1992.
[20]The process by which the syntactic class of an item is obtained from the dictionary is referred to as **labelling**. This clearly involves looking up the item, but the term **lookup** is avoided, since this is reserved for the procedure which obtains the defining gloss of an item.

[sentence number word number item syntactic class of item]

(where *item* is the item itself, or DEL if it has been deleted as a closed class word), as follows:

[1 1 DEL det 1 2 video noun 1 3 games noun 1 4 industry noun 1 5 DEL aux 1 6 growing adj 1 7 fast noun 1 8 DEL conj 1 9 DEL modal 1 10 dominate verb 1 11 DEL det 1 12 toy noun 1 13 market noun 1 14 DEL conj 1 15 become verb 1 16 DEL det 1 17 established adj 1 18 part noun 1 19 DEL prep 1 20 home noun 1 21 entertainment noun 1 21 . punc_s 2 22 DEL det 2 23 1991 number 2 24 computer noun 2 25 games noun 2 26 market noun 2 27 DEL aux 2 28 worth noun 2 29 275 number 2 30 million noun 2 31 pounds noun 2 32 sterling adj 2 33 growing adj 2 34 DEL prep 2 35 500 number 2 36 million noun 2 37 DEL prep 2 38 1992 number 2 38 DEL punc 2 39 half noun 2 40 DEL det 2 41 toy noun 2 42 market noun 2 42 . punc_s 3 43 hardware noun 3 44 sales noun 3 45 DEL modal 3 46 rise noun 3 47 DEL prep 3 48 261 number 3 49 DEL prep 3 50 635 number 3 51 million noun 3 52 pounds noun 3 53 sterling adj 3 54 DEL prep 3 55 1994 number 3 55 . punc_s 4 56 associated adj 4 57 software noun 4 58 sales noun 4 59 DEL aux 4 60 forecast noun 4 61 DEL prep 4 62 645 number 4 63 million noun 4 64 pounds noun 4 65 sterling adj 4 66 DEL prep 4 67 1993 number 4 67 . punc_s 5 68 DEL det 5 69 compact_disc n_1 5 70 market noun 5 71 DEL aux 5 72 worth noun 5 73 345 number 5 74 million noun 5 75 pounds noun 5 76 sterling adj 5 76 . punc_s 6 77 DEL det 6 78 main noun 6 79 competitors noun 6 80 DEL prep 6 81 DEL det 6 82 market noun 6 83 DEL aux 6 84 Sega name 6 85 DEL conj 6 86 Nintendo name 6 86 . punc_s 7 87 nintendo n_1 7 88 DEL modal 7 89 spend verb 7 90 15 number 7 9 1 million noun 7 92 pounds noun 7 93 sterling adj 7 94 DEL prep 7 95 advertising noun 7 96 DEL prep 7 97 Oct name 7 97 DEL punc 7 98 Dec name 7 99 1992 number 7 99 . punc_s]

The detailed labelling of the abstract shown here is not utilised to the full at this stage of the implementation, but could be utilised at a later stage when constraints (involving sentence number, the position and type of punctuation, the type and position of deleted items) might be introduced. At this stage it is only the syntactic word class information which is utilised.

2) Pre-lookup (pre-linkage) processing of labelled abstract

This stage involves two processes: firstly, the identification of compound nominals which already exist in the abstract; secondly, the identification of those heads of existing compound nominals which occur more than once in the original abstract. The first stage yields the following in our example:

> *[[video games industry] [growing fast] [toy market] [established part] [home entertainment] [1991 computer games market] [worth 275 million pounds sterling growing] [500 million] [toy market] [hardware sales] [635 million pounds sterling] [associated software sales] [645 million pounds sterling] [worth 345 million pounds sterling] [main competitors] [spend 15 million pounds sterling] [Dec 1992]]*

The result of the second stage is a list of the salient heads (ie those which occur more than once), alongside their frequency of occurrence in the list of existing compound nominals. The result of this second stage in our example is the following list:

[market 3 sales 2 sterling 4]

3) The Lookup Process

This is the point at which the defining gloss for each salient head is looked up. The gloss is converted from an *ascii* stream of characters into a list of words, which enables the further processing and matching of terms in the gloss. The processing of each gloss involves the filtering of words which do not pertain to the definition (such as *sense*), and the deletion of all closed class words, followed by the labelling of all open class words, as in stage 1. This stage takes as input, the list of salient heads as shown above (containing the items *market, sales* and *sterling*) and produces, for each salient head, a gloss from which all the closed class words have been removed. The glosses for the salient heads thus produced and processed appear as follows (where *GLOSS FOR* is followed by the term, and then by the open class words appearing in its definition):

> *[[GLOSS FOR sterling greatest highest sterling prenominal super superlative highest degree quality superior vs inferior characteristic high rank importance] [GLOSS FOR sales trade sales desperate boost sales commercial enterprise business enterprise business purchase*

sale goods services senses sale sale cut sale sales event selling specially reduced prices sale reduce inventory selling merchandising marketing exchange goods agreed sum money sale particular instance selling made three sales hour selling merchandising marketing exchange goods agreed sum money] [GLOSS FOR market market customers particular product service class social class people sharing common attribute market securities industry securities markets aggregate market always frustates small investor industry people engaged particular kind commercial enterprise grocery store grocery market supermarket marketplace mart market marketplace commerical21 activity whereby good services exchanged competition there market activity behavior specific action pursuit avoided recreational activity senses market market deal market transact deal business market sale produce sale trade deal merchandise engage trade commercialize market make commercial change alter cause change make different]]

The glosses can then be compared against one another in the search for matching items, which will justify the formation of a link between their respective parents.

4) Matching Terms

This stage involves searching for links between the glosses of the different salient nouns, with the aim of finding terms which are common to more than one gloss. As mentioned above, a link between two parent items is held to be justified if their respective glosses contain common terms. The identification of a link provides justification for assuming the parent heads to be semantically related. Matching of a term in the gloss of one L0 parent with a term in the gloss of another L0 parent yields an L1-L1 link, since the glosses involved in the matching process are obtained as a result of 1 lookup process for each original parent term.

In the example being considered, the input at this stage is the list of glosses as shown in 2) above. There are three gloss lists involved, so all possible pairs of gloss lists are compared one with another, and items which appear in both lists of a pair are recorded as common, linking terms.

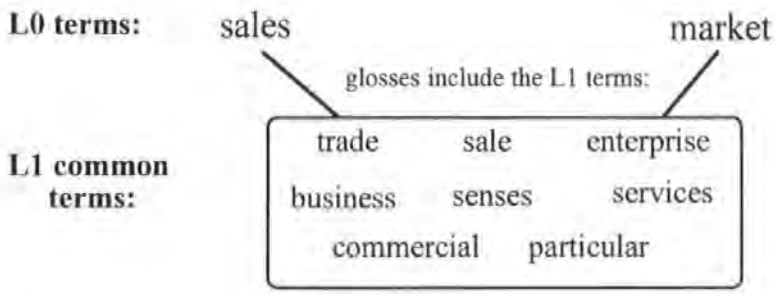
---

[21] The spelling mistake here originates in WordNet itself, and is an example of how errors can occur in such resources.

In our example, there are no linking terms occurring in the gloss for *sterling*[22] (i.e., there are no terms which appear in the gloss for *sterling* as well as the gloss for either *sales* or *market*). There are, however, eight terms which are common to the glosses of the L0 terms *sales* and *market*, and these matched items are the output from this stage, as follows:-

[trade commercial enterprise business sale services senses particular]

These terms which occur in the gloss of more than one salient head are the justification by which the parent items to whose glosses they are common (i.e., *sales* and *market* here) are held to be semantically related. As such, these common terms constitute the L1-L1 links (i.e., between two glosses each pursued to the 1st level of lookup) which link the parent L0 terms *sales* and *market*, as shown below:-



The L1 common terms (L1L1 links) are held to justify the assumption of semantic linkage between the two L0 terms *sales* and *market*, and as such are the justification for applying all the modifiers of both L0 terms to the most salient head term (i.e. that which occurs with the highest frequency). This *cross-application* of modifiers is described in stage 6) below.

5) Weighting the links

Once all the commonalities between the L1 glosses of the salient heads have been identified, the frequency of each matched term (i.e. link) throughout the glosses is recorded. This frequency was intended to be used subsequently in the establishment of a salience weighting value which reflects the strength of the justification for the linkage. However, as Chapter Six shows, and Chapter Seven discusses, the fact that

[22]This is counter to expectations, but results from the version of WordNet used at the time listing *sterling* only as an adjective. Note that version 1.5, used in the final implementation has this problem resolved.

some aboutness expressions exhibit greater amounts of linkage during their formation than others, does not necessarily mean that they end up being judged as more representative of the aboutness of the text.

As discussed in Chapter Seven, it is likely that this information regarding the amount of linkage between different items (as listed in Appendix 4) would be useful, in conjuction with a much more extensive treatment of the weighting issue, which would be expected to include aspects relating to the subject of text genre[23].

6) Selection of head noun and cross-application of modifiers

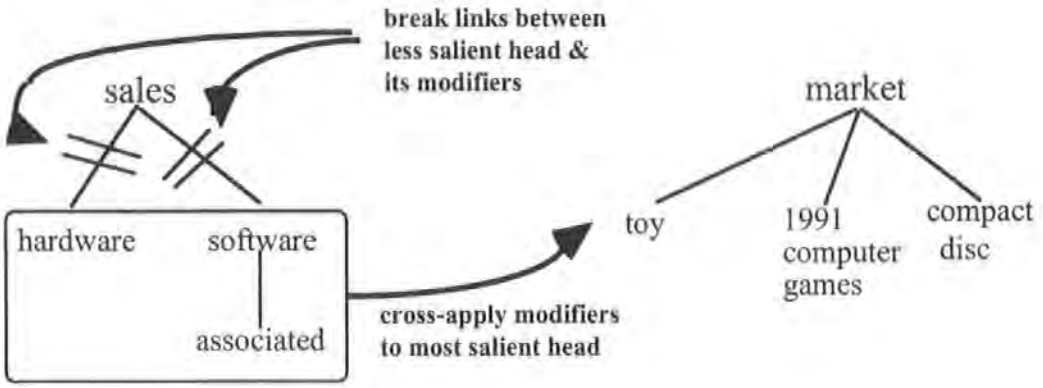
If linkage has been found between two or more items which occur as salient heads in the abstract, then there is justification for combining their modifiers and applying them all to one of the head nouns. The noun selected to be the head of the expression so formed is the one which is involved in L1-L1 linkage and occurs with the highest frequency in the original abstract. This stage identifies the appropriate noun and applies to it all the modifiers which relate either to that noun or to nouns which are semantically linked to it (via commonalities in their L1 glosses, as indicated by a link having been identified).

Thus, based on the above justification for assuming semantic relatedness between the terms *sales* and *market* (as shown in point 4 above), the next stage is to identify all the modifiers applying to each of these words in the original text of the abstract. These can be represented by the following tree structures:-

The frequencies of occurrence of each of *sales* and *market* are compared, and the most salient head (based on frequency of occurrence) turns out to be *market*. Once *market* has been selected as the most salient head, and therefore as the one to be used in the

[23] In other words, we might well expect that different text types would require different contraints governing the weightings accorded to different aboutness terms. This could be the basis for an interesting future study.

output compound nominal expression, the 'links' between the other, less frequent head (i.e., *sales* in the example) and its modifiers are broken, and the modifiers are cross-applied to join the existing ones which apply to *market*.



Thus, all the modifiers of both *sales* and *market* are applied to the most frequently occurring of *sales* and *market*, which is *market*. The final output expression produced at this stage is:

hardware associated software toy 1991 computer games [24] *market*

4.5.3 Performance of this Stage of Implementation

Although the output term does not conform completely to notions of syntactic well-formedness, it does provide a surprising degree of insight into the aboutness of the example abstract. Considering the simplicity of this first implementation, this is encouraging with regard to the possibilities for tuning the technique at a later stage.

It should be noted, however, that out of a sample of ten abstracts, the example abstract was the only one to yield any output terms. As the discussion below reveals, this is partly due to the overly restrictive notion of salience, which needs to be extended to include both premodification and frequency of occurrence of items in the text. A further aspect of this limitation concerns the level at which linkage is sought

[24] We would expect the expression to contain the term *compact_disc* as a modifier at this point, but this term was not listed in the earlier version of WordNet used at this initial stage. It has therefore become eliminated as 'not listed' at an early stage of the processing.

being restricted to the L1 level. The issue of pronoun and name resolution is also involved, and this again is discussed below.

These shortcomings are assessed below, in the context of the overall performance of this initial stage of the implementation. This assessment, which motivates the suggested improvements, refers to the local assumptions pertaining to this implementation, as specified in Section 4.5.1.

4.5.3.1 Improvements Based on Assessment of Performance and Validity of Local Assumptions

This section discusses the performance of the system, relating comments to the local assumptions specified above. It details the improvements which have been added to this initial stage to form the final state of the implementation, to which state the evaluation and final comments in Chapters Six and Seven refer.

The first stage of the implementation has met with some success, in that it generates an expression which does give an indication of the aboutness of the abstract. However, it should be clear that there are several improvements to be made to the implementation described so far, and these are described below, again, in relation to the specific local assumptions which highlighted their need.

Assumption 1

Any streams of consecutive items which are labelled as 'noun', 'number', 'name', 'verb', 'adjective' or 'adverb' are initially identified as compound nominals. The head of the expression is simply taken to be the last item in such a stream.

Validity of Assumption 1

The list of compound nominals already existing in the example abstract (according to this oversimplified definition) is as follows:

> *[[video games industry] [growing fast] [toy market] [established part] [home entertainment] [1991 computer games market] [worth 275 million pounds sterling growing] [500 million] [toy market] [hardware sales] [635 million pounds sterling] [associated software sales] [645 million pounds sterling] [worth 345 million pounds sterling] [main competitors] [spend 15 million pounds sterling] [Dec 1992]]*

There is an obvious shortcoming here, by which it becomes possible for items which are not nouns to be selected as head 'nouns', simply by virtue of their position as the final term in an uninterrupted string of items falling into the above classes. Thus, we find the following anomalies classed as compound nominals:

[worth 275 million pounds sterling growing], [635 million pound sterling],
[645 million pounds sterling], [worth 345 million pounds sterling],
[spend 15 million pounds sterling]

in which all the terminating items have been classed as adjectives.

Improvement: Restriction of Class of Head

In the light of the above finding (and bearing in mind our description of what constitutes a compound nominal, in Chapter Three), the final implementation restricts the class of items which can occur as heads of the final aboutness expressions: the restriction allows only nouns and nominalised verbs.

It is worth noting here that the *final* implementation *does* allow each of these items (i.e., *sterling* and *growing*) to occur *legally* as heads of the expressions. There are two reasons for this: firstly, the newer version of WordNet has incorporated the nominal sense of 'sterling' (as the British monetary unit), and the default assumption of nominality over all other word class assignments ensures that 'sterling' is labelled as a noun rather than an adjective. The second reason concerns the refinement relating to the manner in which the nominalisation of verbs is dealt with, and this aspect is discussed next.

Improvement: Nominalisation of Verbal Information

At the time of final implementation of COMMIX there was no electronic facility which listed noun forms of verbs. Chapter Three discusses at length the issues involved in the nominalisation of verbs, and indicates the complications involved in trying to automate the process.

In the light of this, a simplistic, but nevertheless useful approach is taken to this problem: namely, to present verbs in the gerundive forms. In our abstract this would yield the changes *dominate -> dominating* (rather than the preferred *domination*), *grow -> growing*, and *become -> becoming* (which could be eliminated as a non-crucial verb).

Where cases of ambiguity of word class occur (which include 'noun' as one of the possible interpretations), the default assumption is that the item is a noun. This results in the gerund form *growing*, which, as a nominalised verb form, is treated as a noun rather than an adjective, therefore constituting an allowable head of a compound nominal expression.

One further point is worth noting here in relation to the performance of the initial stage of the implementation. From a theoretical perspective, the 'compound nominal' *[growing fast]* is not an anomaly, since the syntactic word class recorded for the lexically ambiguous item 'fast' is 'noun'. In addition (as noted above), the version of WordNet used at the time this work was done, only listed the term *sterling* as an adjective, and did not recognise its existence as a noun (denoting British currency). If it were listed as a noun, those compound nominals ending with the term *sterling* would be perfectly acceptable in respect of being noun-terminated[25].

Assumption 2

Some representative compound nominal expressions can be generated by basing the assessment of salience of items solely on those items which occur in modified expressions. That is, if a noun occurs in an unmodified form, this occurrence does not count towards the assessment of salience of that item.

Validity of Assumption 2

In the example abstract shown above, this assumption is not overly restrictive, since the most salient head, *market* (as judged by modification), also happens to be the most frequent item, even when its unmodified occurrences are ignored. However, the effect on some of the other abstracts tested in this initial phase is highly restrictive in that items which do occur more than once are not counted. This, along with other restrictions on the search space (resulting from the restriction of linkage) results in a failure of the system to provide any output expressions at all. It is therefore clear at this stage that adopting this assumption even in this initial implementation is overly restrictive, and that the frequency of items even in unmodified form does indeed need to be included as an additional criterion of salience.

---

[25] Note that version 1.5, used in the final implementation has this problem resolved.

The next section shows the amount of linkage which results when the salience of items is based a) on premodification only (the restricted notion of salience), and b) on frequency of occurrence alone (the unrestricted notion of salience).

A Fuller Discussion of Salience

The initial implementation required that for a noun to be assessed for salience in the first instance, it must be premodified in the original abstract. The resulting, restricted list of salient nouns for Text 1 was thus *[market 3 sales 2 sterling 4]*, where the number following a noun is its frequency of occurrence in the abstract.

By contrast, the use of frequency of occurrence as the basis for establishing salience produced the corresponding list *[games 2 toy 2 market 5 worth 2 million 6 pounds 5 sterling 5 sales 2]*, in which all nouns occurring more than once appear along with their frequencies, regardless of whether or not they occur in modified form in the text. Figures 4.1 and 4.2 respectively show the difference in the amount of linkage resulting after the L1 lookup stage, for the restricted and unrestricted assessment of salient nouns respectively.

Figure 4.1 refers to the case where the salient nouns are *market, sales* and *sterling*. It represents the linkage obtained by consulting glosses for each salient noun, and recording words which are common to their glosses. In this and the following linkage diagram (Figure 4.2), the salient nouns are represented as unboxed terms, which are linked by virtue of there being words common to their glosses: these common, linking 'intermediaries' are shown in boxes in the figures. Curved lines represent semantic relatedness (or linkage) as justified by the commonalities in the glosses for the salient nouns (i.e., by the intersection of their sets of gloss words).

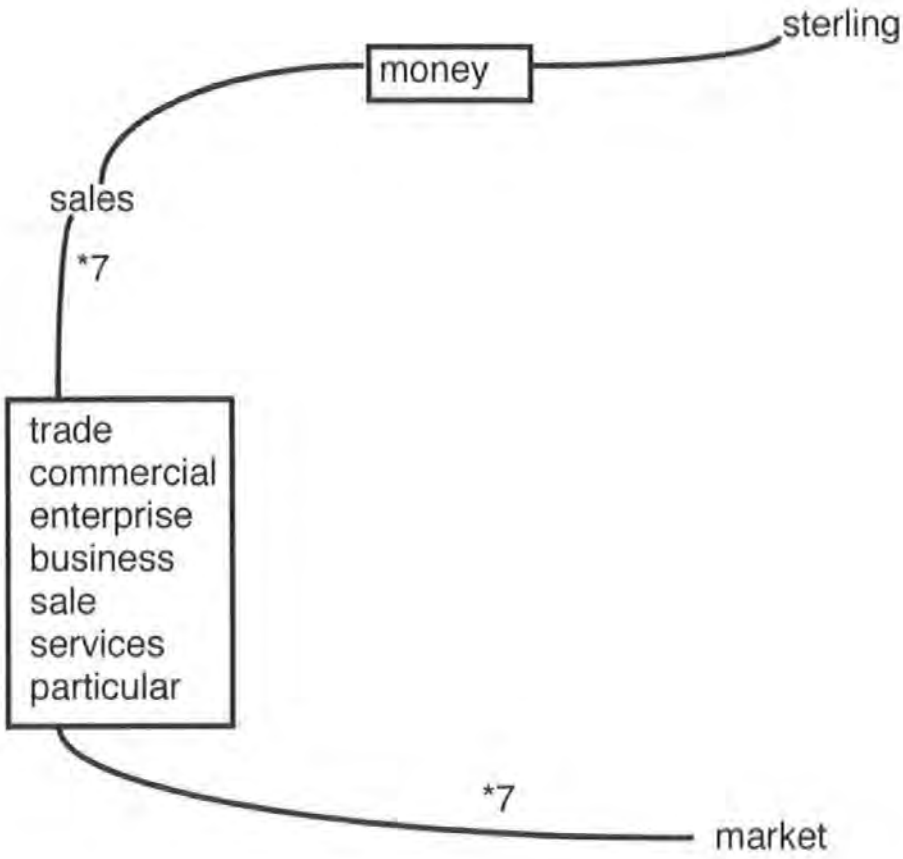


Figure 4.1. L1-L1 links for Text 1, using Restricted notion of Salience (Premodification only)

The associated numbers refer to the number of such links, and therefore represent the number of words which are common to the glosses of the linked parent terms. Thus, for example, the number 7 associated with the links between the salient nouns *market* and *sales* indicates that there are seven words which the glosses for these two items have in common, and these common words appear in the box through which the linkage lines pass.

Changing the basis of the assessment of the salience of individual nouns affects stage 2 of the processing (as described in Section 4.5.2 above), which results in the list of salient nouns produced being no longer *[market 5 sterling 5 sales 2]*, as in the initial stage , but *[games 2 toy 2 market 5 worth 2 million 6 pounds 5 sterling 5 sales 2]*.

Figure 4.2 shows the linkage diagram corresponding to this new situation, which is produced after L1 lookup. It shows that, whereas the frequency of occurrence of the nouns in the original abstract clearly remains the same, the total number of links in

which they are involved is different, since the set now being examined for linkage is a superset of that used in the previous state of the implementation.

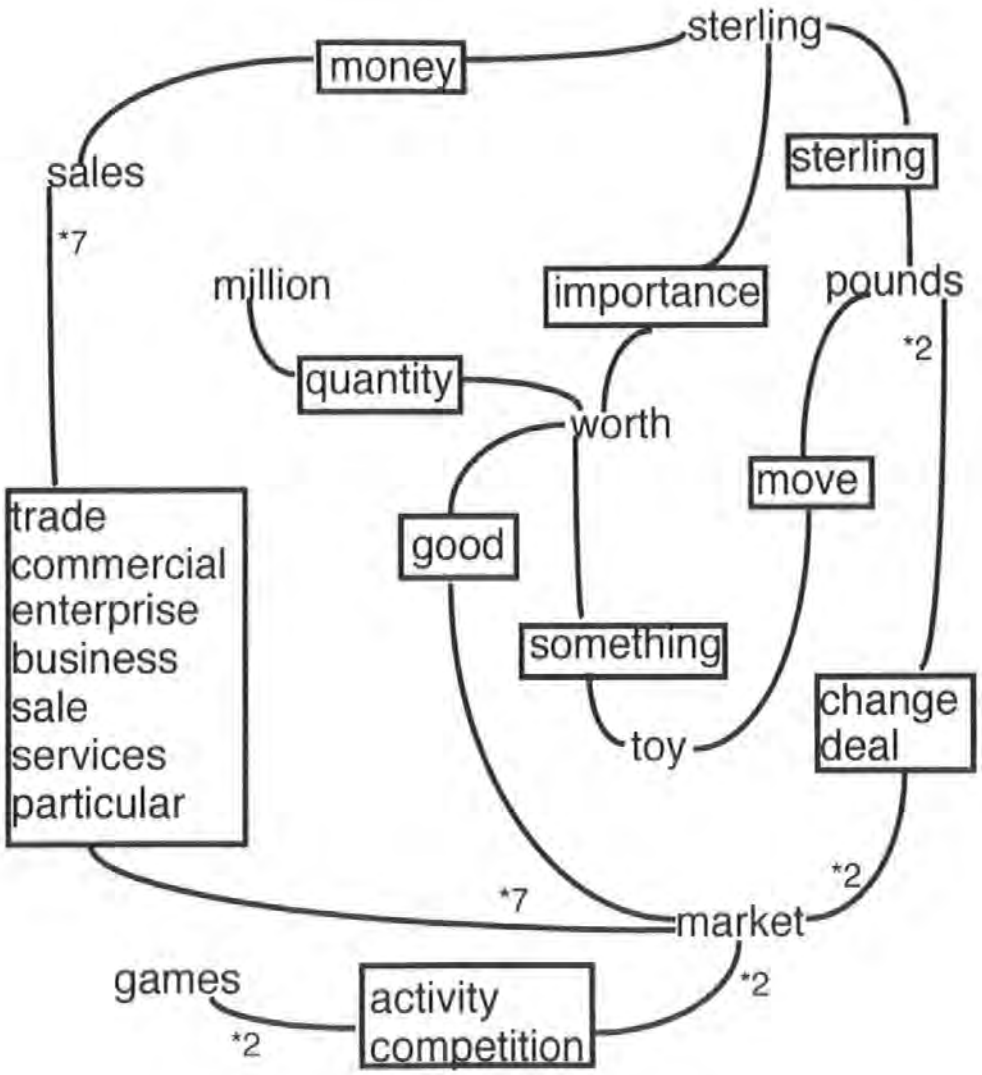


Figure 4.2. L1-L1 links for Text 1 based on Unrestricted Notion of Salience (frequency)

It is to be expected that the linkage produced from the second, unrestricted list (the superset) will be greater than that resulting from the first, restricted list. It is, in principle, possible that the linkage will be the same in the two instances, since the additional terms may not exhibit linkage at all. However, the superset will obviously never produce less linkage than the subset.

Figure 4.2 shows the linkage between the salient nouns occurring in the new list, which becomes established after the L1 lookup procedure[26] . The number of links which connect an item to the rest of the network may be said to be at least partly representative of the salience of that item in the assessment of the overall aboutness of the text. This gives us a second indicator of salience, (the first being the frequency of occurrence of the terms themselves in the text). Comparison with Figure 4.1 indicates that the information from the latter source does not necessarily substantiate that from the noun frequency count.

Consider, for example, the terms *sales, market* and *million*. The noun frequency count alone would lead us to conclude that the term *million* is the most salient, occurring 6 times in the text, and thus is the most generally representative of the aboutness. Conversely, the word *sales* has a frequency of only 2, and on this basis alone would be treated as relatively unrepresentative of the aboutness if this were the only criterion considered. However, if we take a look at the amount of linkage these terms are involved in, we see that *million* has only one link (which connects it, via *quantity* to *worth*), whereas *sales* has 8 links (7 of which link it to *market*, with 1 linking it to *sterling*). In fact, the noun which exhibits most linkage is *market*, with a total of 12 links (7 linking it to *sales*, 2 to *games*, 1 to *worth* and 2 to *pounds*).

It is interesting that the unrestricted list of salient nouns gives us *million* as the most salient at the initial stage (with a frequency of 6), whereas using the restricted list gives us equally *sales* and *market* (each having a frequency of 5). If we are to compare these outcomes on an instinctive level of judgement, it seems clear that it is the result of the restricted approach which yields the most representative nouns. In other words, it seems reasonable to say that *sales* and *market* are far more representative of the aboutness of the abstract than is the word *million*.

When the output expressions produced for the whole set of texts were examined, it was found that in some cases the restricted notion of salience led to the most representative aboutness terms, and in other cases it was the unrestricted notion. Therefore, the final version of the system utilises both, and counts as salient any noun which is either premodified or occurs more than once in the text.

---

[26] Note that the processing at this stage did not look up multi-word lexemes as independent units (see Chapter Five), but treated individual words independently.

Improvement: Salience Based on Premodification and Frequency

The final stage of the implementation, therefore, utilises a refined assessment of salience, which is based on the frequencies of terms as they occur in the complete abstract, as well as the premodification criterion used in the initial stage. Thus, any noun which is premodified, or occurs in the input text with a frequency greater than 1 is considered to be salient.

It is noteworthy here that the indication of salience, or 'importance', is often based on frequency of occurrence alone. For example, Lin (1995), Hahn and Reimer (1998) and Wacholder (1998), all of whom aim to identify salient *concepts* based on frequency of occurrence alone. The method used in COMMIX extends this notion of salience, to include modification of items as an additional indication, which applies particularly well to abstracts.

Assumption 3

It is not necessary to resolve the reference of pronouns and proper names.

a) Pronouns are deleted as closed class items.

b) Proper names are simply labelled as such. An item is assumed to be a proper name if it begins with an upper case letter and occurs within a sentence.

Validity of Assumption 3

In the case of the example text, there are no pronouns used at all, and there are only three instances of the use of names. Note that the month abbreviations *Oct* and *Dec* have been labelled as names by virtue of their mid-sentence positions and capital initial letters. The restriction caused to the search space by this lack of resolution of reference is minimal in this instance.

However, the impact of this restriction on the processing of other abstracts which do have occurrences of pronouns is much greater, and leads to a failure in the recognition of identical referents, and consequently to a failure in the recognition of items which do occur more than once. In many instances this leads to a failure to produce any output expressions at all, since no item has been accorded any greater salience than any other item.

The reference of pronouns and proper names is therefore one which ideally needs to be addressed. It should be clear, however, that a full treatment of pronoun resolution

and replacement of names by types, would constitute an entire research project of its own. Whilst this is clearly not possible in the scope of this thesis, Chapter Five is dedicated to the matter, and presents a brief foray into this realm.

Assumption 4

Sentence-initial capital letters are assumed to be upper case merely by virtue of their position in the sentence, and therefore are replaced by lower case letters.

Validity of Assumption 4

This is generally a useful and valid assumption, since it enables a capital letter to be accorded the same interpretation as its lower case counterpart. The one problem (which arises in the treatment of the example abstract) regards the treatment of sentence-initial names, when we clearly do not want the upper case initial letter to be replaced by its lower case counterpart. This occurs in the example abstract with the proper name *Nintendo* (at the beginning of the final sentence), which after processing by COMMIX becomes labelled as the 'not listed' (i.e., *n-l) nintendo,* and is not recognised as an occurrence of the same item. This exacerbates the problems of reference of names mentioned above.

Improvement: De-sensitise for case of initial letter

For this reason the COMMIX system in its final implementation included a desensitisation to case as regards the initial letters of names. If an item occurs mid-sentence with a capital letter, then its additional occurrence sentence-initially does not result in decapitalisation. For instances where the only occurrence is sentence initially, this may result in its decapitalisation. However, this does not affect the subsequent analysis of linkage in which the item is involved, since WordNet is case-insensitive.

Assumption 5

Restricting the level of lookup to 1 is sufficient to indicate linkage between semantically related items.

Validity of Assumption 5

The processing of the example abstract shows that some linkage is picked up at the L1 level. However, there is no doubt that this restricts the amount of linkage identified, and that more semantic relatedness would be identified by increasing the

lookup level to 2. As Chapter Six discusses, however, this does not necessarily result in less representative output.

Improvement: Extension to Linkage

The final stage of the implementation includes two improvements concerning the linkage between possible related items. Firstly, when looking for links (i.e. overlapping words in definitions), we extend the investigation of the linkage procedure to include the L2 level. Note that when we use WordNet we simply utilise the next level up in the semantic hierarchy. Although this has advantages in terms of processing power (and time), it provides a far less rich basis for establishing linkage. A second extension made here results from paying attention to establishing L0-L1 linkage where appropriate.

In the initial stage of the implementation, the only linkage searched for is L0-L0 or L1-L1. In its improved state, the system has been extended to investigate not only the L2 level (as discussed above), but also across lookup levels. For example, the gloss for the salient head *market* includes the term *industry* (as shown in the description of methodology above), which also appears in the original abstract (i.e., is an L0 term). There is thus justification for forming an L0-L1 linkage between the terms *market* and *industry*, and hence incorporating the modifiers of *industry* (i.e., *video games*) into the aboutness expression whose head is *market*.

Assumption 6

The modifiers within a compound nominal expression may be placed in any order.

Validity of Assumption 6

The output expression produced for the example abstract is:

hardware associated software toy 1991 computer games market

Although being fairly indicative of the content of the abstract, this expression does seem intuitively to contain some erroneous ordering of elements, as well as some redundant and inappropriate elements. It seems clear that the listing of modifying items according to the order in which they appear in the original abstract, although yielding some representative expressions, could be improved by some post-processing ordering of modifiers. Note that this post-processing stage is not included in the final implementation, since it would add a superficial layer to the system's

performance, and this in turn would obscure the evaluation of the system in its 'purest' form. As Chapter Six shows, the results are quite encouraging even in the absence of such a refinement.

We note, however, that whilst such a post-processing ordering facility would enhance the overall performance of the system, attention would need to be paid to the precise methodology. A principled ordering would be in line with theories and generalities relating to adjective type (as discussed in detail in Quirk et al (1985), Chapter 7). We would need to ensure, however, that components of modification units, which themselves comprised compound nominals or ADJ - NOUN pairs retained the relations between themselves as components of the whole, and did not become subject to a generalised reordering process which ignored their status in relation to one another within their modification units. We would expect, then, that the ordering of terms appearing in the output expressions is an aspect which would form the basis of some useful future work.

Assumption 7

At the labelling stage, if there is an ambiguity of word class involving the option 'noun', then this is taken as the default assumption.

Validity of Assumption 7

When the ambiguity of syntactic class is between noun and verb, there is justification for holding this assumption, since the improved methodology translates verbal information into nominal form. However, problems arise when the ambiguity is between noun and some class other than verb. The processing of the example text throws up one such example, the item *fast*, which appears as an adverb in the first sentence, but has been accorded the label 'noun'.

The incorporation of a parsing facility, which could be included in the type of hybrid system discussed in Chapter Seven, would go a long way to solving this particular problem. If we consider the practical working of the system, and the manner in which the aboutness terms are generated and read, however, we note that even if items are accorded incorrect tags at the initial stage of processing, this does not necessarily present as a problem in terms of the overall performance of the system. This matter has been discussed in greater detail in Chapter Three (Section 3.2).

Assumption 8

A single output expression is sufficient to give some indication of the aboutness of the abstract.

Validity of Assumption 8

Even the relatively unprincipled output expression produced by this initial stage of the implementation gives a surprisingly accurate indication of the aboutness of the abstract. However, the abstract is not only about some kind of market, but is also about company spending on advertising particular kinds of products, sales of particular products and so on. A fuller processing of the abstract to indicate such additional aboutness terms is required.

Improvement: More than one Aboutness Expression Generated

This matter relates to the distinction between intensional aboutness versus extensional aboutness. Section 3.2.2 (Chapter Three) discusses this distinction, with its associated implication that, since a text is about more than just one thing alone, the generation of more than one aboutness term would be beneficial. For this reason the final implementation has been extended to generate an unlimited number of aboutness expressions, with the precise number generated being dependent on the amount of linkage identified between items in the text.

One problem with the number of expressions being unlimited, occurs when there are such large amounts of linkage identified that the number of aboutness expressions generated is huge. This problem could be lessened by the incorporation of a full weighting process, which would therefore rank each of the terms generated according to principled criteria. As we discuss in the methodology description in this chapter, the development of a fully principled weighting system would depend on factors such as text type, prospective audience, etc., and would constitute an interesting area for the development of this work.

Assumption 9

In methodological terms, it is sufficient to treat individual words within a noun stream as independent of one another.

Validity of Assumption 9

The terms *video games, computer games* and *compact disc* are all listed as individual lexical items within WordNet. Their definitions are distinct from those of their constituent words, and contain different words to the set formed from the conjunction of the glosses of the constituents. The semantic linkage produced by using the two-word lexemes will be different to that resulting from using the glosses of the individual words.

It should be clear that there are two distinct but related issues involved here: namely lexicalisation and compositionality, both of which were discussed in Chapter Three.

From the processing point of view, there needs to be a comparison of the performance of the system in regard to the treatment of such noun streams. When a stream such as video games or *compact disc* exists in the database as an independent lexical item, its gloss should be at least partly the basis for the identification of semantic relatedness. Chapter Five includes a study which relates to this matter.

Improvement: Dictionary Lookup of all Noun Pairs Included

The final stage of the implementation was extended, so that all noun-noun pairs in the text were the subject of independent lookup, and labelled (and subsequently treated) as multi-term lexemes if they were found to be listed independently.

4.5.4 Summary of Improvements

The improvements which are implemented in the final version of the COMMIX system are summarised as follows:

- inclusion of procedure to nominalise verbal information (re: assumption 1)
- notion of salience is extended to include frequency as well as pre-modification (re: assumption 2)
- extent of search for linkage is refined and extended to L2 level (re: assumption 5)
- extended number of aboutness expressions generated to include all those for which linkage evidence is identified (re: assumption 8)
- extended lookup procedure to include search for multi-term lexical items (re: assumption 9)

4.5.4.1 Other points for consideration for future work

The above discussion has highlighted some additional points for possible future work. The major ones are listed below:-

- principled ordering of modifiers
- principled, full treatment of weighting, probably in relation to text type (see Chapter Seven)
- principled method of dealing with pronoun and names

The last point in the list for possible future work has in fact been the subject of some small-scale investigations over the course of the development of the COMMIX system. Although the resolution of pronouns and reference of names has not been implemented in the final system, some of the effects which their resolution can have on the output of the system have been investigated. These investigations form the basis of Chapter Five, which follows.

Chapter Five

5. An Investigation: Pronouns, Names and Multiple-Word Lexemes

Chapter Four, which described the implementation of the COMMIX system in detail, ended with a specification for the improved final version of the system, the performance of which has been evaluated as described in Chapter Six. The final section of Chapter Four also contained some indications for areas that would benefit from future work, and mentioned in particular is the matter of pronoun reference and the resolution of names.

These topics constitute enormous fields of research in their own right, and as such are precluded from inclusion in the formal implementation, which minimises the emphasis on syntax. However, the effect which the replacement of pronouns and names by their referents and types would have on the output expressions generated was considered to be of significant interest. Therefore, a series of investigations was undertaken, which involved the replacement of pronouns by their referents, and the replacement of names by their type of referent, in the input texts. It is important to note that the replacement was done by hand, and no attempt was made to automate this replacement, for reasons already mentioned. The effect of identifying (and subsequently using the definitions of) any multiple-word lexical items occurring in the texts is also described here.

This chapter describes these investigations and discusses the findings, interpreting the findings relating to pronouns and names in particular, as indications of the relative importance of the two types of replacement on the generation of aboutness terms. These investigations were carried out before the incorporation into COMMIX of the facility to nominalise verbal information. However, since the investigations show output before and after any replacement of pronouns and names, and the same version of COMMIX was used in each case, the comparisons described in this chapter are valid.

5.1 Topics for Investigation

It is the aim of this study to investigate the effect of names, pronouns and multiple-word lexical items (henceforth to be referred to as multi-word lexemes) on the usefulness of the output expressions generated by COMMIX. In the case of pronouns and names, these investigations involve the replacement of the non-specific item by the specific referent. The replacement procedure itself raises several issues of interest, such as:

- what specific referent should be used when more than one exists in the text?
    - do we just use the last one?
- with what should possessive pronouns be replaced?
- where should we plug into the hierarchy of names?

These (and a good many other) questions are part and parcel of the difficulties which have led to the establishment of these topics as fields of study in themselves. It is important to note that the replacements - even when done by hand - are far from uncomplicated or uncontroversial. However, although we bear these points in mind in the ensuing investigations, we are primarily concerned with the effect of any replacement which will encourage a matching of referents when multiple terms of reference are used in the texts to refer to one entity. As such, the actual items used in the replacement process are subjective to the author.

As in Chapter Four, the expectations are presented in the form of mini-hypotheses, or local assumptions, the validity of which then provides the forum for subsequent discussion. In this chapter, there are three such assumptions, relating respectively to pronouns, names and multi-word lexemes.

5.1.1 Assumption 1: Pronouns

The replacement of pronouns by the nouns to which they refer (or the type if reference is to a name) leads to an improvement in the representativeness of the compound nominal expressions generated for the texts under this implementation.

5.1.2 Assumption 2: Names

The replacement of proper names by their types of referents leads to an improvement in the representativeness of the compound nominal expressions generated.

5.1.3 Assumption 3: Multi-Word Lexemes

The treatment of multi-word lexemes as single units (and the consequent looking up of the 'compound' term) yields some linkage information which is not available if the constituent words of a multi-word lexeme are looked up independently.

5.2 Texts Used in the Investigations

The 'video games industry' abstract (Text 1[27] in the data set) which has so far been used as an example, does not contain any pronouns, but does contain names and multi-word lexemes. The 'designer fashion industry' abstract (Text 2) contains both names and pronouns, but no multi-word lexemes, whereas the 'soap opera' abstract (Text 6) includes all three categories.

Although the full subset of ten abstracts which comprise the unrelated set (i.e. Texts 1 to 10) has been tested in these investigations, the behaviour of the set is largely typified, encompassed and summarised by a discussion of the behaviour of these three abstracts. There is, however, one notable exception whose findings contradict those of the other texts. This is the abstract about Derek Walcott, which is Text 7 in the data set, and is discussed in Section 5.3.4 below.

The processing of each of Texts 1, 3, 6 and 7 is discussed below, firstly with regard to the effects which replacement of names and pronouns has on the output. Secondly, we describe the effects of using the independent constituents of multi-word lexemes, rather than the multi-word lexemes themselves, as the basis for the lookup procedure. Note that the effect of the replacement of names and pronouns has been investigated using the multi-word lexemes as the basis for the lookup procedures, rather than the constituent words.

The effect of the use of constituent words as opposed to multi-word lexemes has been investigated using the standard COMMIX methodology, which does not

---

[27] Reference to text numbers refers to their numbering in the data set, as evaluated in Chapter Six. The full set of texts, along with their COMMIX-generated aboutness terms, appears in Appendix 1.

include any replacement of pronouns or names, and this is discussed as a separate issue in Section 5.4.

5.3 Text 1: Video games industry

This abstract has no pronouns, only names.

5.3.1 Text 1 Replacements

The names involved are 'Sega' and 'Nintendo', both of which are replaced by the type noun 'company'. It is the multi-word lexemes which are listed in their own right in the WordNet database which are used here: these are 'video_games', 'computer_games', 'pounds_sterling' and 'compact_disc'. Points a) and b) below give the output expressions respectively for the original abstract where names (N) remain unresolved (U), and for the case where names (N) are replaced (R) by their type of referent. For the sake of clarity, each output expression is presented between square brackets.

5.3.1.1 Text 1 Output Expressions

a) Output Expressions where Names Remain Unresolved (NU)

- [toy 1991 computer_games hardware associated software compact_disc market]
- [worth 275 500 635 645 345 spend 15 million]
- [toy 1991 computer_games compact_disc market]
- [video_games market]

b) Output Expressions where Names have been Replaced (NR)

- [toy 1991 computer_games hardware associated software compact_disc market]
- [worth 275 500 635 645 345 spend 15 million]
- [toy 1991 computer_games compact_disc market]
- [video_games market]
- [hardware associated software company]
- [established company]

5.3.1.2 Comments Relating to Text 1 Replacements

Name replacement yields two additional aboutness terms, [hardware associated software company] and [established company]. In terms of useful output, which is representative of the aboutness of the text, the replacement of names by their type nouns appears to make little difference.

One point should be noted here, which is that the substitutions have involved the same referent type for different specific occurrences of the type: i.e., the word 'company' has been the common replacement for 'Nintendo' and 'Sega'. Whilst it is acknowledged that this might be problematic in terms of larger texts, such replacement seems justified in the genre of abstracts, and particularly in relation to the generation of aboutness expressions. Thus, even though the specific references are different, it is still valid to utilise the common type replacement, since 'company', in terms of aboutness, relates to both examples, and therefore to the general aboutness terms which apply to the text as a whole.

5.3.1.3 Conclusion[28]

Name replacement makes little difference to the usefulness of the output expressions.

5.3.2 Text 3: Designer Fashion Industry

Since this text has both pronouns and names, there are four permutations to be tested, and consequently, 4 sets of output results: a) gives the output when pronouns are deleted and not replaced by any referent (as in the original implementation) and names are simply labelled as names, but ignored as far as the generation of the output expressions is concerned. Section b) gives the output for the case where pronouns are replaced and names remain as they are; c) gives the output where names have been replaced by their type of referent and pronouns are deleted; d) gives the output where both pronouns and names have been replaced by their referents.

5.3.2.1 Text 3 Replacements

The names occurring in this abstract are 'Paul Smith', 'Jasper Conran', 'Vivienne Westward', 'Paris', 'London' and the more problematic 'British Designer of the Year

[28] We stress the point that these comments relate to the particular texts mentioned, although the conclusions apply to the whole set of 10 texts (with the exception of Text 7)

Award'. The pronouns occurring in the text are 'they' (to refer to Paul Smith and Jasper Conran), 'her' (the possessive pronoun referring to Vivienne Westward), 'some' (used to refer to 'some people'), and 'they' (subsequently to refer also to 'some people').

The pronouns were replaced by their referents as given above, and the names were replaced by their type nouns: names of designers were replaced by the noun 'designer'. Although 'Paris' and 'London' are both names, they each occur in the WordNet database and are thus looked up as standard items. Although the title 'British Designer of the Year Award' is a name, there is a degree to which it may be described as compositional, with its constituents all being listed as independent items in WordNet. In regard to this name, the issue of lexicalisation of multi-word lexemes does not really arise, since it is a highly specific title and would not be expected to become a lexeme in its own right.

5.3.2.2 Text 3 Output Expressions

The following four sets of output expressions encompass the various combinations of pronouns (P) (being either deleted (D) or replaced (R)) and names (N) (being either un-typed (U) or replaced (R) by the type of object to which they refer). Thus, for example, the abbreviation 'PD/ NU' refers to the case in which pronouns have been deleted and names remain un-typed.

a) PD/NU

- [designers][29]

b) PR/NU

- [designers]
- [designer fashion clothing industry]
- [designers designers][30]
- [top designers represent people]

---

[29] Single lexical items are excluded from consideration in the final output. They have been identified as salient items, but do not have any identifiable modifiers to include in a compound nominal expression. Although they clearly do represent what the text is about to a degree, they are nevertheless eliminated from the output in the final version of the implementation.

[30] Such exact repetition of lexical items does not occur in the final implementation

- [designers people]

c) PD/NR

- [designers]
- [designers designer]
- [designers designers]

d) PR/NR

- [designer fashion clothing industry]
- [designers]
- [designers designer]
- [designers designers]
- [top designers represent people]
- [designers people]

5.3.2.3 Comments Relating to Text 3 Replacements

The difference made by pronoun replacement can be seen from a comparison of the output from a) and b) shown above, and of that from c) and d). In the former situations names remain untyped, whereas in the latter cases, types have been substituted for names. In each case the comparison is between the cases in which pronouns are both deleted and replaced.

Similarly, the difference made by name replacement can be seen by examining the output from a) and c) (in each of which pronouns are deleted), and from b) and d) (in each of which pronouns are replaced).

The Difference made by Pronoun Replacement

When names remain untyped, there is a substantial difference between versions in which pronouns are, firstly, deleted and secondly, resolved (compare a and b). Whereas a) (in which pronouns are deleted) yields only [designers] as a final output expression, the replacement of pronouns by their referents yields the final expression [designer fashion clothing industry]. Although no formal evaluation has been done

in relation to these investigations, it seems clear that this expression is fairly representative of the aboutness of Text 3.

When names are replaced by their type nouns, there is a substantial difference between versions in which pronouns are deleted and replaced (compare c and d). As in the case above (where names are untyped), there is no useful output in the case where pronouns are deleted, but the same representative expression, [designer fashion clothing industry], is produced when pronoun reference is resolved.

The Difference made by Name Replacement

A comparison of b) and d) indicates that name replacement makes little difference. In b), where names remain untyped (and pronouns have been replaced by their referents), the final output expressions are the same as above, including the representative [designer fashion clothing industry]. When names are replaced by their type nouns the only difference is that the unrepresentative expression [designers designer] is produced. Thus, when pronouns are replaced by their referents there is little difference between versions in which names remain untyped and where names are replaced by their types.

A comparison of a) and c) similarly indicates that when pronouns are deleted there is little difference between the versions in which names are respectively typed and untyped. In other words, for Text 3, it is the pronoun resolution that makes the most difference (and leads to more useful and representative output), regardless of name replacement.

5.3.2.4 Conclusion

Pronoun replacement makes the major difference, whilst name replacement makes little difference.

5.3.3 Text 6: Soap Operas

This abstract has all three categories under investigation: pronouns, names and multi-word lexemes.

5.3.3.1 Text 6 Replacements

The pronouns occurring in this text are 'him' and 'he', which both refer to 'a soap opera writer'; 'them', which refers to 'a large number of characters from Eastenders'; and 'who', referring to 'an actor'. The names occurring in the abstract are 'British

Broadcasting Corporation', 'Eastenders', 'IRA', 'Britain' and 'United States'. Note that in this text the following names have been left untyped, as each element appears in WordNet: 'British Broadcasting Corporation', 'Britain', 'United_States'. Names which have been replaced are 'Eastenders' (by 'soap_opera') and 'IRA' (by 'terrorist').

There is one multi-word lexeme, 'soap opera', which appears as an independent lexeme in the WordNet database.

5.3.3.2 Text 6 Output Expressions

As with Text 2 above, there are 4 sets of output expressions for consideration, since Text 6 also has both names and pronouns.

a) PD/NU

- [soap_opera characters]

b)PR/NU

- [soap_opera writer]
- [killing soap_opera characters]
- [soap_opera killing characters]

c) PD/NR

- [soap_opera characters]

d) PR/NR

- [soap_opera writer]
- [killing soap_opera characters]
- [soap_opera killing characters]

5.3.3.3 Comments Relating to Text 6 Replacements

The difference made by pronoun replacement can be seen from a comparison of the output from a) and b) shown above, and of that from c) and d). In the former case names remain untyped, whereas in the latter case, types have been substituted for names. In each case the comparison made is between output generated when pronouns are firstly deleted and then replaced.

Similarly, the difference made by name replacement can be seen by examining the output from a) and c) (in each of which cases pronouns are deleted), and from b) and d) (in each of which, pronouns are replaced).

The Difference made by Pronoun Replacement

When names remain untyped, there is a substantial difference to the final output expressions, between versions in which pronouns are, firstly, deleted and secondly, resolved (compare a and b). Whereas a) (in which pronouns are deleted) yields only [soap_opera characters] as a final output expression, the replacement of pronouns by their referents yields the final expressions [soap_opera writer], [killing soap_opera characters] and [soap_opera killing characters]. Again, these investigations are not subject to formal assessment, but it seems clear that these expressions do give a good indication of the aboutness of the abstract.

When names are replaced by their type nouns, there is a substantial difference between versions in which pronouns are deleted and resolved (compare c and d). As in the case above (where names are untyped), there is one (reasonably informative) output expression produced when pronouns are deleted, but the same three expressions are produced when pronoun reference is resolved.

The Difference made by Name Replacement

A comparison of b) and d) indicates that name replacement makes little difference. In b), where names remain untyped (and pronouns have been replaced by their referents) the final output expressions are the same as above, comprising the three expressions [soap_opera writer], [killing soap_opera characters] and [soap_opera killing characters]. When names are replaced by their type nouns there is no difference in the output. Thus, when pronouns are replaced by their referents there is no difference between versions in which names remain untyped and where names are replaced by their types.

A comparison of a) and c) similarly indicates that when pronouns are deleted there is no difference between the versions in which names are respectively typed and untyped. In other words, it is the pronoun resolution that makes the difference (and leads to more useful and representative output) regardless of name replacement.

5.3.3.4 Conclusion

Pronoun replacement makes the major difference, whereas name replacement makes no difference.

5.3.4 Text 7: Derek Walcott

As mentioned above, this text is the only one of those containing pronouns or names which behaves atypically from the rest of the set of texts investigated here.

5.3.4.1 Text 7 Replacements

In this case, the names Derek Walcott, and Walcott were replaced by the type 'writer'. The pronouns 'he' and 'his' were also replaced by 'writer'.

5.3.4.2 Text 7 Output Expressions

There are once more 4 permutations to test, and thus 4 sets of output results on which to comment.

a) PD/NU

- [playwright as well first]
- [first first]

b) PR/ NU

- [playwright as well first]
- [first first]

c) PD/ NR

- [epic writer]
- [writer]
- [1992 writer]
- [playwright as well first]
- [first first]

d) PR/NR

- [writer published first book when epic writer]
- [writer published first book when writer]
- [1992 writer published first book when writer]
- [writer prize]
- [playwright as well first]
- [first first]
- [writer poem]

5.3.4.3 Comments Relating to Text 7 Replacements

Difference made by Pronoun Replacement

When pronouns are replaced there is a noticeable difference between versions with names unresolved and names resolved (compare b and d); i.e., name replacement does make a difference (when pronoun reference is resolved).

When pronouns are deleted there is a noticeable difference between versions with names unresolved and names resolved (compare a and c). In other words, name replacement improves output even when pronoun reference is not resolved.

Difference made by Name Replacement

When names are replaced by their types of referents, there is a noticeable difference between versions with pronouns deleted and replaced (compare c and d), although it is difficult to say which of c) and d) is the more representative of aboutness.

When names are not replaced, pronoun resolution makes no difference (compare a and b).

5.3.4.4 Conclusion

Name replacement improves output particularly when pronoun reference is resolved, but also when pronouns are deleted. We note, however, that the replacement for names and pronouns is the same item: i.e., 'writer'. This increases the salience of 'writer', but does not necessarily increase the likelihood of linkage being identified.

Pronoun resolution changes (but doesn't necessarily improve) output when names are also replaced. Pronoun resolution causes no difference in output when names are not replaced.

5.4 Multi-word Lexemes

Two of the texts considered so far contain multi-word lexemes, these being Texts 1 and 6 (video games industry and soap operas respectively). The effects of replacing the multi-word lexeme with its constituent words in the processing of the texts had the same effect in each case. For this reason, only the performance of Text 1 is discussed below. The ensuing discussion, whilst it refers particularly to Text 1, applies also to Text 6.

5.4.1 Text 1: Video games industry

This text contains the terms 'video_games', 'computer_games', 'pounds_sterling' and 'compact_disc', which all exist as independent lookup items in WordNet.

5.4.1.1 Text 1 Output Expressions

a) Output for Linked Version (version with multi-word lexemes looked up)

- [toy 1991 computer_games hardware associated software compact_disc market]
- [500 million]
- [toy 1991 computer_games compact_disc market]
- [video_games market]

b) Output for Unlinked Version (version with constituent words looked up)

- [toy 1991 computer games hardware associated software compact disc market]
- [500 million]
- [toy 1991 computer games compact disc market]
- [hardware 635 million pounds associated software 645 worth 345 spend 15 sterling]
- [hardware associated software pounds]
- [635 million pounds 645 worth 345 spend 15 sterling]

- [pounds]
- [video games market]
- [spend 15 million pounds pounds]
- [worth 345 million pounds pounds]
- [645 million pounds pounds]
- [635 million pounds pounds]

5.4.1.2 Comments Relating to Text 1

As we might expect, the unlinked version includes all the expressions from the linked version. The additional expressions produced by utilising the constituents of the unlinked expressions do not enhance the general representativeness of the expressions.

5.4.1.3 Conclusion

Utilising the unlinked constituents of the multi-word lexemes in the lookup process produces all of the expressions which are produced when the multi-word lexemes themselves are used. The output using the linked version is a complete subset of that produces using the unlinked constituents (i.e., it does not contain any additional expressions). However, the additional expressions produced using the unlinked constituent words do not seem particularly representative of the aboutness of the text. In other words, utilising the constituents of multi-word lexemes rather than the multi-word lexemes themselves yields no particular advantage in the case of this text. Utilising the latter, however, has the advantage of reducing the number of output expressions produced overall.

5.5 Validity of Local Assumptions

This section refers back to Section 5.2, in which the expectations related to these investigations were specified in terms of assumptions.

5.5.1 Names

The assumption was that the replacement of proper names by their types of referents would lead to an improvement in the representativeness of the compound nominal expressions generated.

However, the output produced for Texts 1, 3 and 6 suggests that the replacement of names does not affect the expressions produced in a particularly useful way. Name replacement, then, seems to affect the output only slightly, judged by the output from these 3 texts.

Looking at the output from Text 7, however, gives a different picture. Here we see that the replacement of names does render the output more useful, and that the resolution of pronouns alone (i.e. when names are not also replaced) has no effect. This finding contradicts the general finding from the other texts (both those mentioned specifically above, and others in the test set which produced the same findings).

It seems plausible that an explanation for this finding lies in the characteristics of Text 7, and the replacements made in this text. Text 7 is highly biographical, with many references made to the same referent (i.e., Derek Walcott). Not only do all the occurrences of names refer to him, but we also find that all the pronoun references also refer to him. The result of replacing the pronouns, in effect, becomes a type of name replacement, with the result being a large number of occurrences of the same referent. It comes as no surprise, then, when this produces output expressions which lead to the conclusion that name replacement has the major effect on the output.

5.5.2 Pronouns

The assumption was that the replacement of pronouns by the nouns to which they refer (or the type if reference is to a name) would lead to an improvement in the representativeness of the compound nominal expressions generated for the texts.

This assumption is largely supported by the output expressions, which suggest that in general the replacement of pronouns by their referents does produce expressions which are more representative of the aboutness of a text, at least in the general case.

There are problems which arise in relation to the specific replacement terms used. Specifically, deciding which precise term of reference to substitute for each instance of a pronoun; and how best to handle the replacement of, for example, possessive pronouns, relative pronouns and so on. In selecting the replacements that have been made in these investigations, the choice has been made according to the particular occurrences of the referent in the rest of the text. In cases where there is more than one possibility, the most general (i.e., the hypernym) has been used.

5.5.3 Multi-Word Lexemes

The assumption was that the treatment of multi-word lexemes as single units (and the consequent looking up of the compound term) yields some linkage information which is not available if the constituent words of a multi-word lexeme are looked up independently.

There was no evidence of this being true, although only 2 texts from the 10 texts in this investigation set contained such constructs. In the light of this finding, a follow-up study of this phenomenon, using a larger and varied data set, would be expected to yield some interesting results.

5.6 Conclusion

The general findings of these investigations suggest that the incorporation of a facility for resolving the reference of pronouns would be a worthwhile addition to the COMMIX system. On the other hand, there does not seem to be much justification for the replacement of names by their referents, although we must remember that these comments are made on the basis of just a small number of texts.

The investigations in relation to multi-word lexemes suggest that there is little value to be gained by looking up the constituents instead of the multi-word term. However, since only 2 other texts contained multi-term lexemes, this indication was not felt to be supported sufficiently to be instantiated in the final implementation of the system, as evaluated in the next chapter. The final implementation, then, involves the looking up both of all potential multi-word lexemes, and of their constituent words, and the data from both is utilised in the generation of the output expressions. It is for this reason that some of the sets of output expressions were large, and this was also responsible for some degree of partial overlap between the different expressions generated for a given text.

These investigations have briefly raised some of the common questions which relate to the resolution of pronoun and name reference.

As regards the replacement of names:

- where to 'plug into' a hierarchy of representations (should 'Vivienne Westwood' be replaced by 'designer', 'woman', 'human', etc?)

- when to replace a name with its type, as opposed to looking it up in a database
- how to treat 'compositional' names, such as 'British Designer of the Year Award'

In regard to the replacement of pronouns:

- precisely what to replace pronouns with (whether to select the last, the first, or the most frequent occurrence of its referent)
- whether it is relevant to mark possessive pronouns as possessive (specifically in the context of this work).

5.7 Summary

This chapter has described some investigations into the effect on the performance of the COMMIX system of replacing the names and pronouns which occur in some of the texts. In this regard, the indications are that it is more worthwhile replacing pronouns than names, although there is an indication that the processing of biographical texts might also be significantly enhanced by the replacement of names.

Investigation of the effects of utilising multi-word lexemes, rather than their constituent words, in the lookup process, indicates that the latter might not improve performance, although the paucity of data leads us to include both in the final implementation.

The investigations described herein give an indication, albeit a brief one, of the manner in which the COMMIX system can be used as a tool for investigating a field of linguistic research. As mentioned in Chapters One and Seven, it is hoped that the system might prove useful in the future, for other such studies.

Whilst the investigations described in this chapter are informal, and not subjected to a full evaluation, the success of COMMIX at generating representative aboutness expressions clearly needs to be evaluated.

Chapter Six

6. Evaluation

This chapter describes the evaluation of the performance of the COMMIX system according to two specific criteria, one of which is a novel approach developed by the author. The chapter comprises three main sections. Section 6.1 describes the general approach taken to the evaluation; Section 6.2 describes the processes involved in setting up the evaluation; Section 6.3 describes the evaluation methodology in detail; and the final section presents the results obtained from the evaluation process. Throughout the chapter, the reader is asked to bear in mind the specific aim of the evaluation, which is stated clearly in Section 6.1 3 below.

6.1 General Approach

It should be clear that the output expressions generated by COMMIX can be viewed in two lights. Firstly, and most directly, the work is aimed at generating (compound nominal) noun phrases which capture the aboutness of input text. These expressions constitute, in effect, compact and informative 'mini-summaries', whose purpose is to inform humans of the aboutness of the associated text. Any evaluation of the extent to which they fulfill this purpose requires some metric which reflects human judgements about the degree to which the expressions constitute good representations of aboutness.

The second manner in which the aboutness expressions may be viewed relates to their potential for use in IR. If viewed in this light, they constitute complex indexing terms, whose aim is to adequately reflect the subject areas of the associated text. An evaluation of the extent to which this aim is fulfilled would be best assessed according to the standard criteria adopted for such evaluations: namely, *precision* and *recall.*

Although it would be interesting and desirable to perform the standard IR tests on the output, it was beyond the scope of this thesis to perform both types of evaluation. The emphasis throughout his work has centred on the degree to which the aboutness terms succeed in representing the aboutness of the input text, which corresponds to

the first scenario mentioned above. For this reason, it was decided that the evaluation stage would concentrate on just the first aspect, i.e., the degree to which expressions succeeded in representing aboutness. This was indeed the more challenging of the options, since it involved the development of a methodology and metric for the task. The evaluation procedure and metric described in this chapter, then, relate specifically to the assessment and expression of the degrees to which the aboutness terms were judged by humans to be representative of the aboutness of the texts.

We continue with an elaboration of the standard measures of precision and recall, explaining why these measures are not ideally suited to the type of evaluation performed, and emphasising that the discussions are made in relation to the specific measurement of success at representing aboutness.

6.1.1 Standard Measures of Evaluation: Precision and Recall

With the increasing production of systems which automatically summarize the content of text, comes an increasing concern with the methods of evaluation used to assess their performance. Mani & Hahn (1998) distinguish between *intrinsic* and *extrinsic* methods, where the former examines the text quality of the summary itself, and the latter concerns the usefulness of the summary for a particular task.

It should be clear that the work described in this thesis falls into the category of extrinsic methods: COMMIX generates aboutness expressions with the specific purpose of representing the essence of text. The performance metric typically used in these circumstances involves the notions of *precision* and *recall*, traditionally used to assess the performance of Information Retrieval or Information Extraction systems, and often used to evaluate summarization and indexing systems.

The specific functions of particular systems determines the exact means of calculating the precision and recall, but the underlying notion is the same. The *precision* expresses the number of those items identified by the system as being relevant which are truly relevant. The precision is therefore lowered if a lot of identifications made by the system are wrongly included (i.e., falsely identified as relevant). The *recall* expresses the number of relevant items existing in the database which are correctly identified as relevant by the system. Recall is therefore lowered if the system misses a lot of relevant items (i.e., falsely identifies them as irrelevant).

To put this in terms relevant to Fat Harry, introduced in Chapter One, increasing the precision will save him having to carry around in his handbag lots of information that he does not need. On the other hand, increasing the recall will make sure that his handbag contains plenty of the stuff he does need.

6.1.1.1 Precision and Recall Presuppose 'Correctness'

There is one main problem with using the measures of precision and recall, and this concerns the prerequisite of *correctness*. The potential problem caused by this requirement is not always apparent, notably when used to evaluate tasks to which the notion of a *correct* performance appropriately applies.

Information Retrieval systems, in general (a good summary of which appears in Salton (1991) and the many references therein), aim to match a specific query term against a database which includes a mixture of documents, some of which are relevant to that query, and others of which are irrelevant. Performance in this case is therefore appropriately measured in terms of the precision and recall of the system, since these measures provide a useful comparison between, on the one hand, all the documents identified by the system as matching the query, and on the other hand, all those documents which are judged by a human as being relevant to that same query. Although this does entail the assumption that relevance can be decided objectively (which can sometimes seem over-simplistic), it is generally possible to say which of the set of data items do apply to a given query, and therefore the assumption of correctness is in most of these cases unproblematic.

Information Extraction systems (Sparck Jones, 1994; Pazienza,1997) are generally geared to the aim of extracting specific information from texts belonging to a given genre, within a highly specified and restricted domain. The domain dependence results from the overall methodology, which aims to fill empty slots in pre-defined templates with the relevant extracted information. The high degree of domain-dependence, and the specificity of the task involved in filling in the missing information means that it is usually fairly easy to decide what information from the text should correctly be assigned to each slot. Again, the prerequisite of correctness is not an obvious problem in this endeavour.

Automatic summarization systems concentrate on identifying the important topics or sentences in a text, with the purpose of extracting them from the text and (generally)

concatenating them to form a summary. Where precision and recall are the metrics used to evaluate these systems, we begin to see problems. Although for the purposes of topic-tracking (by which different references to particular topics are identified throughout a (part of a) text), we might cling to the notion of 'correct', it is clear that the condensation of text content into a 'correct' summary has no such easily determined counterpart. This is the main reason for the recent interest in developing more appropriate evaluation metrics for automatic summarization (e.g., Hand, 1997) and topic identification (e.g., Wacholder, 1998).

It should be clear, however, that the COMMIX system falls into none of these categories. As discussed in Section 2.2 of Chapter Two, it can be broadly, and functionally, described as a type of Information Extraction system, in that it extracts information from text automatically. But it is aimed primarily neither at the query-driven identification of relevant documents that dominates the field of Information Retrieval (although its extension in this direction would seem logical), nor at the highly pre-specified, domain dependent template filling tasks typically associated with the field of Information Extraction. Nor is it aimed at reproducing the gist (see Chapter Two) via full-text summaries. Instead, COMMIX operates at the level of indicating aboutness, by means of generating expressions to represent the essence of text. In assessing the success of this system, then, we are primarily interested in the degree to which the aboutness expressions generated for the texts do actually represent what those texts are about. 'What a text is about' is a subjective matter, and it is hard to see how we can justify the notion of 'correctness' as applied to the degree to which aboutness terms represent their text. This information is not captured by the measures of precision and recall.

6.1.2 A Variation on the Notion of Precision: 'Appropriate' rather than 'Correct'

Although the standard measurement of precision is not directly suited to this application, we are of course interested in measuring the extent to which the expressions produced by the system do actually reflect what the text, for and from which they were generated, is about. We can adopt a variation on the notion of precision, in which we refer to the degree to which an aboutness term, or a set of aboutness terms, achieves its maximum potential in regard to representing the aboutness of the corresponding text. The degree of *achievement* in this regard may be

expressed as a measure of its *appropriateness*, or its *success* at achieving its purpose, which is to be *representative* of the aboutness.

In this evaluation, therefore, we exploit the notion of *attainment*, in which aboutness terms are judged according to the extent to which they succeed in representing the text to which they refer. The judgements are made by humans, and expressed as ratings on a scale, the lower end of which represents no achievement, and the upper end of which represents maximum achievement. By associating ratings with points along this scale, we can therefore express opinions about the degree to which an aboutness term or set of terms achieves its full potential.

The actual attainment of a term or set of terms may then be expressed as a percentage of the maximum possible attainment[31], and it is this measure which is adopted as one of the criteria for evaluating the success of the system at generating representative aboutness expressions. It is expressed more formally in Section 6.4.2.1, where it is used in relation to the specific data obtained in the evaluation exercise.

6.1.3 Aim of the evaluation

The aim of the evaluation described in this chapter is to test the following hypothesis. The COMMIX system generates expressions that are representative of the aboutness of a piece of input text. These expressions are significantly more representative of that aboutness than dummy expressions which are also generated by the system, from the same input texts, and have a similar superficial structure. The evaluation also aims to determine the extent to which the genuine aboutness terms do actually represent what their corresponding texts are about.

To this end, the general method employed in this evaluation involves presenting subjects with the texts and their corresponding COMMIX-generated aboutness expressions[32] (including the dummy expressions), and asking them to judge the degree to which each expression is representative of the aboutness of the corresponding text. The overall aim is to determine both the extent to which opinions about the genuine expressions differ from those relating to the dummy expressions,

---

[31] This expression of actual performance expressed as a percentage of total possible performance seems close to the notion of *percent agreement* adopted by Jing et al (1998), which came to light as this thesis was being submitted.

[32] Note that the term 'aboutness expression' and 'aboutness term' are used synonymously to refer to the output expressions generated by COMMIX. As explained in Chapter Two, these terms represent the *essence* of their corresponding input texts.

and the extent to which the genuine aboutness terms actually do represent what the corresponding text is about.

Sections 6.2 - 6.4 below describe the evaluation methodology, and the consequent analysis of results in detail. The former begins with a specification of the starting point of the analysis - the data available, and continues through a detailed description of the stages which constituted the evaluation process. The construction of the questionnaires presented to subjects for their opinions is described, as is the system by which subjects' opinions are recorded. The latter section (6.4) elaborates on the type of analysis selected for the interpretation of the results, and presents and discusses the results themselves.

This evaluation, then, is geared towards the prospective user, and centres around the degree to which people judge the aboutness terms to be successful in representing what the associated text is about. In terms of the overall principle of being user-directed, it is in line with the comments made by Hand (1997) in relation to the evaluation methodology under development for TIPSTER III. Although her comments are made in relation to the evaluation of automatic summarising systems, the principles apply to this work nonetheless. In particular, she notes that the main problem with some of the other (notably statistical) methodologies, lies in the "reliance on the notion of a single 'correct' abstract" (ibid, p. 32), which, as we have discussed, is over-simplistic.

6.1.4 Materials

The data set consists of the following: a set of 20 abstracts, which were selected from the same wide-coverage database, and which fall into two distinct subsets of 10. One subset of ten comprises abstracts which are not related to one another in terms of their aboutness. These abstracts were selected so as to have a broad range of subject area, as well as showing some variation in text type. For example, care was taken to include texts written in an 'open' style (such as the biographic Text[33] number 7, as shown in Appendix 1) as well as a more expository style (e.g., Text number 4). The other subset of ten (Texts 11-20) were selected on the basis of similarity, and are all related to one another in subject matter.

---

[33] The complete set of texts appears in Appendix 1, along with the aboutness terms generated for them.

Another part of the data consists of aboutness terms or expressions generated by the system for each of the 20 abstracts. These terms are compound nominals which are intended to express the essence of what the abstracts to which they refer are about, and are referred to here as *genuine* aboutness expressions or terms. In addition to the genuine aboutness terms there is a set of 'dummy' terms, which have also been generated by the system, from the same input texts. These are also compound nominals, whose constituent words have been selected from the same input texts and juxtaposed at semi-random. The precise method used to generate these expressions is described in Section 6.2.2 below.

6.2 Pre-evaluation Tasks

A number of tasks needed to be performed before carrying out the evaluation procedure *per se*. These involved standardising the number of aboutness terms used, establishing the optimal number of texts to give to each subject, generating appropriate dummy expressions, and controlling for the envisaged effects of a number of possibly confounding variables. The evaluation was planned with possible follow-up stages in mind, which would allow for an additional layer of analysis which may have been required, depending on the results obtained. This whole set of tasks, which needed to be performed in advance of the main evaluation, are referred to as *pre-evaluation* tasks, and are discussed in the remainder of this section.

6.2.1 Standardisation of Number of Aboutness Terms

The number of aboutness expressions produced by the system for individual abstracts is variable. Some abstracts have a larger number of aboutness expressions generated than others, the number generated being related to the amount of semantic linkage identified between different items in the text. There was in fact a wide range in the numbers of expressions generated for texts in the data set: the minimum number was 6, and the maximum was 120, although many of these latter were subsets of the larger expressions.

With such variation in the total number of expressions generated for the different texts, the first task was to standardise the number of aboutness expressions to be tested per text. To this end, a small pilot pre-evaluation task was necessary.

6.2.1.1 Standardisation Pilot Study

The purposes of this pilot study were threefold:

- to standardise the number of aboutness terms per text
- to ascertain the number of aboutness terms that people can judge without their performance being adversely affected by tiredness or boredom with the task
- to ascertain the number of texts which people found they could read and express opinions about, without getting too tired or disinterested.

With these aims in mind, each of five subjects was presented with the whole set of 20 abstracts along with all their associated aboutness terms but not the dummy expressions. They were asked to do the following steps for each text in turn:

- to read the text and get a good idea of what it was about
- to read all the aboutness terms which referred to the text
- to select the 5[34] terms they considered to be most representative of the aboutness of the text.

The results of this task showed a high degree of concensus regarding which were the 5 most representative aboutness terms. In cases of disagreement, the terms were selected according to those which had been 'voted for' the highest number of times. In cases where a tie between 2 or more terms resulted in there being a final set of more than 5, the selection between all terms having equally low votes was made at random, so as to give a total set of 5 aboutness terms for each text.

The next pre-evaluation task involved the generation of 5 dummy expressions for each text, to be added to the list of 5 genuine expressions for presentation to subjects in the final evaluation.

---

[34] It was hypothesised that a total of 10 terms (5 genuine plus 5 dummies) would be a manageable number for subjects to judge. Subjects who performed the evaluation task subsequently agreed that any more would have been too many to judge consistently over the data set.

6.2.2 Generation of Dummy Aboutness Terms

6.2.2.1 Status of Dummy Expressions

It could be suggested that the representativeness of the genuine terms should be compared with the key words which accompanied the abstracts in the database. Proponents of such a suggestion would say that these are the nearest approximations to the aboutness terms (being intended as indicators of content), and should therefore form the basis of any comparison. There are two main counter-arguments relevant here.

Firstly, in the context of the evaluation described here, the ordering of the genuine terms and the dummy terms is randomised, with the intention that subjects should therefore not have any way of distinguishing the two sets, other than on the basis of their representativeness. The problem with using the abstractor-selected key terms in place of these dummies would be due to their very different properties. Whereas the dummy terms are constructed to appear superficially similar to the genuine terms (in terms of length of expression and type of head), the key terms are noticeably different, having been selected from a pre-specified list by the abstrators. As such, they would have been immediately discernible from the genuine terms, and this may well have skewed the results obtained.

The second counter-argument is that the key terms supplied by the abstractors, and the aboutness terms generated by COMMIX, have different purposes. Whilst the overall aim of both is to indicate the aboutness of the texts, the key terms are aimed at classifying the texts according to a predefined structure, whereas the COMMIX-generated terms are aimed at expressing far more compacted information. In terms of the Abox/TBox analogy discussed in Chapter Two, the former present just Tbox information, whereas the latter aim at also including some Abox information.

In the absence of any equivalent representations of this sort, the dummy terms thus constitute the most sensible comparison.

6.2.2.2 Generation of Dummy Expressions

The dummy expressions were also generated by the system. These terms were produced by the random selection and juxtaposition of the same classes of words that constitute the genuine expressions: that is, nouns, verbs, nominalised verbs,

adjectives, and adverbs. These strings varied in length, with the actual length being determined by the selection of a random number up to the maximum length appearing in the corresponding list of genuine aboutness expressions. Thus, if the list of 5 genuine terms had expressions of word length up to 7, then the lengths of the dummy expressions for the same text would vary in length *n*, with *n* being selected at random from 2-7. The length *n* was varied for each of the 5 dummy expressions generated for each text. More than one expression of the same length did occur, but only when a given value for *n* had been selected more than once by the random number generator.

Once 5 dummy expressions had been produced for each text, the data to be presented in questionnaires to subjects comprised the following:

- a set of 20 short texts (the abstracts), 10 of which were related to one another in aboutness, and the other 10 unrelated
- a list of 10 aboutness terms associated with each text, 5 of which were genuine (i.e., produced by the system as aboutness terms for that text), and 5 of which were dummies.

6.2.3 Incorporating Scope for Follow-up Investigation

During the planning stage of the evaluation, the extent to which the results would support the hypothesis was not known. It was felt that if the results were to show that particular expressions were accorded particularly low ratings, then it might be interesting to see to what extent this was due to subjects finding them difficult to understand. It was decided to include a back-up question relating to the clarity of each expression at the same point as subjects were asked to record their opinions of the degree to which the expressions were representative of the aboutness of the relevant texts. If this secondary investigation turned out to be required, it would therefore be a simple matter to include this data in a subsequent analysis. In the end, the results achieved (presented in Section 6.4) were such that this planned-for investigation did not turn out to be necessary.

6.2.4 Construction of Questionnaires

The data to be tested was based on the set of 20 texts, each text having 10 aboutness terms (5 dummy terms and 5 genuine aboutness terms). These were then to be presented in the form of questionnaires to subjects, who would be asked to express their opinions about them. The specific design of the questionnaires raised the issue of how best to deal with the confounding influence which a number of variables might have on the subsequent results.

6.2.4.1 Confounding variables

There are a number of factors which needed to be considered in order to add to the clarity and meaningfulness of the results once they were obtained. In the main they concern one or more of the following: a subject-variation effect; a tiredness or boredom effect; and a learning effect. The factors which might contribute to any such confounding effects are specified below, accompanied by descriptions of the measures taken to minimise or control any effects they may have on the results.

Subject Variation

It is primarily variation across types of subjects which is referred to here, although intra-subject variability might result from, for example, the tiredness/boredom factor mentioned below.

Although uniformity of subjects was not essential to the evaluation task, an effort was made to use subjects who would approach the task with a high degree of seriousness and concentration. If subjects were to approach the task in a trivial manner, the ratings accorded to the aboutness terms would tend towards being randomly distributed, and would fail to provide an accurate appraisal of the performance of the system in generating representative aboutness expressions. The subjects used, then, were mainly postgraduate students or research/teaching staff at the University of Brighton.

Tiredness / boredom factor

The methodology used in the evaluation (as described in detail below) was that of asking human subjects to read the texts, then to read the aboutness terms, and to express their opinions as to the extent to which the aboutness expressions were representative of the aboutness of the corresponding texts. A crucial element of this methodology was to ensure that the subjects read the texts with a high degree of

enthusiasm and seriousness, since an accurate appraisal of the aboutness terms relied on subjects having gained a good mental record of what the text was about.

Tired or disinterested subjects are likely to make trivial decisions when applying ratings to the aboutness terms they are asked to judge. It was thus important to ascertain the optimum number of texts to be presented to each subject, so as to maximise the number of texts each subject would usefully process, whilst minimising the confounding effects which tiredness and boredom with the task would inevitably have on the results obtained.

In fact, an informal pilot trial of 5 subjects showed unanimous agreement that being asked to perform the task for 20 texts was too great a load on concentration, patience and time. This trial showed that subjects were happy to perform the task for 10 texts, and felt they could do so uniformly, without being affected by tiredness or boredom. It was therefore decided that the number of texts in each presentation set would be 10.

Encouraging concentration

In addition to adopting a serious approach to the task, subjects were asked specifically to approach the task in a concentrated manner. Given the difficulty of complying with a request of this sort, one of the tasks which each subject was asked to perform was to answer a question[35] (multiple choice) about each text just after they had read it and expressed their opinions. The instructions given to subjects made it clear that a question should only be read (and answered) after getting a good idea of what the text to which it referred was about. Subjects were asked specifically to avoid reading ahead on the question sheets, since reading the question in advance of reading the text may have concentrated their attention on the answer to the question and prevented them from getting a balanced idea of what the text was about.

The purposes behind this task were twofold: firstly, and primarily, it gave subjects a reason to have to concentrate on the subject matter of the text, and get a good idea of what it was about, rather than just quickly skimming it and using superficial word-recognition to judge the expressions. It provided, in other words, an incentive for subjects to read the texts in a concentrated manner. The second, and secondary,

[35] These questions are listed in Appendix 2.

reason for the task was that it provided a record of subjects' answers to the questions, which was considered to be potentially useful as a confirmation of any particular subject having misunderstood the text (and who may consequently have given particularly atypical opinions). The actual answers given to these questions, then, were not of particular importance except as a possible subsequent indication of any atypical behaviour.

Learning Effects

One of the main possibly confounding factors to be aware of was that due to a learning effect, by which subjects might become successively more adept at the task, and increasingly familiar with (and perhaps accepting of) the type of expressions used. If the order of presentation of texts to subjects were always the same, and there was such a learning effect, then we would clearly expect the results to be skewed in favour of texts occurring towards the end of the presentation sets (provided that sets were not so long as to bring about a tiredness effect). In order to avoid such uncontrolled skewing of results, texts were presented in random order (but see remainder of this section in relation to a controlled investigation of a learning effect).

a) Order of presentation of texts in questionnaires

Whilst wanting to minimise the uncontrolled effects of task-familiarisation on the results, it was felt that it would be interesting to investigate the degree to which familiarisation might affect the results. One aspect to be investigated here is that of familiarisation with the task itself: whether subjects began to find the expressions more representative of aboutness as their familiarity with the task and the type of expression increased. The other aspect of familiarisation to consider is that associated with the aboutness of the texts: whether having read a text with similar (related) aboutness might affect subjects' subsequent performance. We could say, then, that the texts vary according to the variable 'relatedness', by which it is meant that the text belongs to either the set of 10 with related aboutness, or the set of 10 which are unrelated to one another.

In order to investigate whether there is such a learning effect due to the variable 'relatedness', half of the texts presented to each subject were related to one another in terms of their aboutness, and the other half were unrelated.

Since subjects were presented with just 10 texts each (as discussed above), this meant that a particular subject was presented with 5 from each group. The texts were presented either related set first, or unrelated set first, although within these subgroups the ordering was random.

b) Ordering of terms after texts

The learning effect described above in relation to the order of presentation of texts, also applies to the ordering of aboutness terms as they are presented, following the texts to which they refer. The terms were therefore placed in independent random order, with no distinction between dummy terms and genuine terms, and were listed on the questionnaires in this random order, following the text to which they referred.

6.2.4.2 Final Form of Questionnaires

The total set of 20 texts comprised 10 with related aboutness and 10 with unrelated aboutness. The individual texts were numbered from 1-20, with numbers 1-10 having related aboutness, and numbers 11-20 being unrelated to one another.

The number of subjects required to perform the task is related to the number of texts in the data set (20), the observation that each subject could only reasonably be expected to process half of the total number (i.e., 10), and the constraint that half of the texts given to each subject (i.e., 5) were to be from the related group, and half unrelated.

The different permutations of text numbers and orderings required the minimum number of subjects to be 8. Subjects were referred to by letter, from A to H. The specific allocation of text numbers to different subjects is shown in Table 6.1 below.

Subject Ref:	A	B	C	D	E	F	G	H
Text Nos:	1-5	11-15	6-10	16-20	1-5	16-20	6-10	11-15
followed by	11-15	1-5	16-20	6-10	16-20	1-5	11-15	6-10

Table 6.1. Texts Processed by the Different Subjects

The ordering of each group of 5 texts was randomised for each subject reference: this was also the case for the aboutness terms, which therefore appeared in random order after their relevant texts. As discussed above, adopting these permutations of orders of presentation of texts facilitated the possible subsequent investigation of whether there was any significant learning effect when subjects had already met a series of texts which were related to one another in aboutness.

An example of the final questionnaire (as prepared for subject A) appears in Appendix 3. The appendix also contains the task description and instructions given to each subject along with their questionnaire, and the multiple choice question sheet which relates to subject A's text order.

6.3 The Evaluation Procedure

Two groups of 8 subjects were consulted. This therefore gave duplicates of each of questionnaires A to H, and these were consequently labelled A1-H1 and A2-H2.

6.3.1 The Evaluation task

The time taken for subjects to perform the task for the whole set of texts which they were given was not considered to be a confounding variable, and there was no need to associate a time limit governing the procedure. Subjects were therefore given the questionnaires, and asked to return them on completion.

Subjects reported that the whole task took between 50 and 60 minutes, and there was notable consistency across all subjects asked (although 2 subjects returned the questionnaires by mail, and did not comment on the time taken).

Each subject was presented with an instruction page which gave general guidelines as well as a specific description of the task, a questionnaire, and a set of questions, comprising 1 multiple choice question per text. Each questionnaire comprised a set of 10 texts, each of which had its own set of 10 aboutness expressions associated with it: 5 genuine (generated by COMMIX) and 5 dummies.

Subjects were asked to do the following for each text and its list of expressions, one text at a time[36]:

- read the text carefully, until they had a good idea of what it was about

[36] The full task description and instruction sheet appears in Appendix 3.

6.4 Results and Discussion

Each of the 16 subjects recorded their opinions about 10 texts (half the total set), each of which had 10 aboutness terms. The data obtained from the completed questionnaires, therefore, consists of 1600 numbers between 0 and 4 inclusive, which are the ratings accorded in respect of the degree to which each aboutness term was judged to represent what its associated text was about.

There are three main aspects to the analysis of the results. Firstly, there is an overall assessment of the difference in general, across all subjects, between the ratings accorded to the dummy terms and those given to the genuine terms. This analysis in effect examines the difference between the two populations of expressions: the genuine terms on the one hand, and the dummy terms on the other. The second aspect to the analysis concerns the rating totals and medians, analysed according to the different texts. The third aspect relates to the range and distribution of the actual ratings accorded to the different populations of expressions (i.e., to the genuine and dummy terms). In discussing the results of each of these three aspects, the notable findings are highlighted in bold typeface.

6.4.1 Difference between Genuine Term and Dummy Term Populations

This section deals with the extent to which the hypothesis presented in Section 6.1.3 holds true. It shows that the genuine expressions (those generated by COMMIX as aboutness terms) are judged by subjects to be very significantly more representative of the aboutness of their corresponding texts than are the dummy expressions, which are also generated by the system for the same texts. This hypothesis is shown to be supported very strongly.

The selection of a statistical test to express the difference between the two populations of ratings was influenced by the fact that the data are ordinal and that the subjects are independent (they processed different sets of texts and expressions). In these circumstances, the Mann-Whitney test is the most appropriate test of the significance of the results.

6.4.1.1 Mann-Whitney Test

The Mann-Whitney test was performed on the data in order to test for a difference between the ratings accorded by subjects to the genuine terms, and those accorded to the dummy expressions. The test was performed across the whole set of texts, and yielded **a *p* value of <0.0001**, which is highly significant.

On the basis of this result we can say with a high degree of certainty that the difference between the ratings accorded to the dummy terms and the genuine terms is due to factors other than chance. The results thus support the hypothesis that ratings accorded to the genuine terms are very significantly different from those accorded to the dummy terms, and that this difference is therefore due to the nature of the expressions. For each of the 20 texts, **the medians of the genuine terms were consistently higher than those obtained from the dummy terms**. Table 6.2 in section 6.4 below shows the individual medians, which are discussed at greater length at that point.

Calculating the *p* values for each individual text showed a **significant difference in each case**, with the **genuine terms being consistently judged as more representative** than the dummy terms. The least significant difference between dummy and genuine term ratings was found to be that associated with Text 17, for which the *p* value was 0.0405. Although this is close to the cutoff point for significance ($p<0.05$), it shows that even this poorest result confirms that there is a significant difference between the two sets of expressions which is not simply attributable to chance. Text number 8 shows the next smallest significance of difference, with $p=0.03$), and in each case the genuine terms were judged more representative than the dummy terms.

The actual size of the differences between ratings accorded to genuine and dummy terms is presented in Figure 6.1 below, expressed as *rating totals*. The rating total for a particular text applies to the accumulated ratings accorded by all subjects to either the set of 5 genuine terms, or to the set of 5 dummy terms. For a particular text, then, each of 5 terms could have been given a maximum rating of 4, giving a maximum possible rating for the set of 20. Taken across the 8 subjects who judged each text, this gives a maximum rating number of 160 for each set of 5 terms, either dummy or genuine.

Figure 6.1. shows, for each text, the actual size of the *difference* between the rating totals for the genuine versus the dummy expressions. Since the genuine terms scored higher on each occasion, the difference is shown as the 'genuine' total minus the dummy total. Although the *y* axis is shown extending to 100, the scale clearly extends to a maximum difference of 160, which would result for a text if all its genuine terms were given maximum ratings by all subjects, and all its dummy terms were given zero ratings. In a similar vein, a rating difference of 0 for a particular text would mean that the genuine and dummy terms had fared equally well or badly.

The chart clearly shows that there are no texts for which this is the case, and the absence of any negative values indicates that fact that the genuine terms always had higher ratings than the dummies. We can also see that Text 17 has the smallest actual difference between sets of terms, and Text 6 has the largest difference.

It is interesting to note at this stage that the 6 largest differences (in decreasing order) occur in Texts 6, 4, 5, 12, 16 and 15, this being 3 from each of the related and unrelated sets of texts[1] (related in terms of what the texts are about). Similarly, the 2 smallest differences, seen in Texts 17 and 8 are 1 from each of the subsets of texts.

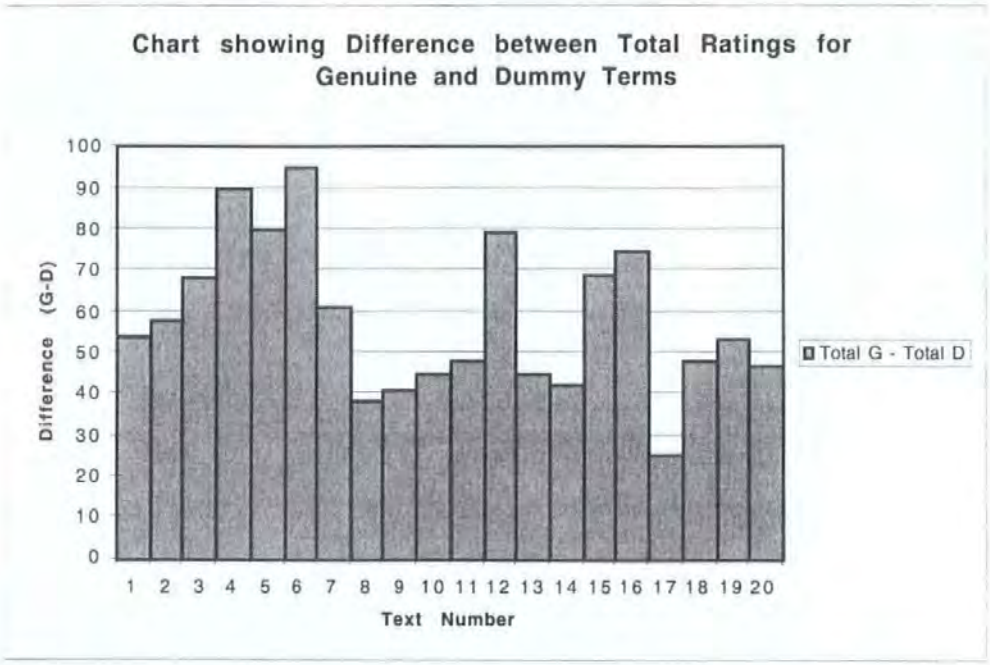


Figure 6.1. Chart Showing Actual Differences between Totals for Dummy and Genuine Terms, for each Text

[1] The reader may remember that Texts 1-10 were unrelated in terms of their aboutness, whereas Texts 11-20 were related.

Whilst it is interesting to see the range of differences between the dummy and genuine term ratings, the most important fact to note is that the difference between the two sets of terms is significant for each individual text in the data set, with the genuine terms having consistently higher ratings. Texts for which the significance of the difference is lowest and highest are of interest, particularly in the light of the findings presented in the ensuing sections of this chapter. Some possible reasons for these differences are discussed in Chapter Seven, which contains a broader scope of discussion, including matters relating to the performance of the system and the data set itself.

6.4.2 Rating Totals and Medians for Different Texts

Given the highly significant difference between the ratings accorded to the two sets of terms, across the whole set of texts, it is interesting to look at the total ratings, across subjects, but for individual texts. Therefore, for each text, the ratings accorded by all subjects to the 5 genuine, and the 5 dummy terms were respectively totalled as described in the preceding section, and both the total and the median recorded, with each set having a maximum total rating of 160 (as detailed above).

These accumulated ratings appear in Table 6.2. This shows, for each text, the actual totals of all ratings for both dummy terms (D-Total) and genuine terms (G-Total), for each text. It also shows this amount as a percentage of the maximum total rating which could have been accorded (i.e., if all subjects had given all terms the maximum rating of 4). The median rating per text, averaged across subjects, is also given.

As an example, consider Text number 6. The rating total for the dummy terms combined is only 21 (out of a possible 160), whereas the total for the genuine terms is 116 which, as shown, is 72.5% of the maximum possible ratings. The table also shows that the median combined rating per subject is only 1.5 for the dummy terms, and 14 for the genuine set (each of the medians being on a scale of 0-20).

Text	D-Total	D-Total as % of max.	D-Median	G-Total	G-Total as % of max	G-Median
1	17	10.63	2	71	44.38	10
2	16	10.00	1	74	46.25	10
3	13	8.13	1	81	50.63	10
4	22	13.75	1.5	112	70.00	15
5	18	11.25	2.5	98	61.25	14
6	21	13.13	1.5	116	72.50	14
7	19	11.88	2	80	50.00	8
8	35	21.88	2.5	73	45.63	8
9	12	7.50	0.5	53	33.13	5.5
10	31	19.38	3	76	47.50	9
11	29	18.13	3	77	48.13	9.5
12	12	7.50	1	91	56.88	12
13	30	18.75	3	75	46.88	8
14	32	20.00	4	74	46.25	10
15	14	8.75	0.5	83	51.88	9
16	29	18.13	3	104	65.00	15
17	26	16.25	2.5	51	31.88	7
18	6	3.75	0	54	33.75	7.5
19	25	15.63	4	78	48.75	11.5
20	23	14.38	2	70	43.75	9.5

Table 6.2. Rating Totals, Percentage Totals and Medians for Dummy (D) and Genuine (G) Terms

Expressing the total ratings as percentages of the maximum possible gives us a good idea both of the success of the expressions in representing aboutness, and provides a benchmark for comparison. These percentage ratings, showing the performance of sets of dummy terms and genuine terms for the different texts, are more clearly represented in chart form. Figure 6.2 shows, for each different text, the totals of all the ratings accorded to the dummy expressions and to the genuine expressions, expressed as percentages of the maximum possible total ratings.

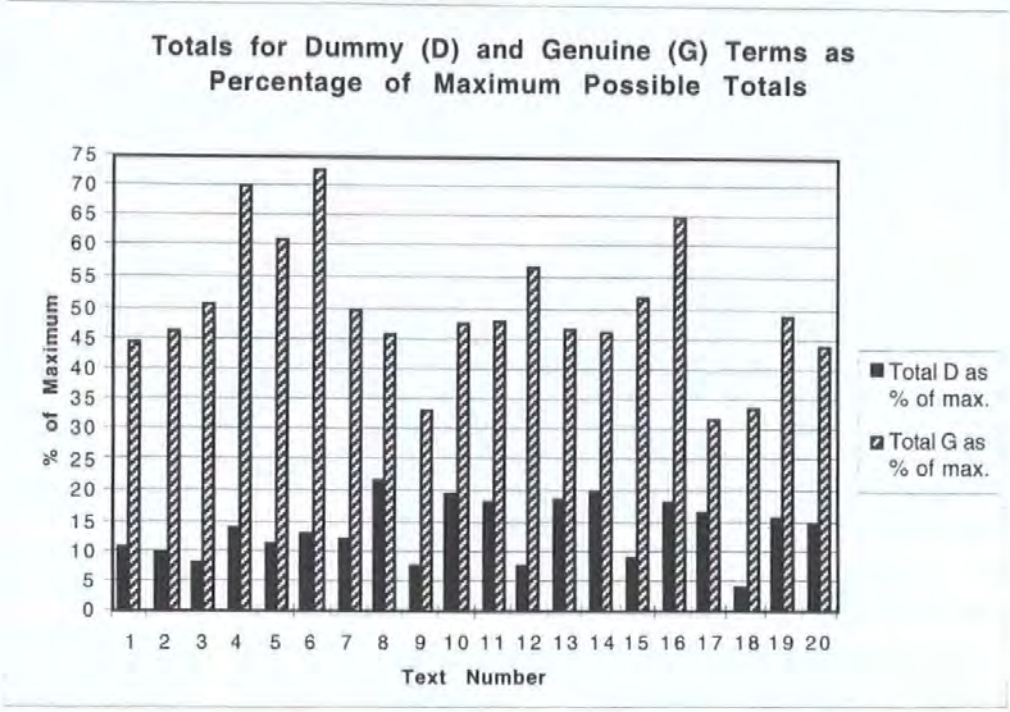


Figure 6.2. Chart showing Actual Rating Totals as Percentages of Maximum Possible Totals, for Dummy and Genuine Terms

6.4.2.1 Percentage Attainment Values: Rating Totals as Percentages of Maximum Possible Totals

We introduce here a formal expression of this notion of the actual rating total as a percentage of the maximum possible total which could be assigned. The term to be used for this measure is *Percentage Attainment Value,* which may be abbreviated to *PAV*. The concept, which was introduced in Section 6.1.2, is straightforward. The accumulated rating total of a set of aboutness terms (be they genuine or dummy terms) which relate to a particular text may be expressed as a percentage of the maximum possible accumulated total. The degree to which a set of aboutness terms attains this maximum possible level is expressed as its PAV level.

The general case, then, can be described by the formula:

$$PAV = \frac{\sum_{s}\sum_{t}^{n} R}{n \times R_{max} \times s} \times 100$$

where R is the actual rating for a term t,

n is the number of terms in the set relating to a given text,

R_{max} is the maximum rating on the scale,

s is the number of subjects who have assessed that set of terms.

It should be clear that the maximum possible total would result if every subject selected the highest rating number for each of the aboutness terms they judged in relation to a particular text. In our case, each term has an R_{max} of 4, and $s = 8$. As we have already said, this gives a maximum possible accumulated rating of 160 per set of either dummy or genuine terms, for a particular text. It should also be clear that the measure can additionally be used in relation to *individual* aboutness terms (rather than sets of 5). In this case, clearly n becomes 1 rather than 5, giving a maximum possible accumulated rating of 32 per expression.

Figure 6.2 expresses the total ratings for sets of 5 each of dummy and genuine terms, calculated according to the above formula, and therefore accumulated across all subjects. It thus shows the totals for each text as PAVs (i.e., as percentages of the maximum possible accumulated rating of 160). The 100% rating would clearly be the ideal case for all the genuine expressions, and would mean that all the terms had been accorded the highest possible ratings by all subjects. Conversely, the ideal case for the dummy expressions would be 0.

Figure 6.2 clearly shows that all of the genuine terms had an accumulated rating of >30% of the maximum possible ratings, with only 4 texts (numbers 1, 9, 17 and 18) falling below 45% of the maximum rating. These percentages of the maximum possible accumulated rating numbers constitute the PAV levels discussed above. As an example, consider the case of Text 7: here we obtain a PAV of 50% for the genuine expressions. This is interpreted as meaning that for this text the combined opinion across all subjects was that the genuine aboutness terms attained 50% of the maximum rating possible.

Table 6.3 below expresses the number of texts for whose genuine aboutness terms the ratings come above a variety of PAV levels. Rather than expressing an individual PAV for each text, the major groupings into attainment levels of 30%, 45%, 60% and 70% are given. It should be noted that there is nothing special about these particular attainment levels, but they constitute useful benchmarks for expressing the general findings of a PAV analysis of this data. These groupings express the number of different texts whose genuine aboutness terms attained *at least* the benchmark PAV levels, which therefore constitute a measure of the minimum level of attainment. This Table also expresses (in the last column) the number of texts falling into these benchmark categories as a percentage of the data set. Thus, for example, if the genuine terms for 15 of the 20 texts reached a PAV of 45%, then we would record this as 75% of the texts having a PAV of at least 45%.

% Attainment Value (PAV)	No. Texts attaining PAV (out of set of 20 texts)	No. Texts as % of data set
>= 30 %	20	100 %
>= 45 %	15	75 %
>= 60 %	4	20 %
>= 70 %	2	10 %

Table 6.3. Percentage Attainment Values for Combined Ratings of Genuine Terms, for whole Text Set

We can see from this Table that **75% of the texts reached a PAV level of at least 45%** of the maximum possible ratings which could have been accorded to its genuine aboutness terms.

Figure 6.2 also shows the PAV levels for the groups of dummy expressions for the whole text set. This shows that, with one exception, all the texts had **PAV levels for the dummy expressions of less than 20%** (with the exception being Text 8, with a PAV of 21.88%).

6.4.2.2 Medians of Ratings for Dummy versus Genuine terms

The chart in Figure 6.3 shows the distribution of median ratings, across subjects, for the different texts. The median of each set of 5 aboutness terms was taken across subjects. Each is therefore on a scale between 0 and 20, since 20 is the maximum

rating total which any one subject could give to any group of 5 dummy or genuine terms.

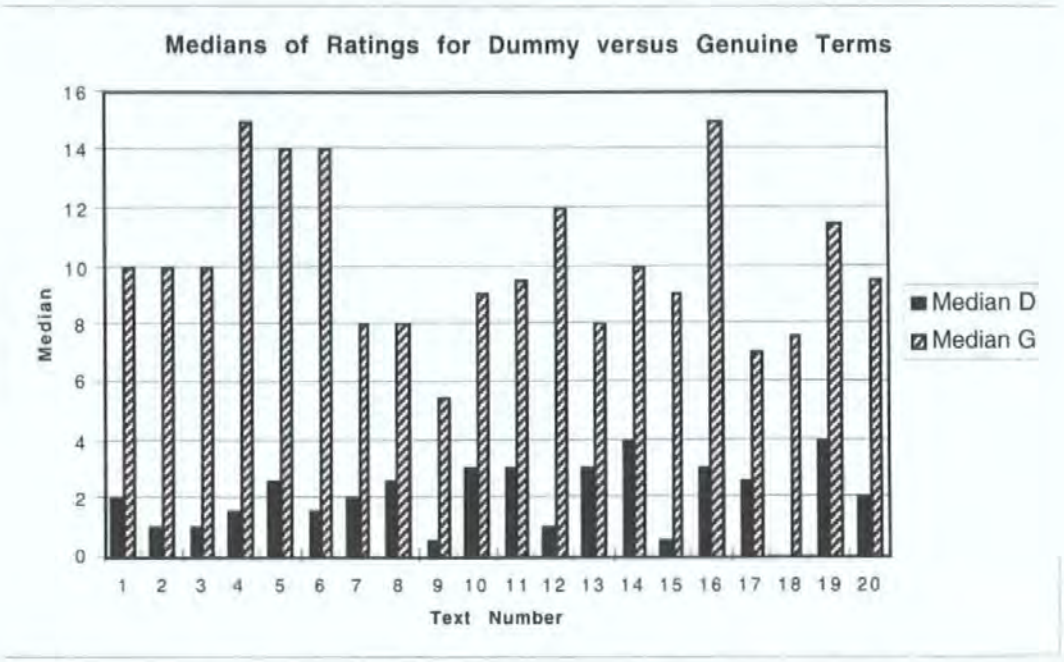


Figure 6.3. Chart showing Medians of Ratings for Dummy versus Genuine Terms, across subjects, for different texts

It is clear that Texts 4 and 16 jointly have the highest median rating for their genuine terms, with Texts 5 and 6 close behind. The text with the lowest median rating is Text 9, followed by Texts 17, then 18 as the next lowest. The size of the median is an indication of the success of the genuine terms at representing the aboutness of the text to which they refer. We can therefore conclude that on the basis of their medians, Texts 4 and 16 are those whose genuine aboutness terms best represent what they are about. Similarly, we can say that Text 9 is the one whose genuine aboutness terms least represent what it is about, although we should remember that the difference seen is significant for the pairs of sets relating to each text.

6.4.2.3 Comparison of PAVs with Medians and Differences between Dummy and Genuine Terms

A comparison of Figure 6.2 and Figure 6.3 shows that the four texts with the highest PAVs for their genuine terms are the same as the four best attainers as expressed by their medians. These are Texts 6, 4, 16 and 5 (in this decreasing order of attainment according to PAV levels, with the order being Texts 16 and 4 equally,

followed by 5 and 6 equally according to the median scores). The fifth best attainer is also the same in both cases, this being Text 12.

Similarly, the 3 texts having the lowest PAVs are the same as the 3 having the lowest medians. These are, in ascending order, Texts 17, 9 and 18 according to their PAV levels, which are in the ascending order of 9, 17 and 18 according to their medians.

According to both the PAV levels, and the median ratings, then, the texts which produced the highest ratings for their genuine aboutness terms were Texts 4, 5, 6 and 16, whilst Texts 9, 17 and 18 produced the lowest. These findings also coincide partially with the ordering relating to the size of differences between dummy and genuine terms (as shown in Figure 6.1, Section 6.4.1). The largest difference between the ratings occurs for Text 6, decreasing through Texts 4, 5, 12 and 16, with the smallest differences being noted for Texts 17, 8 and 9 (in ascending order).

We can see that there is a set of 5 texts which are consistently the highest attainers according to all three criteria: the PAV level, the median and the difference between dummy and genuine ratings, although the orderings differ for the different criteria. Similarly, Texts 17 and 9 are consistently found at the lower end of the scale for each of these same criteria (even though Text 9 was found at the third lowest position according to the criterion of difference between its dummy and genuine ratings). The highest 5, then, are Texts 4, 6, 16, 5 and 12; the lowest performers are Texts 17, 9 (and 18).

The next section deals with the findings associated with the distribution and frequencies of ratings accorded to individual terms (rather than accumulated totals of ratings), after which the final section compares all the findings and presents some generalised results. These results are discussed further in Chapter Seven, which looks at the findings both in relation to the particular texts, and in regard to the aboutness terms themselves.

6.4.3 Distribution of Different Ratings

This section looks at the actual rating values accorded to the genuine and dummy expressions, in terms of the frequencies of use of the different rating numbers. We begin by looking at the overall percentages of dummy terms and genuine terms which were accorded each of the possible rating values.

6.4.3.1 Ratings Accorded to all Genuine and Dummy Terms

Table 6 below shows the overall percentage of dummy and genuine terms which were accorded the different rating numbers by subjects. These figures are combined to cover all the ratings for all texts and by all subjects.

Rating Number	% of Dummy Terms	% of Genuine Terms
0	64.63	13.50
1	23.13	26.25
2	8.50	24.63
3	2.38	19.13
4	1.38	16.50

Table 6.4. Percentage of Dummy and Genuine Terms Accorded Different Ratings

Table 6.4, and Figure 6.4, which represents it, show that whereas **almost 65% of the dummy terms were accorded a rating number of 0, 60.25% of the genuine terms were rated at 2, 3 or 4.**

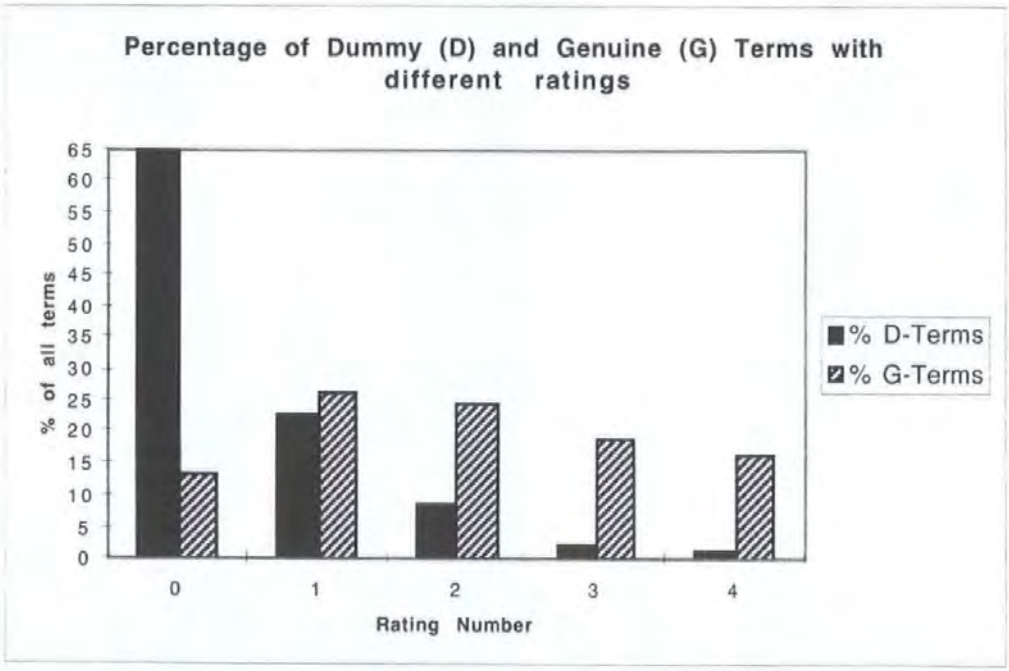


Figure 6.4. Chart showing Percentages of all Dummy (D) and Genuine (G) Terms with Different Rating Values

Interpreted into the words used to define the meanings of these numbers (and presented to subjects in their questionnaires), we can say that subjects were of the opinion that **60.25% of the genuine aboutness terms were either reasonably, very, or extremely representative of what the related text was about**. We can also say that almost **36% of the genuine terms** were judged as being in the top 2 categories of representativeness: **either very, or extremely representative**.

In contrast, this data shows that almost 88% of the dummy terms were judged as being either not representative at all of the aboutness of their texts, or only slightly representative.

With this overall picture in mind, in which **over 60% of the genuine terms are judged as being reasonably, very, or extremely representative of the aboutness of the text**, we turn in the next section to look just at the genuine terms, and the distribution of ratings accorded to them across subjects, but for particular texts.

6.4.3.2 Ratings Accorded to Genuine Terms of Different Texts

We concentrate here on the ratings associated with the genuine aboutness terms generated by COMMIX. Table 6.5 gives the frequencies with which the different rating numbers were accorded to the genuine terms of the different texts. These frequencies represent the number of times a subject selected that particular rating number to express the degree to which it represented the aboutness of its related text. Since each text was process by 8 subjects, and each subject had to record an opinion for each of 5 genuine aboutness terms, the frequencies for terms rated between 0 and 4 sum horizontally to 40 (note the final column is a combination of terms rated 3 and 4).

We can use the information presented in this table most usefully to compare the performance of the texts with particularly high and low ratings respectively. In this regard, it is useful to look at the frequencies associated with ratings 4 and 3, since these represent opinions that the aboutness terms were, respectively, extremely and very representative of the aboutness of the corresponding texts.

The final column of Table 6.5 shows the combined frequencies of terms accorded ratings of 3 and 4. We can see that terms associated with Texts 6, 4 and 16 each had over half of their aboutness terms rated as either very or extremely representative. Texts 5 and 12 are also noteworthy in this respect.

At the other end of the performance scale we see that Texts 17 and 18 both have frequencies of just 6 out of the 40 terms judged as being very representative of their aboutness.

Text No.	Terms Rated 0	Terms Rated 1	Terms Rated 2	Terms Rated 3	Terms Rated 4	Terms Rated 3 or 4
1	5	14	10	7	4	11
2	5	13	8	11	3	14
3	5	10	11	7	7	14
4	1	7	7	9	16	25
5	3	6	12	8	11	19
6	1	8	5	6	20	26
7	6	6	14	10	4	14
8	6	15	6	6	7	13
9	14	10	8	5	3	8
10	5	13	10	5	7	12
11	7	12	8	3	10	13
12	4	6	12	11	7	18
13	1	16	14	5	4	9
14	9	6	12	8	5	13
15	6	10	8	7	9	16
16	1	7	8	15	9	24
17	10	15	9	6	0	6
18	8	16	10	6	0	6
19	7	5	15	9	4	13
20	4	15	10	9	2	11

Table 6.5. Frequencies of Different Ratings accorded to Texts (for Genuine Terms)

6.4.3.3 Ratings Accorded to Unrelated versus Related Sets of Texts

Before summarising the findings of this evaluation, it is worth referring briefly to the distinction made between the performance in respect of texts belonging to the subset having related aboutness (Texts 11-20), and those with unrelated subject matter (Texts 1-10). As discussed earlier, there was the possibility that subjects' performance

would be improved by having previously processed texts with similar subject matter.

Table 6.6, along with its associated graphic representation in Figure 6.5, shows the percentages of genuine and dummy terms from each of the related and unrelated text sets, which were accorded the different ratings.

Rating number	G- R %	G - U %	D-R %	D -U %
0	14.25	12.75	62.25	67.0
1	27.00	25.50	23.25	23.0
2	26.50	22.75	10.50	6.5
3	19.75	18.50	2.75	2.0
4	12.50	20.50	1.25	1.5

Table 6.6. Percentages of Genuine (G) and Dummy (D) terms, from Related (R) and Unrelated (U) texts, showing different rating numbers.

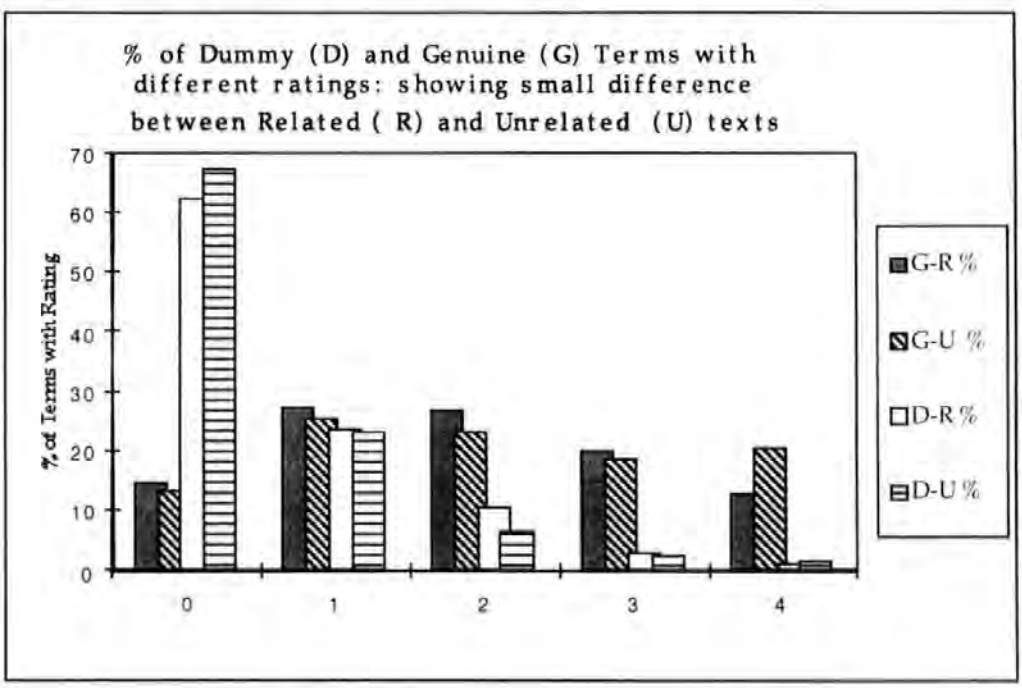


Figure 6.5. Chart showing Percentages of Dummy and Genuine Terms having different Ratings, for Related (R) and Unrelated (U) text sets.

As the second column of Table 6.6 shows, for the genuine terms pertaining to the texts from the *related* set (i.e., Texts 11-20), that 14.25% of them were rated at 0. The last 3 rows of this column inform us that a total of 58.75% of the genuine terms

pertaining to the related set were accorded ratings of 2, 3 or 4 and were therefore judged as being either reasonably, very or extremely representative of the aboutness of their corresponding texts.

The same data for the genuine terms pertaining to the set of *unrelated* texts appears in the third column of Table 6.6. This shows only a difference of 1.5% in the number of terms rated at 0, and difference of only 3% for the terms rated at 2, 3 or 4 (with a total of 61.75%).

Figure 6.5 shows at a glance the degree of similarity between the pairs of unrelated and related terms for each of the genuine terms and the dummy terms.

It does seem, therefore, that **any learning effect due to familiarity with the aboutness of the texts was not sufficiently large to affect subjects' behaviour**.

This observation is supported by the results of the **Mann-Whitney test**, which showed that **texts from the related and unrelated sets were equally distributed** amongst those with the largest differences between genuine and dummy ratings, and those with the smallest.

6.4.4 Summary of Results

The data obtained from the evaluation described in this chapter has been analysed according to three criteria. Firstly, the emphasis was on the difference between ratings for the genuine and dummy terms. Secondly, we looked at the rating totals, for which the notion of percentage attainment values was introduced, and at the medians, with both being taken across subjects and in relation to particular texts. The third aspect of the analysis was in relation to the distribution of the actual ratings given to the genuine terms in particular.

The analysis of the data using the Mann-Whitney test has indicated that there is a highly significant difference ($p < 0.0001$) between the ratings accorded to the genuine aboutness terms and those accorded to the dummy terms. It has shown that the difference is significant for each individual text, and that in all cases the genuine terms were judged as consistently more representative of the aboutness of their texts than were the dummy terms.

The analysis of percentage attainment values in Section 6.4.2.1 has shown that 75% of the texts reached a PAV level of at least 45% of the maximum possible ratings which could have been accorded to its genuine aboutness terms.

A set of 5 texts has been identified, which are consistently the highest attainers according to the criteria of PAV level, median, and the difference between dummy and genuine ratings. These are Texts 4, 5, 6 ,12 and 16. Conversely, a set of 3 notably low attainers has also been identified: these being Texts 9, 17 and 8.

Even though there are some texts whose genuine terms were accorded relatively low ratings, the different analyses described in this chapter have supported the hypothesis that the genuine aboutness terms are rated very significantly higher than their dummy counterparts. We have also shown positive results which indicate that over 60% of the genuine terms were considered to be either reasonably, very, or extremely representative of the aboutness of the texts for which they were generated.

6.5 Summary of Chapter

This chapter has described in detail the evaluation methodology used to assess the performance of COMMIX. It has discussed the measures of precision and recall, and has explained why these metrics are not suited to an evaluation which claims to assess the degree to which aboutness expressions succeed in representing their corresponding texts. We have introduced instead the notion of *percentage attainment value*, which describes the actual performance of a term, or set of terms, expressed as the degree to which it attains its maximum possible potential, as judged by human subjects, according to a number of different rating values.

The chapter has included a detailed discussion of the issues involved in the planning and execution of the evaluation, which took the form of a questionnaire given to subjects. The results obtained have been analysed in three mutually compatible ways, and each has shown encouraging results. We have seen a highly significant difference between ratings accorded to the genuine terms and the dummy terms, and in all cases have obtained significantly higher ratings for the genuine terms. Over 60% of the genuine terms were considered to be either reasonably, very, or extremely representative of the aboutness of the texts for which they were generated. The PAV analysis has shown that 75% of the texts reached a PAV level of at least 45% of the maximum possible ratings which could have been accorded to its genuine aboutness terms.

We take these results as being encouraging, particularly in the light of the simplicity of the implementation of COMMIX at its basic level of performance.

The next and final chapter discusses the performance of the system, both in the light of the findings presented here, and in regard to its intentional status as a 'bare bones' approach to the generation of aboutness expressions to fill the 'essence' gap reported in Chapter Two. It discusses the findings in relation to characteristics of some of the texts, as well as to specific aboutness terms. It also discusses some possible future extensions and pinpoints some interesting areas for further research.

Chapter Seven

7. Discussion, Future Directions and Conclusion

The scope of this final chapter is threefold: firstly, we continue from Chapter Six, discussing the implications of the findings of the evaluation described therein. Section 7.2 concerns the performance of the system and comprises two parts: we look first at the actual performance of the system, in the light of the intentional simplicity of the approach. A number of possible add-ons, which would be expected to enhance the performance of the system, are then discussed, after which some possible future directions which this work could take are suggested. This leads to the final section, in which we discuss the work in relation to the original aims, and comment on the extent to which these aims have been fulfilled. The chapter ends with a general conclusion of the value of the work described in this thesis, and the contribution it makes to the general field of content specification and generation, which is seen as an essential part of any IR system.

7.1 Implications of the Evaluation

Chapter Six presented the evaluation methodology and the results, and gave some interpretation of the findings. In addition, it identified texts whose aboutness terms fared particularly well, and others whose terms performed particularly badly in the rating exercise. It was found that the aboutness terms associated with Texts 4, 5, 6, 16 and 12 fared particularly well, and that those associated with Texts 9, 17 and 18 fared notably badly. Given that the design of the evaluation accomodated for the possible confounding effects of other variables (such as subject tiredness, subject variation and any learning effect), a particularly poor or good performance may be said to be related to properties of either the texts or the aboutness terms generated for them, or both.

The current chapter begins by looking at the characteristics of these aforementioned texts and their genuine aboutness terms, and suggests possible reasons for the different levels of performance. Note that in the ensuing sections of this chapter, the expression 'aboutness terms' refers by default to just the genuine terms associated

Chapter Seven

7. Discussion, Future Directions and Conclusion

The scope of this final chapter is threefold: firstly, we continue from Chapter Six, discussing the implications of the findings of the evaluation described therein. Section 7.2 concerns the performance of the system and comprises two parts: we look first at the actual performance of the system, in the light of the intentional simplicity of the approach. A number of possible add-ons, which would be expected to enhance the performance of the system, are then discussed, after which some possible future directions which this work could take are suggested. This leads to the final section, in which we discuss the work in relation to the original aims, and comment on the extent to which these aims have been fulfilled. The chapter ends with a general conclusion of the value of the work described in this thesis, and the contribution it makes to the general field of content specification and generation, which is seen as an essential part of any IR system.

7.1 Implications of the Evaluation

Chapter Six presented the evaluation methodology and the results, and gave some interpretation of the findings. In addition, it identified texts whose aboutness terms fared particularly well, and others whose terms performed particularly badly in the rating exercise. It was found that the aboutness terms associated with Texts 4, 5, 6, 16 and 12 fared particularly well, and that those associated with Texts 9, 17 and 18 fared notably badly. Given that the design of the evaluation accomodated for the possible confounding effects of other variables (such as subject tiredness, subject variation and any learning effect), a particularly poor or good performance may be said to be related to properties of either the texts or the aboutness terms generated for them, or both.

The current chapter begins by looking at the characteristics of these aforementioned texts and their genuine aboutness terms, and suggests possible reasons for the different levels of performance. Note that in the ensuing sections of this chapter, the expression 'aboutness terms' refers by default to just the genuine terms associated

with the text/s under discussion, since it is primarily these that are the focus of attention.

7.1.1 Differences between Texts

The findings of Chapter Six showed that the expressions relating to Texts 4, 6 and 16 were the top three performers, based on both their median rating values and their PAV levels. These 3 texts each had over half of their aboutness terms rated as either *very* or *extremely* representative. Texts 5 and 12 were also found to be noteworthy in both these respects. The poorest performance occurred in association with Texts 9 and 17, followed by Text 18 as the third lowest performer.

This section concerns some possible explanations for the particular texts producing the results they did, and includes some discussion on the characteristics of the high and low-performing texts.The following sub-sections discuss each of the high and low-performing texts and terms specifically in relation to both the text style and to the amount of linkage occurring between its items. To some extent the comments represent 'gut reactions' to the texts, which have been expressed by some of the subjects, and are shared by the author. The discussion of style of texts in this section, then, is speculative, and reflects some general points about the texts, but it is not based on a formal analysis of the texts (see Future Directions below).

The comments which relate to the amount of linkage refer either to the frequency of links at the L1 level[38], or at the L2 level, or both combined, and refer to **Table 7.1** below, and to Appendix 4. Appendix 4 lists, for each text, all the words which become linked during its processing by COMMIX, along with the number of links in which each is involved.

Table 7.1 shows a summary of the information presented in Appendix 4, and presents, for each text, the number of links found between words at the different levels. As Appendix 4 makes clear, some of the words feature in both lists, but with different frequencies, since amounts of linkage differ at the different levels. The reader is reminded that the full set of texts and terms appear in Appendix 1, to which reference should be made over the course of the ensuing discussion.

---

[38] The different levels of linkage referred to here are discussed in detail in Chapter Four.

Text No.	No. L1 Links	No. L2 Links	Total No. L1 + L2	PAV Level (%)
1	20	20	40	44.4
2	18	64	82	46.3
3	75	80	155	50.6
4	56	80	136	70.0
5	12	26	38	61.3
6	105	139	244	72.5
7	145	147	292	50.0
8	4	12	16	45.6
9	14	40	54	33.1
10	186	265	451	47.5
11	14	22	36	48.1
12	50	110	160	56.9
13	108	122	230	46.9
14	68	176	244	46.3
15	68	144	212	51.9
16	50	92	142	65.0
17	34	44	78	31.9
18	26	38	64	33.8
19	10	36	46	48.8
20	88	131	219	43.8

Table 7.1. Number of Links occurring during Processing of Different Texts (also showing PAV level).

7.1.1.1 High performers

In the ensuing discussion which relates to the characteristics of particular texts, we refer to their *readability*, but use this criterion in an informal way. It should be stressed that the term is not used in its formal sense (Klare, 1976), where it gives measures, typically, of syntactic and semantic complexity. The positive use of this

term is intended to convey the subjective view that the text is *easy to read*, and does not consist of what seem to be unrelated sentences with different topics following on from one another.

Text 4

Text 4 is easily readable, and is typical of expository text, in that it successively builds on the information that has gone before (as discussed in Chapter Three), allowing the reader to get a clear idea of what the text is about. We see none of the topic-hopping that occurs in the lower-performing Text 18.

In the case of Text 4, the amount of linkage, as shown in **Table 7.1**, is neither particularly high nor low, with its position in the linkage table being about half way down the list of decreasing amounts of linkage. However, it is notable (as can be seen in Appendix 4) that the lexical items which have contributed most highly to the linkage are 'development', 'advances' and 'considerations'. If we then look at these terms along with their modifiers in the text, we see that the combination gives a good picture of what the text is about. Whether or not this is related to the sense of integrity of the text, and its general description as expository in style, is an interesting point to speculate, but goes beyond the realm of this work (see Future Directions below).

Text 6

Text 6 is also clearly readable, and shows a high degree of continuity of topic throughout. There are many occurrences of lexemes which occur in more than one sentence, giving scope for a lot of linkage across sentences.

It is interesting to note that this text, which has the highest overall PAV level for its set of aboutness terms, does show a high amount of linkage, having the 3rd highest figure for total linkage. Also, as Appendix 4 shows, the words *death*, *characters* and *ratings* are involved in notably high amounts of linkage.

Text 5

Text 5, although not particularly expository in style, is an informative abstract, being fairly readable, and leaving the reader with a good sense of what it is about. A few of the words occur multiple times, across different sentences, and this may contribute to the overall relatively high performance of its aboutness terms. As Appendix 4 and Table 7.1 show, Text 5 exhibitsonly a small amount of linkage, although it is equally

second in terms of the median ratings of its aboutness terms, and fourth in terms of its PAV level.

Text 12

Text 12 is not quite as readable as the other high-performing texts, but presents a far more coherent picture than any of the low-performers. It is not particularly expository in style, and sentences tend to change topics, and do not follow on smoothly from one another.

In terms of its linkage, Text 12 is not notably high or low in relation to the other texts, occuring about half way down the list of decreasing linkage. Nor do any individual words show particularly large amounts of linkage, with the highest individual linkage occurring at the L2 level, for *institutions* and *commitment* (from Appendix 4).

Text 16

Again, as an informal comment, it is notable that Text 16 also has a continuity of topic, which makes it readable as a continuous piece of text. In this text, the word *union* appears in the first 3 of the 4 sentences, with *progress* also occurring 3 times. Although it has the 3rd highest PAV level, the amount of linkage exhibited is not particularly high, and it occurs half way down the list of decreasing linkage. Appendix 4 shows that there are no particularly highly linked words, with the maximum linkage occurring for the word *policy*, at the L2 level.

7.1.1.2 Low performers

Texts 9 and 17 are the two lowest performers in terms of both PAV level and median rating value.

Text 17

There is nothing particularly striking about text 17, except that it does not 'hang together' well, and is not particularly readable. It may be that the use of the occurrence of *the campaign*, and *the result* without qualification as the topics of two of the three sentences, has decreased the performance level in this case. Text 17 tends to be indicative rather than informative, with its three sentences jumping between topics.

Text 17 exhibits a greater degree of linkage than Text 9 (another low performer), at both the L1 and L2 levels. It occurs 13th in the overall linkage table, and 12th in the

L1 linkage table. However, this difference in linkage amount does not appear to be reflected particularly in their overall performance levels, as measured by their PAV levels and median ratings.

Text 9

Although fairly readable, Text 9 appears rather disjointed, relating a number of different facts, in a non-expository style. Much of the information is numerical, and is (sic) wasted on the efforts of this system.

Some particular problems may have contributed to the low performance of its aboutness terms. For example, *Heidelberg* is not listed by WordNet, whereas *Mannheim*, yielding the same entry as *Germany*, becomes linked and appears in the final output expressions. The lexical item *cope* has been wrongly assigned the syntactic class of noun (as in a type of brick), and does not even appear as a verb in WordNet at all.

However, the lexical item *waste* occurs 4 times, and links with *refuse* (since the default reading is as a noun). With the addition of some of the enhancements described in Section 7.2 below, the output would be likely to produce something much more representative, such as *Germany waste disposal strategies*, since the items *waste* and *strategies* have already been identified as salient head nouns, and already occur as heads of some of the aboutness terms.

Text 9 exhibits a relatively small amount of linkage. If we consider linkage at the L1 level, we see only 14 links (occurring equal 16th out of 20 in the L1 linkage table), and 15th out of 20 in the combined L1 and L2 linkage table.

Text 18

Text 18 exhibits a similarity to Text 17, lacking continuity of topic, and being fairly low in readability. The pronoun *it*, used on two occasions to refer to *the Treaty* may be partly responsible for a lack of linkage, although as the listing of linked items in Appendix 4 shows, the lexeme *Treaty* is linked anyway at the L2 level. The major factors involved in the poorer performance of this text, then, are more likely to be the switching of topics that occurs in the 2 sentences which comprise the second half of the text.

It is interesting to compare the high-performing Texts 16 and 12, to the low-performing text which is related to it in subject matter: text 17. The high performers have more continuity of topic, and repeats of particular lexical items: as noted above, in Text 16 we have 3 occurrence of each of *union* and *progress*. It is notable that Text 16 (and to some extent Text 12) seems far more cohesive than any of Texts 9, 17 or 18.

This section has given an indication that texts which are written in a 'continuous' style, with continuity of topic which can be identified by the methods used, and being to some degree expository, yield better output results than those which are less cohesive and more descriptive.

7.1.2 Characteristics of Aboutness Terms

In this section we turn our attention to the aboutness terms themselves. The main characteristic of the aboutness terms which might be expected to affect subjects' ratings is length. Throughout the analysis of the evaluation, which is described in Chapter Six, there was some evidence, amongst some of the subjects, of a tendency to find the longer expressions less clear to understand. The following section looks at the ratings given, across subjects, to some of the longer expressions and discusses whether or not there is any associated low performance.

7.1.2.1 Length of Expression

Table 7.2 below shows the lengths of all the genuine aboutness terms associated with each of the texts, with the 5 numbers in column 2 corresponding to the lengths of the 5 different terms. It also lists the total word number occurring across the 5 terms, and gives the medians for each set.

As this Table shows, there is clearly an uneven distribution of word lengths per expression, across each set of 5 terms, but a general impression of the comparative length of terms can be associated with the total length, taken across all the 5 terms for each text. This gives an idea of the ordering of texts in terms of the length of expressions associated with them.

Text No.	No. Words in Aboutness Terms.	Median Words per Term	Total No. Words in all 5 Terms
1	2 5 5 5 8	5	25
2	7 4 4 5 3	4	23
3	2 3 2 3 3	3	13
4	7 4 4 5 8	5	28
5	8 9 4 8 5	8	32
6	3 4 6 5 6	5	24
7	4 6 3 2 4	4	19
8	5 3 5 2 9	5	24
9	5 3 3 4 3	3	18
10	3 4 4 5 7	4	23
11	2 4 4 5 3	4	18
12	3 4 7 4 5	4	23
13	5 6 4 5 4	5	24
14	4 6 3 5 6	5	24
15	4 4 5 4 3	4	20
16	5 5 5 6 4	5	25
17	3 3 7 6 4	4	23
18	6 5 8 8 6	6	33
19	2 3 3 3 6	3	17
20	3 4 3 4 5	4	19

Table 7.2: Lengths of Different Genuine Aboutness Terms for Different Texts

Table 7.2 shows a decreasing order of length of expression as follows, with the longest expressions occurring in Text 18, and the shortest in Text 3:

18 > 5 > 4 > {1, 16}[39] > {6, 8, 13, 14} > {2, 10, 12, 17} > 15 > {7, 20} > {9, 11} > 19 > 3.

Comparison of the individual longest terms with the ratings accorded to those terms, across all subjects (which is achieved by looking at the results of the questionnaires),

[39] Texts listed together in brackets have the same total number of words across their 5 aboutness expressions

does not indicate that there is any tendency for subjects to give the longer expressions lower ratings.

Text 4, for example, has genuine aboutness terms of lengths 7, 4, 4, 5 and 8 words respectively. The data collected in the questionnaires shows that the longest expression (of length 8 words) was accorded ratings of 3 or 4 by half of the subjects, and has an individual PAV[40] (as described in Chapter Six) of 53.1%. Similarly, the 9-word aboutness term generated for Text 5 was given a rating of either 3 or 4 by 75% of the subjects who judged it, and had an individual PAV of 81.2%.

At this point it does not seem relevant, or indeed particularly useful, to present an exhaustive list of the PAV levels of each individual aboutness term. Just by looking at the overall distribution of ratings and length of terms, it seems clear that there is **no particular relation between rating value and length of expression**. Even if we look at the lengths of expressions across the highest and lowest performers, we see little indication of any connection. Thus, from the ordering of overall lengths accumulated across the 5 terms, and their median values (shown in Table 7.), we certainly do not find the best performers having the lowest numbers of words. On the contrary, the best performing texts tend to be in the top half of the lists of accumulated lengths (and medians). In terms of length, Text 4 appears in 3rd position (equal third, along with Texts 1, 6, 8, 13, 14, 16 according to its median); Text 5 in 2nd position according to accumulated length (and 1st position according to its median); Text 16, equal 4th, along with text 1; Text 6, equal 6th, along with Texts 8, 13 and 14, according to accumulated length (and equal third according to its median). On the other hand, although we do find that Text 18, being one of the lowest performers as regards the representativeness of its aboutness terms, has the largest accumulated total number of words in its 5 terms (2nd lowest according to its median), the other two low performers, Texts 9 and 17, are respectively at positions equal 17th (equal last), and equal 10th as regards their length.

In general, then, it seems fair to say that there is little, if any, relation between ratings given to expressions and the length of those expressions.

---

[40] Individual PAV levels are calculated as a % of their maximum attainment. In the case of our data, this maximum rating is clearly different from that which applies to a set of 5 terms, being described as $R_{max}\ x\ s$ (i.e.4 x 8) = 32.

7.1.3 Amount of Linkage in Different Texts

The discussions of individual texts in Section 7.1.1 above has indicated that there does not seem to be any particularly striking association between the amount of linkage identified between lexical items in a text and the subsequent extent to which its aboutness terms end up being judged as representative of the aboutness. We do not need any formal analysis to tell us this: a look at

Table 7.1 shows that the highest amount of linkage at both the L1 level alone, and the combined L1 and L2 level, occurs for Text 10, with Text 7 second in line in both cases. Both of these texts are neither particularly high nor low performers when judged by their aboutness terms. They both occur just above the 45% PAV level (Text 7 being 8th in the PAV league table, with a PAV of 50%, and Text 10 being 11th, with a PAV of 47.5%).

Comparing the linkage occurring in the high performance set of texts with one another, and the low performance set with one another gives the following generalisation. Whilst the high-performing texts 4, 6, 12 and 16 all show higher amounts of linkage than texts 9, 17 and 18, Text 5, which is one of the higher performers, shows less linkage than any of the low-performers. Although none of the high-performers has anything like the extent of the linkage seen in Text 10, we could suggest that the better performing aboutness terms tend to have higher amounts of linkage, but that high amounts of linkage do not necessarily produce highly representative aboutness terms.

7.1.4 Some Speculations about Style

The texts in the data set were, by intention, varied in both subject matter and style. Although no formal analysis of genre has accompanied this work, it seems that those texts which are clear to read, and those written in expository style produce better results. However, Texts which might have been expected to perform less well, such as the highly biographic Text 7, actually has a PAV level of 50% overall, and shows the second highest amount of overall linkage. So it seems that although texts written in expository style tend to produce better results, this property is not necessarily required for successful output to be produced.

7.1.5 Summary of Differences between High and Low Performers

Whilst we acknowledge the observation that the comments made above in relation to text style and 'readability' are informal, we can nevertheless make the following generalisations:

- the higher performing texts tend to be written in a more continuous style, with less topic-hopping, and may be described as being more expository, but this is not a requirement for representative output to be produced
- the amount of linkage involved in the generation of the aboutness terms is not necessarily related to their overall performance in the rating exercise carried out by subjects in the evaluation stage
- the rating values accorded to aboutness terms seems independent of the length of the term, and subjects rate even fairly long expressions as 'very' and 'extremely' representative of the aboutness of the corresponding texts.

7.2 Performance Enhancement

It has been made clear throughout that an aim of this work has been to investigate the extent to which the shallow semantic processing methodology can be usefully exploited in the generation of representative aboutness expressions. This section, therefore, which discusses the issue of performance enhancement, begins with some comments regarding the deliberately unencumbered approach taken. Section 7.2.2 discusses some of the negative consequences resulting from this approach, whilst Section 7.2.3 emphasises ways in which the performance could be improved.

7.2.1 Simplicity with No Add-ons

The encouragement of simplicity in this work has been driven by the wish to present an uncluttered picture of what can be attained without the addition of a number of performance-enhancing facilities.

It was an aim from the outset to ensure that the system remained completely domain-independent. The only knowledge base used is the definitions associated with the WordNet lookup facility. It should be emphasised here that the original aim was to use a normal on-line dictionary, such as LDOCE. Due to problems in acquiring this, it was decided to use the freely available WordNet, which was

becoming widely available at the time of implementation. Although only the basic synset lookup facility has been utilised, the use of WordNet has the advantage that its full semantic properties would be able to be exploited to further depths in any future work extending beyond the end of that undertaken in this thesis.

Another aim during the course of this work has been to minimise the use of syntactic processing, and to maximise the use of the semantic information implicit in the definitions of words. In this regard, we can say that the only syntactic processing done is the simple dictionary lookup of word class, and in cases of ambiguity the default assumption has been that the item is a noun (the reasoning behind which assumption is fully explained in Chapter Four).

Although an element of syntax has been involved in the nominalisation procedure, again, described in Chapter Four, the deliberate intention to refrain from using any full parsing process has been upheld throughout the development of the COMMIX system.

7.2.2 The Price of Simplicity

As was expected, there are a number of drawbacks to adopting such a simple approach, which have been discussed in detail in previous chapters. The main ones are mentioned in summary form here.

7.2.2.1 Ambiguity

By applying the system to the processing of short abstracts, particularly those which have been professionally written, the problems associated with ambiguity are reduced. This does not prevent their interference totally, and there are some occurrences of incorrectly interpreted items. For example, the word *won* is interpreted by default as a noun (being the monetary unit in North and South Korea), rather than the past tense of the verb, which is its correct interpretation in the text. Similarly, the name *Smith* has been interpreted in its general sense of 'trained worker' rather than being a specific name corresponding to the person mentioned in the text. And *Major* fails to be recognised as the Prime Minister of the time, but is interpreted as 'a commissioned military officer'.

These, and some additional cases of misclassification of ambiguous words and names do not necessarily cause problems regarding the overall processing of the texts. Even though an item may be misclassified, it is only if it then becomes involved

in the creation of links when it should not have done, or fails to when it should have done, that there may be ramifications for the overall performance. The results of the evaluation (discussed in Chapter Six), therefore, suggest that the low levels of ambiguity of word sense and word class can be tolerated.

7.2.2.2 Unordered Output

Apart from ensuring that the pooled modifiers occur before the head, and that, for individual groups of modifiers, relationships between nested modifiers and heads are maintained, the output expressions are not subjected to any principled post-ordering. This leads to some expressions having components which appear in a strange order. For example, for Text 10 we get *spending households high debt levels* (whereas we might prefer, for example, *households high debt spending levels*), and in Text 6 we see *characters Eastenders large number killing* (whereas *Eastenders large number characters killing* might be preferred).

Even with this 'erroneous' ordering, the expressions do indicate the aboutness to a fair degree. The head is still representative of the overall type of thing the text is about, and the individual multi-term modifiers also retain their modification relations. It seems clear, however, that a principled re-ordering would be likely to improve the performance of the COMMIX system.

7.2.2.3 Decapitalisation

Another problem caused by the simple approach taken occurs in regard to capitalisation. The over-simplified assumption used in this work is that sentence initial capitals can be decapitalised, unless the same word appears non-sentence initially elsewhere, with a capital. Thus, for eaxample, if a name occurs only in sentence-initial position, then it is decapitalised. For those names which *do* appear in WordNet, this is not a severe problem, since WordNet itself is not case sensitive (although COMMIX is), and the consequence is just that such names will appear decapitalised if they occur in any of the output produced. However, for names which have another meaning as normal lexical items (such as *Smith, Thatcher* and *Major*) the definition itself will be (wrongly) used if these items are not recognised as names. We may thus end up with inappropriate linkage being established, with the consequent possible lowering of performance.

7.2.2.4 Use of WordNet

Whilst WordNet is a useful resource, its definitions of words are not as rich as those we find in a dedicated dictionary, such as LDOCE. The consequences of using WordNet in its limited capacity are clear: much of the linkage that could have been ascertained by using a standard dictionary has not been identified using WordNet as a resource.

We note, however, that whatever the lexical resource used, there is no guarantee of linkage being itentified in all cases where humans would judge items to be semantically related. As Wilks points out: dictionaries, as texts, are fallible (Wilks, 1996, p.98).

7.2.2.5 A Technical Problem Concerning Punctuation

A technical problem which remains unsolved concerns the means by which Pop11 (the language of implementation) can be made to distinguish between its own syntax and the appearance of apostrophes, colons or semicolons in its input. Thus, any occurrences of apostrophes, generally in association with the use of the genitive, and colons or semicolons, need to be removed from the input. However, this is the only 'tampering' necessary, and the text input is otherwise completely un-pre-processed.

7.2.3 Incorporating some Additional Features

There are a number of additional features which could be added to the system in order to enhance its performance, some of which are mentioned below.

7.2.3.1 Maximise Use of WordNet Facilities

Whilst the work described in this thesis has deliberately used WordNet in a minimal capacity (see Chapter Four), this resource could clearly be exploited more fully. We could, for example, utilise the extensive semantic network properties to provide hypernyms at a number of different 'levels' of representation (increasing the 'lexical chaining'), and thus increase the generality of the aboutness terms produced. Similarly the output expressions could be extended in number to include alternative synonyms, which could be used both in the head and the modifying sections of the terms.

This substitution of synonyms and hypernyms, which would increase the scope for the generation of aboutness terms, would be a fairly simple and easily implemented

add-on feature. Such an addition would be of benefit if the COMMIX system were subsequently incorporated into a full IR facility, which would require users' terms to match against system-generated aboutness expressions.

7.2.3.2 Name and Company Listings

The approach adopted to the identification of proper names in this work is algorithmic, and relies on the over-simplification that any item not listed in the dictionary, but occurring non-sentence initially, and being capitalised, is a proper name. This is clearly superficial, and could be improved in a number of possible ways.

One method of improvement might be to integrate with COMMIX a system such as NYMBLE (Bikel et al, 1997), which identifies proper names with a high degree of accuracy. This is a probabilistic learning system, which relies neither on finite lists of names nor specific token and pattern matching. Its incorporation with COMMIX could be facilitated as described in Section 7.3, where it could be expoited as an independent preprocessing module within a larger text engineering environment. This would increase the reliability with which proper names were identified and labelled, and would be particularly useful in unambiguously identifying phrasal names, such as the 'European Defence Community' (Text 16) or the 'British Designer of the Year Award' (Text 3). It would not, however, contribute to the solution of the problem of co-reference, since the approach does not utilise world knowledge. Thus, although, for example, 'Vivienne Westwood', 'Paul Smith' and 'Jasper Conran' would be correctly identified as named entities, we would still require additional knowledge to co-refer each of these entities to 'designers' (Text 3). Indeed, Bikel et al (ibid) recognise the need for this type of information, and specify their wish to incorporate, amongst other things, lists of organisations, person names and locations.

An alternative system to utilise in respect of proper names would be NOMINATOR (Wacholder, Ravin and Choi, 1997). These authors utilise minimal parsing techniques, with pattern matching heuristics (including matching to commonly used (English) abbreviations), and specific 'shallow' heuristics aimed at ambiguity-resolution. The latter rely on the specification of particular types of 'ambiguity operators' (ibid, p. 205), which, they point out, tends to render the system at least partially domain dependent, and requiring large human input times for the development of these heuristics.

The use of either NYMBLE or NOMINATOR in the pre-processing task of COMMIX would clearly improv e the reliability of the identification and disambiguation of proper names, and it seems likely that a resource which combines the benefits might soon be developed.

In addition, the system could be linked to a gazetteer, and a source of titles, companies and people, which could be of benefit both in the identification and labelling of names (according to types), and in extending the work on pronoun resolution.

7.2.3.3 Principled Ordering of Modifiers

As discussed in Section 7.2.2 above, the output could be enhanced by including a post-processing stage which re-orders the modifying sub-phrases in a principled way. We would need to ensure that the components of nested compound nominals are kept together in the manner of the current implementation. If the components were to become separated from one another we might end up breaking the implicit semantic relations which hold within particular modification units.

7.2.3.4 Resolution of Pronouns

We have seen in Chapter Five that the lack of a pronoun-resolution facility has negative consequences both for the establishment of salience, and for the identification of linkage. Consequently, the system produces less terms than we might otherwise see generated (c.f., Boguraev and Kennedy, 1997, p. 4).

The addition of a pronoun resolution facility could improve the degree to which the aboutness terms do represent the aboutness of the text, although, as we have seen, this is not necessarily the case. In any such extension of COMMIX, we would ideally want to minimise the syntactic input at this stage, in line with the overall methodology employed. To this end a method such as that developed by Lappin and Leas (1994) and extended by Kennedy and Boguraev (1996) might prove a useful additional facility

7.2.3.5 Salience and Weighting

Although the processing of a text does involve some treatment of the notion of salience, in terms of both frequency and premodification, there is clearly scope for a fuller treatment of salience, which would, for example, bias the judgement of salience in favour of sentence-initial items (as mentioned in Chapter Two).

In the development of COMMIX we have been loathe to implement *ad-hoc* weighting measures, although this would have been easy to do. Beneficial weighting measures could be based on the amount of linkage associated with particular items in the text. For example, we might want to combine the frequency information with the weighting information, and could also include synonymy information in an extended treatment of weighting particular items in the texts. As has been mentioned previously, a good treatment of weighting, in relation to COMMIX, would be expected to be related to characteristics of the abstracts of different types of texts. Generalities such as the disproportional representativeness of the first sentence would be suitable for inclusion. However, the application of any specific style-dependent weighting criteria would clearly add the requirement for a formal analysis of the style of the input text.

The consideration of synonyms and hypernyms, possibly extending the use of WordNet to longer lexical chains, combined with a process of anaphora resolution and proper noun recognition, would be expected to give a strong base for the extension of the criteria for salience, and the development of a principled weighting procedure. This would then improve the conciseness and representativeness of the aboutness terms, and would facilitate the selection of the most appropriate aboutness terms.

This section has made just a few suggestions for ways in which the performance of the system could be improved by the addition of extra features. There is no limit, in theory at least (although there would be in terms of processing power required) to the number of such features which could be exploited. This work has given a clear and uncluttered picture of the basic level of performance which can be obtained from adopting a simple approach, with minimal resources, and leaves the way open for a number of extensions and studies.

7.3 Future Directions

In this section we distinguish between the practical applications of COMMIX, and its potential to be used as a research tool in further studies.

7.3.1 Practical Applications

To some extent, the COMMIX system could be described as an elaborate indexing system, whose generated terms (i.e., the aboutness terms) represent what a text is

about in a far more concise and informative way than typically occurs in automatic indexing systems. There are two main practical application to which COMMIX could usefully be directed. Firstly, in line with its current scope, as a facility for generating concise, representative aboutness terms which indicate text content. Secondly, as an element of a full IR system. These two aspects are discussed briefly below.

7.3.1.1 Integration with other Indexing Methods

The facility provided by COMMIX represents the essence of the text at a more informative level than that typically associated with indexing terms or key expressions. This facility could be integrated along with more traditional indexing methods, to give a 'hybrid' indexing system. Such a hybrid system could also include the traditional statistical methodologies, producing the standard output indexing expressions. The two types of expressions (indexing and essence) could be used as independent, alternative re-presentations of what the associated text is about, which could be selected by system users. Alternatively, (or additionally), the methodologies could be integrated, with information from both sources utilised to provide a cross-checking of one another.

This type of hybrid approach would have the potential for extension, with, perhaps, an automatic feedback between modules responsible for the different types of analyses. This would facilitate the automatic refinement of the final output terms, yielding a composite set of indexing/aboutness expressions.

In our consideration of possible extensions to COMMIX we have centred on the use of COMMIX in a fully automatic capacity. We could, however, envisage its use as a *semi-automatic* system, similar to, for example, SISTA (Whitehead, 1984). In such a capacity it would be useful as a tool which produced elaborate and informative aboutness terms for post-editing by professional indexers.

7.3.1.2 Extension into an Information Retrieval Facility

Although it has been recognised in the field of IR that the processing of query terms is a more successful approach than the processing of text (Sparck Jones and Tait, 1984), there is undoubtedly some advantage to be gained in having a concise representation of what each text in a data set is about.

The practical aspect of applying COMMIX processing to each text in a database would depend on the size of the database involved, and it would not necessarily be

sensible, or even feasible, for every text to be processed to yield aboutness terms in the way described in this thesis. However, the production of aboutness terms to represent the essence of texts could be a useful secondary stage in the IR endeavour, in which an initial set of retrieved texts could be subsequently processed to produce aboutness terms. These could then be used in a refined matching process against the original query.

This type of multi-stage approach could be further refined by the integration of query processing, in a manner similar to that suggested by Sparck Jones and Tait (1984). Complex user queries could easily be processed, using the COMMIX approach, so as to produce compound nominal 'query' expressions. These could then be used in a matching process against aboutness terms produced by COMMIX for each of a potentially relevant set of texts, identified at a first pass of the database. As stressed in this work, this would be expected to perform best for short texts, such as abstracts, since the problems of ambiguity (mentioned above) would be exacerbated in longer texts.

7.3.1.3 Extension to Longer Texts

The COMMIX technique has been restricted in usage to short abstracts, for reasons such as ambiguity and pronoun resolution, as already mentioned. The processing of longer texts would be possible, but would require a richer and more reliable text labelling facility. There are now increasing numbers of lexical resources becoming available, the inclusion of which would facilitate the extension of COMMIX to longer input texts. The incorporation of COMMIX into an extensive text engineering facility, such as GATE (Cunningham, Wilks and Gaizauskas, 1996), would provide a number of pre-processing stages which could be applied to any input text. Thus, for example, the parser and coreference resoloution modules, or 'creoles' (ibid, p. 1058), could be utilised to produce a more accurately labelled text. Utilising such an architecture as GATE would enable the incorporation of specific creoles (e.g., a nominalisation creole such as NOMLEX, Macleod et al, 1998), and additional facilities as they become available. In this type of environment, the labelling of the input text would clearly be far richer and comprehensive than the small-scale labelling performed by COMMIX.

This kind of extensive text analysis would add a strong syntactic element to the labelling stage of the processing. We would, however, continue to be able to adopt a

shallow semantic approach to the construction of compound nominals by the lexical overlap method described. With more reliable and informative labelling of input text, the extension of the approach to far longer texts would be expected to yield useful results. Such use would require a richer treatment of salience, which would benefit from an analysis of the text in terms of its discourse structure. We would expect to find an optimal length of document to which the methodology could usefully be applied. It seems clear that very long texts would be precluded due to the enormity of the task of searching for linkage between very large numbers of word pairs. We might expect, however, that the methodology could be extended to the full text of research papers (typically less than 10 pages). In this case, it could very usefully be applied as a checking facility, aimed either at corroboration/verification of the aboutness terms generated from the abstract, or applied directly to the extended text itself. Indeed, a comparison of the different aboutness terms generated from sets of research papers according to these different means would form the basis of some very interesting future research.

7.3.2 Research Tool

As has been noted at various points in this thesis, a secondary motivation behind the development of COMMIX has been its potential use as a research tool. Chapter Five has presented an example of such a usage. Additionally, in this respect, we mention two particularly relevant directions, geared respectively at improving the quality of abtracts, and a future investigation of the phenomenon of the compound nominal.

7.3.2.1 Specification of Improved Abstracting Practice

It would be interesting in a subsequent study to relate the findings from this work to a formal analysis of the text type. Such an analysis could be in terms of genre, employing a formal discourse analysis of the texts (perhaps in line with the work carried out by Rino (1995) in relation to physics abstracts). Particular findings of such a study could be used to specify more fully the criteria required of a good abstract across all genres, both in terms of the abstract as an end in itself, and in terms of providing a better basis for the application of a system such as COMMIX.

7.3.2.2 Study of Compound Nominals

This work has shown how the compound nominal is a highly versatile and useful construct for representing large amounts of information in a concise way. The

presentation of information in nominal form forces a reader to view the information as an entity (albeit often an abstract one). This has distinct advantages for the expression of complex topics (including queries to be made to an IR system). The COMMIX system has the potential to be used as a tool in the study of the compound nominal. It could be used in a number of ways: for example, to investigate the limits of acceptability of length of compound nominal expressions; or to study the effects of reordering 'chunks' of modifiers. It would also be interesting to follow up the indication we found whilst developing the evaluation, that people find it hard to produce compound nominals consistently for texts. This could be linked to an investigation of how easy people find it to construct representative compound nominals. Each of these studies would be expected to yield interesting results, and might lead to a comparison of human-generated versus COMMIX-generated aboutness terms for the same texts. We stress that this type of evaluation was not carried out here due to the very strong indication that the difficulties encountered by people attempting this task would fail to provide any consistency amongst the human-generated terms.

7.4 Conclusion

This final section summarises the aims behind the work described in this thesis, and the extent to which these aims have been fulfilled. We discuss the contribution the work has made to the field of content representation, with specific reference to the identification of the *essence gap,* and the provision of an automatic means of filling that gap.

7.4.1 Fulfilment of Aims

This work began with the identification of a significant gap in the modes commonly used to re-present textual information. We saw in Chapter Two how the whole field of content re-presentation can be usefully divided, on a functional level, between modes which purport to *reproduce* text content, and modes whose aim is to *indicate* content. We have expressed this distinction at the conceptual level, specifying the difference between the *gist* of a text and its *aboutness.* It is within this scenario that we identify a missing mode of expressing aboutness, which we term the *essence gap.*

The thrust of the work has been double-stranded. We firstly proposed that the *compound nominal* would be a highly useful construct for the expression of aboutness,

where the aboutness is *complex*. In Chapter Three we have discussed the characteristics of complex 'topics' which a text might be described as being about. A study of the features of the compound nominal has shown that this is indeed a particularly useful mode of expression of aboutness.

The second strand of the work has been the implementation of a proof-of-concept system, which automatically generates compound nominal aboutness expressions from input text. The aim in this respect was to implement a fully workable system to perform this task across all domains, taking natural language as input, without needing to specify large amounts of either syntactic or semantic information, but rather, exploiting the lexical resource of an on-line dictionary (although we ended up using WordNet).

As Chapters Four (detailing the implementation) and Six (showing the evaluation and its results) have shown, an implementation based on simple lexical overlap can produce some significantly encouraging results. Even in the absence of any enhancement facilities (of the type discussed earlier in this chapter), over 60% of the aboutness terms generated by the system for its input texts have been judged by human subjects as being either **reasonably, very,** or **extremely** representative of the aboutness of their associated texts.

In regard to the usefulness of the method of lexical overlap as applied to this task, we have seen that this method does succeed in establishing the necessary semantic linkage which forms the basis for the construction of these complex aboutness terms. It has, however, been shown that the representativeness of the expressions generated by COMMIX is not directly related to the *amount* of linkage involved in the generation of particular expressions. Rather, it seems that the identification of **some** linkage bears fruitful results.

The evaluation also revealed that subjects do not find it particularly difficult to judge the appropriateness of the aboutness terms, even though they comprise highly compact expressions.

We have shown, then, that although we might expect improved performance with the addition of a number of enhancement facilities, even the simplest implementation of this methodology has met with some significant success.

A further contribution made by the thesis relates to the evaluation of this type of work. Chapter Six discusses at length the unsuitability of the measures of precision and recall for the purposes of establishing the degree to which an expression represents what a text is about. This thesis has developed and presented a more appropriate measure for establishing the degree to which an expression (or set of expressions) succeeds in being representative of a text. This measure is aimed at establishing the extent to which expressions attain their maximum potential, as expressed by quality judgements based on human ratings. The degree of success is expressed as a percentage of the total possible attainment, and the metric is consequently referred to as the *percentage attainment value (PAV).* We have shown how the performance related to different texts, or expressions, can then be compared, either by direct comparison of individual *PAV* levels, or by identifying 'benchmark' categories of attainment.

We remember that a secondary aim of this work was to provide a system which could be used as a tool for the study of various linguistic phenomena. Chapter Five has shown an example of how the COMMIX system might be used in such a way; in this case, to investigate how the performance of the system differs when pronouns and names are replaced by their referents and types. This investigation has indicated that the absence of pronoun reference in particular might cause the system to identify less linkage, with some consequent under-performance (although the amount of output is not necessarily dependent on the **amount** of linkage identified).

We refer finally to the hypothesis presented in Chapter One, which stated:

shallow semantic processing based on the method of lexical overlap, applied to the definitions of distinct terms occurring in an existing text, constitutes an effective means of generating novel compound nominal expressions to represent the aboutness of text.

We have seen that the compound nominal as a construct may indeed be usefully exploited, at least in English, in the pooling of different pieces of information which pertain to specific and salient items occurring in a short piece of text, which in this case has been the abstract. The method used in their construction has been shown to be effective for this purpose. The aboutness terms formed by this method are rated very significantly better than dummy terms, and over 60% of them have been judged as being either reasonably, very, or extremely representative of the aboutness of their corresponding texts.

This work has shown that some highly encouraging results are possible, using a simple and shallow semantic methodology, with minimal recourse to syntactic methods. We have also discussed the potential enhancement of the system which would be expected to follow the incorporation of a variety of add-on facilities. COMMIX has the potential to be used not only in the generation of aboutness expressions which indicate the content of text, but also in the processing of user queries to generate 'aboutness queries' which might be used to represent complex topics of interest.

We finish, then, by reiterating the suggestion that the COMMIX approach could be usefully incorporated into various hybrid systems, particularly in the fields of Automatic Indexing and Information Retrieval, and we end in the hope that Fat Harry will soon have a bit more time on his hands.

Appendix 1: Text Set, with Aboutness Terms used in Questionnaires

This appendix contains a listing of all the texts used in the data set. Each text is followed by a section headed **Aboutness Terms**, which gives first the 5 dummy terms used, and then the 5 genuine terms used. Their ordering was randomised in the questionnaires.

TEXT 1

The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling. The main competitors in the market are Sega and Nintendo. Nintendo will spend 15 million pounds sterling on advertising over Oct-Dec 1992.

Aboutness Terms

Dummy Terms:
Dec 500 software 1991
associated million sales toy worth 1994
main forecast associated entertainment
Nintendo sales dominating 1993 million 15 entertainment games - becoming
advertising fast industry entertainment

Genuine Terms:
growing games
video games compact disc market
video games associated software sales
associated software compact disc market
1993 645 million pounds_sterling associated software sales

TEXT 2

Manufacturing success in recession depends on identifying changing needs and quickly responding with new products. Customers and suppliers should be involved in product development. Greater profits will result from smooth product introduction but old stocks may hinder the process. Computer systems are essential to effective management but some companies are being badly advised. Operational simplicity should be the guideline.

Aboutness Terms

Dummy Terms:
guideline development operational profits
hinder badly greater success result identifying
simplicity products customers greater manufacturing essential - identifying smooth result
badly smooth products
computer success process customers

Genuine Terms:
identifying changing needs recession depending manufacturing product
effective management computer systems
effective management operational simplicity
smooth product introduction computer systems
manufacturing computer systems

TEXT 3

The designer fashion industry in Britain is under threat. Designers Paul Smith and Jasper Conran have withdrawn from the British Designer of the Year Award because they felt the award took attention from the problems facing the industry. Vivienne Westward is taking her fashion collection to Paris rather than have a London show. 100,000 jobs in the clothing industry have been lost since 1989. Some believe that top designers represent only a tiny fraction of the industry. They point to partnerships between manufacturers and designers as the way forward.

Aboutness Terms

Dummy Terms:
representing Paris Britain tiny Jasper
withdrawn London Britain
point representing believing industry
fraction 1989 forward industry award
designer Award forward threat manufacturers

Genuine Terms:
British fashion
London clothing industry
British designers
fashion Year award
problems fashion collection

TEXT 4

Environmental considerations are encouraging the development of dry running pumps to eliminate vapour contamination. However removing the sealing fluid may cause localised overheating. Cooling by inert gas injection may impair exhaust condenser efficiency. Other advances in pump technology are aimed at increased efficiency, flexibility, reduction of downtime and low maintenance costs.

Aboutness Terms

Dummy Terms:

gas localising
environmental efficiency considerations cause inert
overheating encouraging impairing advances
inert maintenance fluid aiming pumps encouraging
dry flexibility contamination sealing

Genuine Terms:

eliminating vapour contamination dry running pumps development
pump technology environmental considerations
environmental pump technology advances
low maintenance pump technology advances
environmental eliminating vapour contamination dry running - pumps development

TEXT 5

An opera house is planned for Salford Quays, England. The Salford Quays development includes housing, a cinema, hotel, cafes and offices. An 1200 seat opera house is to be added. The building will cost 30-60 million pounds sterling. The architect is Michael Wilford. The theatre will be used for opera, dance and concerts and will include an art gallery for Salford's Lowry collection. A few people consider such a development an extravagance because adjoining Salford Quays is the depressed Ordsall estate.

Aboutness Terms

Dummy Terms:

pounds extravagance Salford 1200 theatre cinema gallery cafes - Ordsdall
used architect building Lowry housing cafes
Lowry dance housing 1200 used Ordsall
art development concerts sterling including gallery
30 -60 architect cafes depressed opera

Genuine Terms:

Salford Quays development including 1200 seat opera house
Salford Quays development including Salford Lowry collection art - gallery
1200 seat opera house
Salford Lowry collection art 1200 seat opera house
Salford Lowry collection art development

TEXT 6

Disposing of soap opera characters presents problems. A soap opera writer is suing the British Broadcasting Corporation for dismissing him when he proposed removing a large number of characters from EastEnders by killing them with an IRA bomb. Deaths of soap opera characters are normally infrequent. Large scale turmoil created by death is out of key with the normal banal rhythm of such productions. It is unusual in Britain to replace an actor who dies, although this is done in the United States. The death of a soap opera character creates publicity and may improve viewer ratings.

Aboutness Terms

Dummy Terms:
viewer normally dies
Broadcasting dismissing normally characters improving
unusual banal publicity killing
character scale unusual key productions Britain infrequent
improving normally deaths character key

Genuine Terms:
soap opera deaths
soap opera character deaths
dismissing British Broadcasting soap opera writer
characters Eastenders large number killing
normal banal soap opera characters deaths

TEXT 7

Derek Walcott has won the 1992 Nobel Prize for Literature. Walcott was born in 1930 in St Lucia in the Caribbean. He went to the University of the West Indies. He published his first book when he was eighteen. Much of his poetry is influenced by his sense of the sea. His epic poem, Omeros, is a Caribbean version of Homer. He is a playwright as well as a poet. He teaches creative writing at Boston University but lives half the year in Trinidad. It is the first time a Nobel prize has been awarded to an English-speaking Caribbean.

Aboutness Terms

Dummy Terms:
time published born
Homer writing awarding year half
half prize won epic playwright version poem Homer
sense University sea lives Prize going time
year Omeross

Genuine Terms:
epic Homer Caribbean version
literature 1992 Nobel Homer Caribbean version
1992 Nobel book
Caribbean first
Homer Caribbean first book

TEXT 8

An inquiry is being held into allegations of scientific fraud at Leeds University which led to the man revealing the fraud losing his job. A biochemist, Chris Chapman, employed at Leeds General Hospital, revealed fraudulent research at the University which wasted thousands of pounds sterling. Chapman was technically made redundant from the National Health Service trust which employed him, but a witness is prepared to testify that his dismissal was deliberate.

Aboutness Terms

Dummy Terms:
Chapman University employed biochemist Service
scientific thousands Leeds
thousands Hospital man Health inquiry
trust Service Hospital Chapman deliberate testifying scientific University witness wasted

Genuine Terms:
National Health Service scientific fraud
man scientific fraud
Leeds National Health Service trust
man fraud
allegations scientific fraud holding Leeds National Health Service trust

TEXT 9

Heidelberg, Germany, is developing strategies to cope with its refuse. Heidelberg has a population of 120,000. In 1991 it produced 75,000 tons of rubbish. 25,000 tons was sent to waste disposal sites in France. France will no longer accept German refuse. Heidelberg has responded with a 3-point plan, to avoid creating rubbish, to recycle it and to incinerate it. In cooperation with Mannheim, Heidelberg will compost 55,000 tons of waste while Mannheim incinerates the waste for both cities. Citizens are being given an incentive to produce less waste by a rent reduction.

Aboutness Terms

Dummy Terms:
longer compost 1991
Mannheim strategies france reduction
rent 120,000 German
75,000 citizens plan rubbish cities 1991
less point reduction three plan

Genuine Terms:
france waste disposal sites sending
avoiding creating waste
Mannheim cope strategies
three point cope strategies
three point strategies

TEXT 10

The basic unit of the economy is the home. The home is not merely a place for rest and refreshment. It has become an entertainment centre housing a vast amount of technology. Nearly 70% of people own their homes and it is their largest source of marketable wealth. The government of the 1980s encouraged people to use the capital value of their homes to borrow money to spend on consumer goods thus triggering the consumer boom and leading to the recession. The present economic problem arises from the high debt levels of households and their reluctance to spend more on consumer goods.

Aboutness Terms

Dummy Terms:
money centre triggering homes capital
source levels
people money households percent capital wealth
becoming households borrowing
nearly people problem

Genuine Terms:
entertainment centre homes
consumer high debt levels
entertainment centre consumer goods
spending households high debt levels
households high debt marketable wealth largest source

TEXT 11

The French voted narrowly in favour of the Maastricht treaty on European union in the referendum of Sep 20th 1992, but the size of the minority against the treaty made it a pyrrhic victory. Senior French politicians, not least President Mitterrand, were almost unanimous in their support for the treaty, and their task in the aftermath of the referendum will be to woo the support of their own voters so as to assure their own political future.

Aboutness Terms

Dummy Terms:
least President
senior treaty
wooing French referendum European own victory voting
victory treaty European narrowly
aftermath Maastricht President

Genuine Terms:
Maastricht treaty
European union Maastricht treaty
Sep 20 th 1992 referendum
French Sep 20 th 1992 referendum
French own voters

TEXT 12

The narrow French vote in favour of the Maastricht treaty on European union in the referendum on Sep 20th, 1992, confirms France's commitment to Europe but also sounds two warnings. The large minority that did not support the treaty should act as a reminder that progress towards European union should be gradual. European institutions will have to observe the will of the leaders and the population of member states. The 'no' vote can also be interpreted as a vote of no confidence in President Francois Mitterrand and as an expression of the French wish for change.

Aboutness Terms

Dummy Terms:
favour Sep institutions
sounds support
vote confidence wish French referendum
sounds european two 20 population progress
President French 1992

Genuine Terms:
European union progress
Maastricht narrow French vote
Sep 20 th 1992 Europe confirming French commitment
european narrow french vote
Sep 20 th 1992 large minority

TEXT 13

The uncertainty engendered by the French referendum on the Maastricht treaty on European union was only one factor in the financial crisis which led to the collapse of the European Monetary System (EMS). The EMS, designed to be flexible, was being treated by member states like a single European currency before the necessary economic convergence had been attained. The resulting rigidity was compounded by some member states' insistence, for reasons of national pride, in maintaining their currencies at too high a parity. The renegotiation of the EMS is now a priority.

Aboutness Terms

Dummy Terms:
insistence reasons European necessary pride
currency union
treated EMS insistence
European led treated union single engendering
resulting states

Genuine Terms:
national pride European Monetary System
European union Maastricht necessary economic convergence
Maastricht necessary economic convergence
national pride single European currency
member states financial crisis

TEXT 14

The French ratification of the Maastricht treaty on European union has gone some way towards assuring the future of a united Europe, but the treaty still has to clear the hurdle of the vote in the British Parliament, expected to be close, and ratification in the German Bundestag. Leaders must find a compromise that will satisfy the Danes. The timetable for the implementation of the treaty is already set, beginning with the introduction of a single European market in Jan 1993, and early priorities must be the reform of the common Agricultural policy and the restructuring of the European Monetary System.

Aboutness Terms

Dummy Terms:
beginning clear European market
European Parliament set Europe ratification Danes
introduction early leaders Jan
expected priorities European market
Bundestag French

Genuine Terms:
European union Maastricht treaty
common Agricultural European union Maastricht treaty
common Agricultural ratification
Jan 1993 single European reform
early Jan 1993 single European market

TEXT 15

The French referendum on the Maastricht treaty on European union resulted in a victory for the treaty, but the 49 percent minority who voted against the motion represent a vote of no confidence in President Mitterrand and in the status quo in French politics. There are four main lessons to be drawn from the result: President Francois Mitterrand must announce his retirement; the presidential term must be reduced from seven to five years; political alliances should be reconsidered; and European Community affairs must be decided democratically and in public.

Aboutness Terms

Dummy Terms:
seven reconsidering voting term
referendum motion drawn five
affairs reduced union representing
victory President 49 minority politics
drawn affairs victory

Genuine Terms:
European Community French politics
President Mitterand French politics
President Mitterand European Community affairs
French 49 percent minority
French political alliances

TEXT 16

The French ratification of the Maastricht treaty on European union is only the latest step in a long progress towards a united, federal Europe, and goes some way towards reversing France's initial vetoing of the formation of a European Defence Community in 1954. The Dutch too vetoed a scheme for closer European union, which in many ways prefigured Maastricht, in 1962. The way is now clear for progress, which will begin with moves towards monetary union. Political progress will depend on developing a workable EC-wide foreign policy and ensuring that the EC can survive expansion.

Aboutness Terms

Dummy Terms:
French foreign vetoing way now federal
ratification Europe Maastricht scheme ways
Defence now Dutch union
formation now moves 1962 foreign
ensuring Europe treaty french federal expansion progress

Genuine Terms:
European union Maastricht latest step
closer European union political progress
European union Maastricht French ratification
closer European monetary union moves beginning
closer European union formation

TEXT 17

President of the French National Foundation for Political Science Rene Remond believes that the narrow French referendum victory in favour of ratifying the Maastricht treaty on European union proves the effectiveness of the use of a referendum. The campaign has shown that political campaigns can still focus on the issues. The result has also indicated the need for a recognition of fundamental political realignments.

Aboutness Terms

Dummy Terms:
issues narrow proving National showing
president still Political result indicating Foundation effectiveness focus
referendum narrow proving
Political Maastricht
issues political National Science union

Genuine Terms:
fundamental political referendum
fundamental political recognition
Political Science French National fundamental political recognition
narrow French referendum French National Foundation
narrow French referendum use

TEXT 18

A House of Commons vote on returning the legislation of the Treaty on European Union to the House achieved 319 votes for and 313 against. The Liberal Democrats voted for it as they felt it would help Britain get out of the recession. Defeat might have forced the resignation of Prime Minister John Major. Sir Edward Heath said Lord Tebbit and Lady Thatcher's interfering attempts to strengthen rebellion was incredible.

Aboutness Terms

Dummy Terms:
returning might Commons getting
felt attempts Commons
European Democrats felt rebellion defeat might
Union recession Britain
Thatcher Commons Liberal

Genuine Terms:
Liberal Democrats House achieving 319 legislation
Prime Minister John Major legislation
Prime Minister John Major House achieving 319 votes
strengthening rebellion Lady Thatcher interfering help Britain getting
European Union House achieving 319 votes

TEXT 19

The government won a majority of three on its motion to return the Maastricht Bill to the Commons. Prime Minister John Major accused Labour leader John Smith of fraud in voting against the motion while declaring himself pro-Europe. John Major asserted the importance of Britain playing a central part in the development of Europe. He considered a free-market and a wider community responsive to its citizens essential to EC development.

Aboutness Terms

Dummy Terms:
majority fraud
Europe voting importance citizens accused minister
asserted motion essential won
voting central development accused
accused Minister part central

Genuine Terms:
EC motion
Maastricht return motion
Maastricht EC development
Maastricht Europe development
EC development citizens Britain playing importance

TEXT 20

The Prime Minister John Major must abandon the Maastricht Treaty on European Union to save the British economy. Major must restate his economic principles since the collapse of sterling in the Exchange Rate Mechanism (ERM) destroyed his previous policy. He must clarify public spending plans. Public services must be improved. Increased public spending means increased taxes, but these should be taxes on purchases. Taxation should replace government borrowing. Interest rates should be reduced. Recovery depends on sterling's independence from the ERM and Britain cannot, therefore, be party to the Maastricht Treaty.

Aboutness Terms

Dummy Terms:
majority citizens considered
three voting leader fraud bill
abandon European Treaty Major collapse
ERM Britain borrowing taxes depending
party improved clarifying public restating replacing

Genuine Terms:
British economic principles
previous Exchange Rate Mechanism
sterling interest rates
replacing government sterling collapse
clarifying public spending economic principle

Appendix 2: Questions Associated with Texts

This appendix contains a list of all the multiple choice questions - one pertaining to each of the 20 texts. Each subject was only presented with the questions relating to the texts which appeared on their particular questionnaire, and questions were presented in the same order as the texts in the questionnaires.

The instruction section appeared at the top of the set of questions given to each subject.

Questions

Please check carefully to ensure that the question number is the same as the number of the text you have just read.

Then select one or more of answers *a, b, c, d* and circle the corresponding letter/s. For example, in the first question, if you think that both answers *a* and *c* are correct, then you should put circles around the letters *a* and *c*.

1. Is the market in computer games ...

Answers:

a	*in decline*	*b*	*increasing*
c	*driven by advertising*	*d*	*none of a-c*

2. What should companies be doing in order to succeed?

Answers:

a	*buying new computers*	*b*	*consulting existing customers*
c	*improving stock management*	*d*	*none of a-c*

3. Where did Vivienne Westward show her work?

Answers:

a	*London and Paris*	*b*	*London only*
c	*Paris only*	*d*	*none of a-c*

4. What might reduce the efficiency of exhaust condensers?

Answers:

a	*cooling the condensers*	*b*	*injecting gas into them*
c	*removing sealing fluid*	*d*	*none of a-c*

5. What do you think is on the Ordsall estate?

Answers:

a	*a theatre*	*b*	*an opera house*
c	*an office block*	*d*	*none of a-c*

6. In Britain, soap opera characters are killed off ...

Answers:

a	*very often*	*b*	*rarely*
c	*more often than in the US*	*d*	*none of a-c*

7. Who / what is mentioned as being a big influence on Walcott's work?

Answers:

a	*Homer*	*b*	*the sea*
c	*the Caribbean*	*d*	*none of a-c*

8. Where was the original alledged faud at the University or the Hospital?

Answers:

a	*Leeds University*	*b*	*Leeds General Hospital*
c	*a NHS trust*	*d*	*none of a-c*

9. Which of the following are new strategies for dealing with waste?

Answers:

a	*burning it*	*b*	*burying it*
c	*composting it*	*d*	*none of a-c*

10. What is a new function of the home?

Answers:

a	*a place to recover*	*b*	*an escape*
c	*a place for enjoyment*	*d*	*none of a-c*

11. What will Mitterand have to pursuade voters to support in the future?

Answers:

a	*a referendum*	*b*	*European Union*
b	*a new treaty*	*d*	*none of a-c*

12. In what manner should France proceed into European Union?

Answers:

a	*enthusiastically*	*b*	*gradually*
c	*confidently*	*d*	*none of a-c*

13. Which of the following contributed to the collapse of the EMS?

Answers:

a	*French uncertainty*	*b*	*too much flexibility*
c	*unrealistic rates*	*d*	*none of a-c*

14. Which countries are named as being problematic for European Union?

Answers:

a	*Britain*	*b*	*Germany*
c	*Denmark*	*d*	*none of a-c*

15. What does the referendum result suggest?

Answers:

a	*Mitterand should retire*	*b*	*Mitterand is popular*
c	*French politics needs revising*	*d*	*none of a-c*

16. How many different attempts are European Union are referred to?

Answers:

a	*one*	*b*	*two*
c	*three*	*d*	*none of a-c*

17. How would Rene Remond describe the referendum as a tool?

Answers:

a	*unreliable*	*b*	*reliable*
c	*useful*	*d*	*none of a-c*

18. How did the Liberal Democrats vote?

Answers:

a	*Against the treaty*	*b*	*For the treaty*
c	*Enthusiatically*	*d*	*none of a-c*

19. Who was accused of fraud?

Answers:

a	*John Major*	*b*	*John Smith*
c	*the Labour party*	*d*	*none of a-c*

20. What does the writer think taxes should be?

Answers:

a	*increased*	*b*	*stay the same*
c	*on income*	*d*	*none of a-c*

Appendix 3: Evaluation Instructions for all Subjects, and Example of Full Questionnaire (Subject A)

Instructions

Please read both sides of this sheet.
You have been given here 4 things:

- this instruction page
- a separate page, on the reverse of this sheet, headed **The Task**, which describes the task as a series of steps. It is worth keeping this section in view, so you can refer to it easily during the task should you need to do so
- a set of 10 short pieces of text, each having a number between 1 and 20, and each one followed by a page headed **Aboutness terms** ... These are stapled together, with a front sheet bearing the heading **Texts and terms**
- a separate sheet of paper, headed **Questions**, on which there is a set of 10 multiple choice questions. There is one question for each text, and the questions have the same numbers as the texts to which they refer. **Do not read these questions before you start the task**

First of all, please read the rest of this instruction sheet **and** the task description overleaf.
This is not a timed exercise. The whole task should take about one hour to complete, although you may find that you need slightly more or less than that.
It is important to remember during the task that there are no correct or incorrect answers or opinions. The important thing is to read each text carefully, and get a good idea of what it is about, which will enable you to express your own opinions accurately.
Do not worry if you end up expressing similar opinions lots of times. The purpose of the task is to express as accurately as possible each opinion you are asked to give.
Please make sure that you read the task description on the following page carefully, and follow the steps in the right order. In particular, please make sure that you do not read ahead on the question sheet: only read a question when you are asked to do so.
When you have completed the task, please return the whole package to me as soon as you can, either in person, or via the internal mail.
Thank you very much indeed for your time.

Jenny Norris,
IT Research Institute,
Watts Building.

Please turn over

Read the other side first!

The Task

Please do each aspect of the task described below for each text in turn, one text at a time, completing the whole task before going on to the next text. When you are asked to read and answer a question about a text, make sure that you have read the text and got a good idea of what it is about before you read and answer the question related to that text. At this stage, please also avoid reading the questions that relate to texts which you haven't yet read.
It is important to realise that you are not being asked to express an opinion about the syntactic (or grammatical) 'correctness' of any of the aboutness terms. What you **are** being asked is to express to what degree the terms convey what the text is about, and to what degree you find them easy to understand.

Instructions for Completing the Task

Before beginning the task, find the section headed **Texts and terms**, and write your name, or if you prefer, just your initials, at the top of the 1st page, where indicated. Do this also on the sheet headed **Questions**. PLEASE DO THIS NOW.
Next, follow the stages below for the first text, and then for subsequent texts, one at a time. Make sure that all the stages are completed for each text in turn, and that you only go on to the next text when you have completed the whole task for the text in hand.

1. Find the section headed **Texts and Terms**, read one text (x) carefully, and try to get a really good idea of what it is about.
2. Find the section headed **Aboutness terms relating to Text x**, which is on the following page, and have both the text and the aboutness terms showing. For each of the terms listed there you are asked to express two things:
    - how well you think the term expresses what the text is about. When you have decided, record your opinion as one of the options given, by circling the appropriate number. (If you forget what the numbers mean, just look across to the definitions listed underneath the text)
    - how clear, or easy to understand, the term is. Again, record your opinion as one of the options given, by circling the appropriate number.
3. Do this for each of the 10 aboutness terms shown for that text. Use the whole range of numbers, including the extremes of 0 and 4, but only if you feel them to be appropriate. Do not worry if you end up with a lot of the same numbers circled.
4. Then look at the question sheet, and read and answer the question which relates to the piece of text you have just read, making sure that the question number corresponds to the number of the text.

Repeat the above steps (1-4) for the next text, until you have done the task for each of the 10 texts.

Appendix 3: Example Questionnaire (Subject A)

Your Name (or initials):

Texts and Terms (A)

TEXT 4

Environmental considerations are encouraging the development of dry running pumps to eliminate vapour contamination. However removing the sealing fluid may cause localised overheating. Cooling by inert gas injection may impair exhaust condenser efficiency. Other advances in pump technology are aimed at increased efficiency, flexibility, reduction of downtime and low maintenance costs.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 4

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

not at all ——————>—————— extremely

environmental efficiency considerations cause inert

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

dry flexibility contamination sealing

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

eliminating vapour contamination dry running pumps development

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

environmental pump technology advances

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

pump technology environmental considerations

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

gas localising

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

inert maintenance fluid aiming pumps encouraging

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

overheating encouraging impairing advances

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

environmental eliminating vapour contamination dry running pumps - development

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

low maintenance pump technology advances

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 4 on your question sheet.

TEXT 2

Manufacturing success in recession depends on identifying changing needs and quickly responding with new products. Customers and suppliers should be involved in product development. Greater profits will result from smooth product introduction but old stocks may hinder the process. Computer systems are essential to effective management but some companies are being badly advised. Operational simplicity should be the guideline.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 2

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

not at all ——————> —— extremely

effective management operational simplicity

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

badly smooth products

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

computer success process customers

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

simplicity products customers greater manufacturing essential identifying - smooth result

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

hinder badly greater success result identifying

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

smooth product introduction computer systems

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

manufacturing computer systems

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

guideline development operational profits

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

identifying changing needs recession depending manufacturing product

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

effective management computer systems

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 2 on your question sheet.

TEXT 3

The designer fashion industry in Britain is under threat. Designers Paul Smith and Jasper Conran have withdrawn from the British Designer of the Year Award because they felt the award took attention from the problems facing the industry. Vivienne Westward is taking her fashion collection to Paris rather than have a London show. 100,000 jobs in the clothing industry have been lost since 1989. Some believe that top designers represent only a tiny fraction of the industry. They point to partnerships between manufacturers and designers as the way forward.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 3

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

	not at all ——————>—————— extremely				
fraction 1989 forward industry award					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
representing Paris Britain tiny Jasper					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
designer Award forward threat manufacturers					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
British designers					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
withdrawn London Britain					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
problems fashion collection					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
fashion Year award					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
London clothing industry					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
British fashion					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
point representing believing industry					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 3 on your question sheet.

TEXT 5

An opera house is planned for Salford Quays, England. The Salford Quays development includes housing, a cinema, hotel, cafes and offices. A 1200 seat opera house is to be added. The building will cost 30-60 million pounds sterling. The architect is Michael Wilford. The theatre will be used for opera, dance and concerts and will include an art gallery for Salford's Lowry collection. A few people consider such a development an extravagance because adjoining Salford Quays is the depressed Ordsall estate.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 5

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

not at all ——————>—————— extremely

Salford Lowry collection art 1200 seat opera house

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Lowry dance housing 1200 used Ordsall

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Salford Lowry collection art opera

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

30 -60 architect cafes depressed opera

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Salford Quays development including 1200 seat opera house

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Salford Quays development including Salford Lowry collection art gallery

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

used architect building Lowry housing cafes

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Salford Lowry collection art development

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

pounds extravagance Salford 1200 theatre cinema gallery cafes Ordsall

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

art development concerts sterling including gallery

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 5 on your question sheet.

TEXT 1

The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling. The main competitors in the market are Sega and Nintendo. Nintendo will spend 15 million pounds sterling on advertising over Oct-Dec 1992.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 1

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

	not at all ——————>—————— extremely				
associated software compact disc market					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
video games associated software sales					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
Dec 500 software 1991					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
Nintendo sales dominating 1993 million 15 entertainment games becoming					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
growing games					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
advertising fast industry entertainment					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
main forecast associated entertainment					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
video games compact disc market					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
1993 645 million pounds sterling associated software sales					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
associated million sales toy worth 1994					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 1 on your question sheet, then return to the next text.

TEXT 11

The French voted narrowly in favour of the Maastricht treaty on European union in the referendum of Sep 20th 1992, but the size of the minority against the treaty made it a pyrrhic victory. Senior French politicians, not least President Mitterrand, were almost unanimous in their support for the treaty, and their task in the aftermath of the referendum will be to woo the support of their own voters so as to assure their own political future.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 11

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

	not at all ——————>—————— extremely				
French own voters					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
European union Maastricht treaty					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
wooing French referendum European own victory voting					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
senior treaty					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
aftermath Maastricht President					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
French Sep 20 th 1992 referendum					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
own political Maastricht treaty					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
victory treaty European narrowly					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
Sep 20 th 1992 referendum					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
least President					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 11 on your question sheet.

TEXT 15

The French referendum on the Maastricht treaty on European union resulted in a victory for the treaty, but the 49 percent minority who voted against the motion represent a vote of no confidence in President Mitterrand and in the status quo in French politics. There are four main lessons to be drawn from the result: President Francois Mitterrand must announce his retirement; the presidential term must be reduced from seven to five years; political alliances should be reconsidered; and European Community affairs must be decided democratically and in public.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 15

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

	not at all ——————>—————— extremely				
French 49 percent minority					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
President Mitterand French politics					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
victory President 49 minority politics					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
referendum motion drawn five					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
drawn affairs victory					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
French political alliances					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
European Community French politics					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
affairs reduced union representing					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
seven reconsidering voting term					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
President Mitterand European Community affairs					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 15 on your question sheet.

TEXT 12

The narrow French vote in favour of the Maastricht treaty on European union in the referendum on Sep 20th, 1992, confirms France's commitment to Europe but also sounds two warnings. The large minority that did not support the treaty should act as a reminder that progress towards European union should be gradual. European institutions will have to observe the will of the leaders and the population of member states. The 'no' vote can also be interpreted as a vote of no confidence in President Francois Mitterrand and as an expression of the French wish for change.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 12

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

	not at all ——→—— **extremely**				
european narrow french vote					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
sounds european two 20 population progress					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
sounds support					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
Maastricht narrow French vote					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
President French 1992					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
favour Sep institutions					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
Sep 20 th 1992 large minority					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
Sep 20 th 1992 Europe confirming French commitment					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
vote confidence wish French referendum					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4
European union progress					
How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 12 on your question sheet.

TEXT 14

The French ratification of the Maastricht treaty on European union has gone some way towards assuring the future of a united Europe, but the treaty still has to clear the hurdle of the vote in the British Parliament, expected to be close, and ratification in the German Bundestag. Leaders must find a compromise that will satisfy the Danes. The timetable for the implementation of the treaty is already set, beginning with the introduction of a single European market in Jan 1993, and early priorities must be the reform of the common Agricultural policy and the restructuring of the European Monetary System.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 14

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

not at all ——————>—————— extremely

beginning clear European market

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

common Agricultural ratification

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Jan 1993 single European reform

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

introduction early leaders Jan

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Bundestag French

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

early Jan 1993 single European market

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

European Parliament set Europe ratification Danes

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

beginning clear European market

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

common Agricultural European union Maastricht treaty

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

European union Maastricht treaty

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 14 on your question sheet.

TEXT 13

The uncertainty engendered by the French referendum on the Maastricht treaty on European union was only one factor in the financial crisis which led to the collapse of the European Monetary System (EMS). The EMS, designed to be flexible, was being treated by member states like a single European currency before the necessary economic convergence had been attained. The resulting rigidity was compounded by some member states' insistence, for reasons of national pride, in maintaining their currencies at too high a parity. The renegotiation of the EMS is now a priority.

Now read the aboutness terms on the following page, and for each one circle one number between 0 - 4 for each of the questions.

0 means that the expression is ***not at all*** *representative of what the text is about, or that the expression is* ***not at all clear*** *or easy to understand.*

1 means that the expression is ***slightly*** *representative of what the text is about, or that the expression is* ***slightly clear*** *or easy to understand.*

2 means that the expression is ***reasonably*** *representative of what the text is about, or that the expression is* ***reasonably clear*** *and easy to understand.*

3 means that the expression is ***very*** *representative of what the text is about, or that the expression is* ***very clear*** *and easy to understand.*

4 means that the expression is ***extremely*** *representative of what the text is about, or that the expression is* ***extremely clear*** *and easy to understand.*

Aboutness Terms relating to Text 13

Please read each of the following expressions, and for each one decide:

- to what degree it represents what the text you have just read is about
- how easy you found the expression to understand.

Record each opinion by circling the appropriate number.

not at all ——————> extremely

European union Maastricht necessary economic convergence

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

European led treated union single engendering

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

national pride European Monetary System

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Maastricht necessary economic convergence

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

national pride single European currency

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

treated EMS insistence

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

resulting states

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

member states financial crisis

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

currency union

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

insistence reasons European necessary pride

How representative of what the text is about?	0	1	2	3	4
How clear to understand?	0	1	2	3	4

Now answer question 13 on your question sheet.

Your Name (Or initials):

Questions (A)

Please check carefully to ensure that the question number is the same as the number of the text you have just read.
Then select one or more of answers *a, b, c, d* and circle the corresponding letter/s.
For example, in the first question, if you think that both answers *a* and *c* are correct, then you should put circles around the letters *a* and *c*.

4. What might reduce the efficiency of exhaust condensers?
Answers:

a	*cooling the condensers*	*b*	*injecting gas into them*
c	*removing sealing fluid*	*d*	*none of a-c*

2. What should companies be doing in order to succeed?
Answers:

a	*buying new computers*	*b*	*consulting existing customers*
c	*improving stock management*	*d*	*none of a-c*

3. Where did Vivienne Westward show her work?
Answers:

a	*London and Paris*	*b*	*London only*
c	*Paris only*	*d*	*none of a-c*

5. What do you think is on the Ordsall estate?
Answers:

a	*a theatre*	*b*	*an opera house*
c	*an office block*	*d*	*none of a-c*

1. Is the market in computer games ...
Answers:

a	*in decline*	*b*	*increasing*
c	*driven by advertising*	*d*	*none of a-c*

continued overleaf

11. What will Mitterand have to pursuade voters to support in the future?

Answers:

a	*a referendum*	*b*	*European Union*
b	*a new treaty*	*d*	*none of a-c*

15. What does the referendum result suggest?

Answers:

a	*Mitterand should retire*	*b*	*Mitterand is popular*
c	*French politics needs revising*	*d*	*none of a-c*

12. In what manner should France proceed into European Union?

Answers:

a	*enthusiastically*	*b*	*gradually*
c	*confidently*	*d*	*none of a-c*

14. Which countries are named as being problematic for European Union?

Answers:

a	*Britain*	*b*	*Germany*
c	*Denmark*	*d*	*none of a-c*

13. Which of the following contributed to the collapse of the EMS?

Answers:

a	*French uncertainty*	*b*	*too much flexibility*
c	*unrealistic rates*	*d*	*none of a-c*

Appendix 4: Linkage occurring in Different Texts

This Appendix shows, for each different text, the words which are involved in linkage. The words involved in linkage at each of L1 and L2 lookup levels are listed underneath the relevant text number. The frequency of linkage associated with each of the words, at each level of lookup, appears alongside each word.

TEXT 1: L1 Linkage		L2 Linkage	
industry	5	industry	6
market	5	market	6
sales	5	games	5
toy	2	sales	2
worth	1	forecast	1
fast	1		
games	1		
Total L1:	**20**	**Total L2:**	**20**

TEXT 2: L1 Linkage		L2 Linkage	
stocks	5	systems	15
product	4	stocks	12
result	3	product	10
systems	3	results	7
profits	3	success	7
		simplicity	7
		essential	6
Total L1:	**18**	**Total L2:**	**64**

TEXT 3: L1 Linkage		L2 Linkage	
designer	27	designer	14
attention	14	attention	12
threat	8	show	10
facing	5	award	10
show	5	collection	9
award	4	fraction	8
foreward	4	facing	7
fashion	4	threat	4
industry	2	industry	4
fraction	1	Smith	1
collection	1	foreward	1
Total L1: 75		**Total L2:**	**80**

TEXT 4: L1 Linkage

development	12
considerations	11
reduction	10
costs	8
advances	8
cooling	5
fluid	1
efficiency	1
Total L1:	**56**

L2 Linkage

reduction	16
development	16
advances	15
cooling	10
considerations	9
costs	7
overheating	5
efficiency	1
fluid	1
Total L2:	**80**

TEXT 5: L1 Linkage

gallery	4
house	2
opera	2
development	2
housing	1
sterling	1
Total L1:	**12**

L2 Linkage

house	11
gallery	9
housing	3
development	3
Total L2:	**26**

TEXT 6: L1 Linkage

death/s	39
key	16
characters	12
rhythm	9
ratings	8
writer	8
killing	7
corporation	3
opera	2
bomb	1
Total L1:	**105**

L2 Linkage

death/s	56
characters	21
ratings	14
killing	14
soap	11
key	11
publicity	7
rhythm	6
opera	3
writer	2
bomb	1
Total L2:	**139**

TEXT 7: L1 Linkage		L2 Linkage	
book	22	time	23
prize	21	version	22
writing	21	prize	21
version	21	book	15
time	18	sense	14
first	13	first	12
year	12	writing	9
poem	7	year	9
half	3	going	9
sense	2	half	8
University	2	poem	3
awarding	2	University	1
going	1	awarding	1
Total L1:	**145**	**Total L2:**	**147**

TEXT 8: L1 Linkage		L2 Linkage	
research	1	trust	4
University	1	fraud	4
trust	1	Chapman	2
fraud	1	University	1
		revealing	1
Total L1:	**4**	**Total L2:**	**12**

TEXT 9: L1 Linkage		L2 Linkage	
strategies	3	rubbish	9
reduction	2	refuse	9
plan	2	sending	4
recycling	2	reduction	4
waste	2	cooperation	3
sending	1	strategies	3
cooperation	1	population	3
refuse	1	recycling	2
		waste	2
		tons	1
Total L1:	**14**	**Total L2:**	**40**

TEXT 10: L1 Linkage		L2 Linkage	
homes	50	homes	64
valve	25	levels	33
unit	23	unit	30
people	15	amount	24
levels	15	source	23
amount	14	value	21
government	13	housing	18
source	9	government	16
boom	8	goods	1
housing	5	people	9
goods	4	leading	5
consumer	3	consumer	5
reluctance	1	percent	5
leading	1	boom	2
Total L1:	**186**	**Total L2:**	**265**

TEXT 11: L1 Linkage		L2 Linkage	
voters	3	task	4
referendum	3	aftermath	3
voting	2	victory	3
minority	2	minority	3
treaty	2	size	2
favour	1	referendum	2
future	1	treaty	2
		future	1
		favour	1
		voting	1
Total L1:	**14**	**Total L2:**	**22**

TEXT 12: L1 Linkage		L2 Linkage	
vote	7	vote	21
minority	7	commitment	16
population	6	institutions	16
expression	6	minority	15
confidence	6	population	12
commitment	5	progress	7
wish	4	warnings	7
referendum	3	expression	6
institutions	3	referendum	5
progress	2	treaty	2
treaty	1	wish	2
		confidence	1
Total L1:	**50**	**Total L2:**	**110**

TEXT 13: L1 Linkage

currency	26
states	11
member	8
convergence	6
reasons	6
insistence	5
factor	5
led	4
system	4
treaty	2
referendum	2
collapse	2
crisis	1
Total L1:	**82**

L2 Linkage

currency	23
system	16
member	14
collapse	13
convergence	11
factor	9
crisis	8
reasons	5
insistence	5
states	3
referendum	2
Total L2:	**99**

TEXT 14: L1 Linkage

beginning	14
introduction	12
implementation	7
ratification	5
vote	5
system	5
Parliament	3
hurdle	3
priorities	3
policy	3
future	3
treaty	2
reform	1
market	1
timetable	1
Total L1:	**68**

L2 Linkage

introduction	24
system	23
beginning	20
policy	18
vote	18
reform	18
hurdle	16
ratification	9
market	6
implementation	6
Parliament	4
clear	4
priorities	3
treaty	3
future	2
still	2
Total L2:	**176**

TEXT 15: L1 Linkage

vote	12
alliances	9
voting	7
minority	7
politics	6
treaty	5
term	5
public	5
affairs	5
confidence	3
referendum	2
lessons	2
Total L1:	**68**

L2 Linkage

vote	32
alliances	19
public	19
minority	18
politics	13
term	9
affairs	9
treaty	5
referendum	5
voting	5
lessons	4
confidence	3
victory	3
Total L2:	**144**

TEXT 16: L1 Linkage

step	11
formation	9
beginning	8
progress	6
expansion	4
community	3
policy	3
scheme	2
treaty	2
ratification	2
Total L1:	**50**

L2 Linkage

progress	16
formation	14
beginning	14
policy	13
expansion	9
scheme	8
ratification	6
step	6
community	4
treaty	2
Total L2:	**92**

TEXT 17: L1 Linkage

recognition	8
use	7
focus	4
foundation	4
referendum	3
need	3
president	3
campaigns	2
Total L1:	**34**

L2 Linkage

need	8
use	7
focus	7
recognition	6
foundation	5
victory	4
president	3
referendum	2
campaigns	1
effectiveness	1
Total L2:	**44**

TEXT 18: L1 Linkage	
votes	8
voting	6
legislation	4
getting	3
resignation	2
house	2
attempts	1
Total L1:	**26**

L2 Linkage	
votes	14
house	7
voting	6
resignation	4
legislation	4
treaty	3
Total L2:	**38**

TEXT 19: L1 Linkage	
motion	4
development	3
bill	2
importance	1
Total L1:	**10**

L2 Linkage	
development	10
motion	9
importance	5
bill	5
essential	4
majority	3
Total L2:	**36**

TEXT 20: L1 Linkage	
services	15
principles	11
taxes	8
mechanism	7
rates	7
economy	7
sterling	6
policy	6
collapse	5
party	5
spending	5
public	3
plans	2
treaty	1
Total L1:	**88**

L2 Linkage	
services	24
mechanism	18
party	15
policy	12
principles	12
public	9
rates	9
collapse	8
plans	6
borrowing	5
spending	5
economy	4
sterling	2
treaty	1
Major	1
Total L2:	**131**

Appendix 5: Published Papers

This appendix contains two publications arising from the research described in this thesis.

Norris, J. (1996) Compound Nominal Generation for Information Retrieval : The COMMIX System, *Proceedings of the Workshop on Language Engineering for Document Analysis and Recognition (LEDAR)*, AISB 1996 Workshop Series, 2nd April 1996, pp. 48-55. Sussex University, Brighton, UK.

Norris, J. (1997) Extracting the Essence from Text: a Computational Approach, *Proceedings of the International Workshop on Lexically Driven Information Extraction*, July 16th 1997, pp. 63-80. Frascati, Italy.

In addition to these publications, the work was described in its early stages at:

the *Workshop on the Unified Lexicon* at the Speech and Language Technology (SALT) club meeting, 15-17th December, 1993, held at St. Aidan's College, University of Durham.

In this talk, the methodology as described in Chapter Four of this thesis was presented, along with some early results.

Compound Nominal Generation for Information Retrieval: The COMMIX System.

Jennifer Norris
Information Technology Research Institute, University of Brighton
Lewes Road, Brighton BN2 4AT
Jenny.Norris@itri.bton.ac.uk

Abstract.
This paper describes a novel approach to the content analysis of a piece of text, and specifically addresses the problem of providing a concise description of the subject matter of abstracts. It describes the COMMIX system, whose practical purpose is to analyse a given abstract and generate elaborate, novel and concise compound nominal terms which express the 'aboutness' of the abstract without relying on simple reproduction of sections of the text or key lexemes occurring within the text. The system is still under development, but some initial results are presented and briefly discussed in this paper.

Introduction.

Having read a piece of informative text it is not always easy to specify precisely what that text is about. It may be about a simple topic, or may cover several distinct but related things. In many cases a specific topic is not overtly named, but is a complex 'object' constructed throughout the course of the text. In such cases, attempts to describe the 'aboutness' of the text may result in an extended piece of prose (when there is no corresponding individual lexical item).

This clearly presents a problem from the Information Retrieval point of view. A user with a complex query should be able to refer to that query in a compact and reproducible way, allowing them to retrieve relevant documents without the enormous waste of time inherent in searching techniques based on the Boolean combination of sets. Information retrieval systems should be able to utilise compact and concise query expressions as well as more straightforward terms. This may be achieved by the system generating its own reliable and concise content expressions for each document (in the case of this work, for each abstract) against which to match the query term provided by the user.

The compound nominal is an appropriate type of construct for use in pursuit of this aim: it allows for a very high degree of compaction of information into an expression comprising a head noun along with what may be a highly complex modifying expression. The work described in this paper concerns the construction of novel compound nominal terms which express the aboutness of an abstract (and therefore the document to which it refers), and which can be used for whole or partial matching against complex user queries. The bulk of the paper specifies the motivation, assumptions, general approach taken and specific methodology employed, along with some initial results, but begins with a brief discussion of the notion of a complex query and the compound nominal as a mode of its expression.

1. Complex Queries.

A user with a complex query has a lot of detail to convey in the query term, and may formulate an expression to convey it in as concise a way as possible. For example, the user may want to know how much money is spent annually on the industry associated with video games. If the user were to ask a direct question of the system, (i.e. "How much money is spent annually on the industry associated with video games?") this would require at least some degree of parsing, and would still not provide the system with any usable concise search term. An alternative strategy is to state the query as a 'thing' or topic about which more information is required. This is not as complicated a task (for a human) as it might at first seem, and the user can have in mind the wording of the directive: "Tell me what you know about 'query' ", which in the above example might yield the query term 'video games industry spending'.

This type of nominal expression has received much attention over the past twenty years or so, and has attracted various terminology, including 'complex nominals' (Levi, 1978) and 'nominal compounds' (Bauer, 1979). It is not clear, however, that these different terms are used consistently, partly due to a lack of clear definitions of the linguistic phenomena under study. For this reason the term used in this work is **compound nominal**.

Compound nominals, then, offer a linguistic mechanism for expressing complex queries in a way which is concise, and requires their expression in the form of nouns. They are not difficult for people to construct, and users with complex queries may be advised to express queries in the form of the general "Tell me what you know about 'query' " instruction. If applied to the analysis of the 'topic' of a document, their use can significantly improve the 'aboutness' expressions whose aim is to describe the content of that document.

1.1 Matching a complex query to relevant documents.

Although it is reasonable to assume that an abstract will contain terms which are more concise than those appearing in the corresponding extended text, it is rare to find such a degree of conciseness typical of the type of highly compacted compound nominal query term advocated here. Such a high degree of compaction of information is unusual in running text, being of a different style, and being more difficult to parse.

The complex topic itself (about which a user has a query) is, then, unlikely to be expressed in relevant documents (or their abstracts) in terms of that topic, but rather as information about some simpler topic. Thus, for example, an article about video games may refer to various pieces of information about video games: it may tell us that the associated industry is growing, or that it is taking over the toy market, or that specific amounts of money are spent annually on it by specific companies. But it is unlikely to utilise very highly specific terms of reference for the information it covers: thus, the text is unlikely to contain such terms as 'video games industry growth', 'toy market domination' (or indeed the highly specific 'video games industry toy market domination') or 'video games industry spending', even though each of these topics may be judged as being what the article is at least partly about.

2. Motivation.

This section describes the practical and theoretical motivation behind the system.

2.1 Practical motivation.
The practical motivation centres around improving access to relevant information by enabling an IR system to generate highly specific expressions representing the content of a given abstract. Current methods which rely on the Boolean combination of sets relating to parts of a query are lacking in both precision and recall (i.e. they lead to the retrieval of irrelevant documents, and do not retrieve all relevant ones). It seems clear that the accurate specification of 'what an abstract (or other document) is about' in concise terms will improve access to relevant information. The 'aboutness concepts' (of a piece of text) referred to in this work are a type of complex 'topic', and lend themselves well to the concise expression of compound nominals. Thus, the COMMIX system has been developed with the aim of generating a number of compound nominal expressions to represent what a given piece of text (an abstract) is about. A front end user query facility would enable it to be used to match compound nominal query terms provided by users against those generated by the system as representations of the aboutness concepts of the abstracts.

2.2 Theoretical Motivation.
The theoretical motivation behind this work has been to explore the hypothesis that there is sufficient information present in the dictionary definitions of individual lexemes occurring in a piece of text, to establish semantic relatedness between terms, and thereby produce novel compound nominal expressions which represent the aboutness of the text.

3. Underlying Assumptions.
Throughout this work the following assumptions have been assumed to hold true:
1) There are concepts relating to complex referents, which an text is about, and which are developed throughout the course of that text, but which are not necessarily accorded linguistic 'labels' overtly within the text.
2) Compound nominals offer a linguistic means of compacting large amounts of information into premodified nominal form, and facilitate the expression of such complex conceptual entities.
3) There is a direct relationship between the salience of a term and its frequency of occurrence within a piece of text. To some extent the salience of an item is reflected in its position in the text, particularly when it occurs in the first sentence. The salience of a complex concept is assumed to be directly related to the combined strengths of the linkages leading to it in a semantic network.
4) Any verbal information (except that associated with modal or auxilliary verbs) can be expressed and utilised in nominal form in the fomulation of compound nominal terms which express the aboutness of a piece of text.

4. General Approach
The general approach adopted in this work centres on the hypothesis that semantic relatedness can be established by matching words which appear in the definitions of distinct parent terms. Once semantic relatedness has been established, compound nominal terms representing aboutness concepts can be automatically generated by identifying the most salient nouns in the original abstract and applying to those nouns (or their hypernms, or synonyms) some or all of the modifiers which apply to semantically related nouns, as well as noun forms of associated verbs. The system can then be used as a tool for investigating and specifying the constraints on the generation of meaningful compound nominal aboutness terms using this approach.

4.1 Establishing Semantic Relatedness.

The assumptions specified above, in conjunction with the general approach, can be expressed as a set of axioms which may be applied to any abstract. Thus, given an abstract in which nouns N1 and N2 appear, with respective modifiers M1 and M2, the application of the above assumptions leads to the following assertions:-

a) If N1 = N2 (= N), then given a piece of text in which M1N1 and M2N2 both feature, a compound nominal term M1M2N can be constructed and deemed at least partly representative of the aboutness of that text. In this case the relation N1:N2 is the highest possible (because they are the same noun[41]), and may be referred to as R1.

b) If N1 and N2 are different nouns, but N2 appears in the definition of N1, then they can be assumed to be semantically related, (to a weaker degree than in a)) and on that basis, there is justification for applying at least some of the modifiers of N2 to N1 in the construction of compound nominal aboutness expressions.

c) If N1 and N2 are different nouns, but there is a word, W, which is common to their definitions, then they can be assumed to be semantically related, although to a weaker degree than in a) and b). In this case there is justification for applying the modifiers of both nouns to each noun independently.

The degree of relatedness between N1 and N2 varies according to the number of words which are common to their definitions: the more terms which the definitions of two terms have in common, the greater is assumed to be their semantic relatedness.

These axioms may be summarised as follows:

- N1 = N2 => strongly related, with strength R1
  -> M1M2N
- N1 /= N2 but N2 **MEMBER_OF** def (N1)
  => strength R2, where R2 < R1
  -> M1 (some of M2) N1
- N1 /= N2 & W **MEMBER_OF** def (N1) **INTERSECTION** def (N2)
  => strength R3 where R3 < R2
  -> M1M2N1 & M1M2N2

R **PROPORTIONAL_TO** no. words in def (N1) **INTERSECTION** def (N2)

5. Method

The methodology is based on the assumptions and general approach described above. For the sake of clarity, the description of the methodology for processing an abstract will refer to the following example abstract[42]:

> *The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling.*

---

[41] Ambiguity of word sense tends not to be a problem in the genre of human-produced abstracts (being short, fairly concise texts), since an abstractor will use a synonym or paraphrase rather than create an ambiguity.

[42] The original text to which this abstract refers appeared in The Observer of Sunday 11th October 1992.

The main competitors in the market are Sega and Nintendo. Nintendo will spend 15 million pounds sterling on advertising over Oct-Dec 1992.

The processing of an abstract falls into the following stages:-

1) Sorting and labelling of the original abstract.

This involves the deletion of all closed class words and the labelling of all open class words according to their syntactic word class (as listed in the WordNet database). If an item is not listed in the database, it is retained and labelled '*n_I*'. Full labelling of the abstract involves numbering items according to both sentence number and word number, and leaving a trace of deleted items, indicating the sort of item that has been deleted. Labelling results in a list of constituents of the form

[sentence number word number item syntactic class of item]

where 'item' is the item itself, or DEL if it has been deleted as a closed class word. For the example abstract, the first 2 sentences thus labelled are as follows:

1 1 DEL det 1 2 video noun 1 3 games noun 1 4 industry noun 1 5 DEL aux 1 6 growing adj 1 7 fast noun 1 8 DEL conj 1 9 DEL modal 1 10 dominate verb 1 11 DEL det 1 12 toy noun 1 13 market noun 1 14 DEL conj 1 15 become verb 1 16 DEL det 1 17 established adj 1 18 part noun 1 19 DEL prep 1 20 home noun 1 21 entertainment noun 1 21 punc_s 2 22 DEL det 2 23 1991 number 2 24 computer noun 2 25 games noun 2 26 market noun 2 27 DEL aux 2 28 worth noun 2 29 275 number 2 30 million noun 2 31 pounds noun 2 32 sterling adj 2 33 growing adj 2 34 DEL prep 2 35 500 number 2 36 million noun 2 37 DEL prep 2 38 1992 number 2 38 DEL punc 2 39 half noun 2 40 DEL det 2 41 toy noun 2 42 market noun 2 42 . punc_s ... etc.

This detailed labelling of the abstract is not utilised to the full at this stage of the implementation, but will be required at a later stage when constraints (involving sentence number, the position and type of punctuation, the type and position of deleted items) are introduced. In this implementation it is only the syntactic word class information which is utilised.

2) Initial processing of labelled abstract.

This stage involves two processes: firstly, the identification of compound nominals which already exist in the abstract; secondly, the recording of the frequencies of all nouns which occur more than once (and are thus deemed to be 'salient' to some degree). The first stage yields the following in our example:

[video games industry] [growing fast] [toy market] [established part] [home entertainment] [1991 computer games market] [worth 275 million pounds sterling growing] [500 million] [toy market] [hardware sales] [635 million pounds sterling] [associated software sales] [645 million pounds sterling] [worth 345 million pounds sterling] [main competitors] [spend 15 million pounds sterling] [Dec 1992]

The result of the second stage is a list of the salient nouns (ie those which occur more than once), alongside their frequency of occurrence in the abstract. The result of this second stage in our example is the following list:

[games 2 toy 2 market 5 worth 2 million 6 pounds_sterling 5 sales 2]

3) Lookup salient nouns and Filter glosses.

This is the point at which the defining gloss for each salient noun is looked up. The processing of each gloss involves the filtering of words which do not pertain to the definition (such as *senses* in *there are four senses of ...*), and the deletion of all closed class words, followed by the labelling of all open class words, as in stage 1. This stage takes as input, the list of salient nouns shown above and produces, for each one, a gloss from which all the closed class words have been removed. For the

purposes of this example, the glosses for 'sales' and 'market' should suffice to show how relatedness (linkage) is established.

GLOSS FOR sales:

trade sales desperate boost sales commercial enterprise business enterprise business purchase sale goods services sale sale cut sale sales event selling specially reduced prices sale reduce inventory selling merchandising marketing exchange goods agreed sum money sale particular instance selling made three sales hour selling merchandising marketing exchange goods agreed sum money

GLOSS FOR market:

market customers particular product service class social class people sharing common attribute market securities industry securities markets aggregate market always rustates small investor industry people engaged particular kind commercial enterprise grocery store grocery market supermarket marketplace mart market marketplace commerical activity whereby good services exchanged competition there market activity behavior specific action pursuit avoided recreational activity senses market market deal market ransact deal business market sale produce sale trade deal merchandise engage trade commercialize market make commercial change alter cause change make different

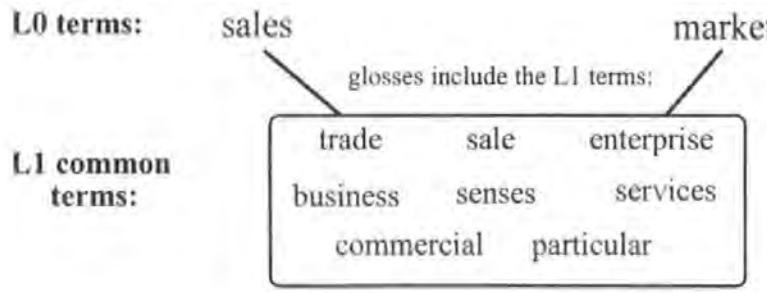
The glosses are then compared against one another in the search for matching items, which will justify the formation of a link between their respective parents.

4) Matching Terms.

This stage involves searching for terms which are common to more than one gloss. As mentioned above, a link between two parent items is held to be justified if their respective glosses contain common terms. In the example, there are seven words which are common to the glosses of the parent (L0[43]) terms *sales* and *market*:-

[trade commercial enterprise business sale services particular]

These common terms constitute the L1L1 links which link the parent L0 terms *sales* and *market*, as shown below:-



The L1 common terms (L1L1 links) are held to justify the assumption of semantic linkage between the two L0 terms *sales* and *market*, and as such are the justification for applying all the modifiers of both L0 terms to the most salient of the two parent nouns.

5) Weighting the links.

Once all the commonalities between the glosses of the salient nouns have been identified, the frequency of each matched term (i.e. link) throughout the glosses is recorded. This frequency is to be used subsequently in the establishment of a salience weighting value which reflects the strength of the justification for the linkage. This

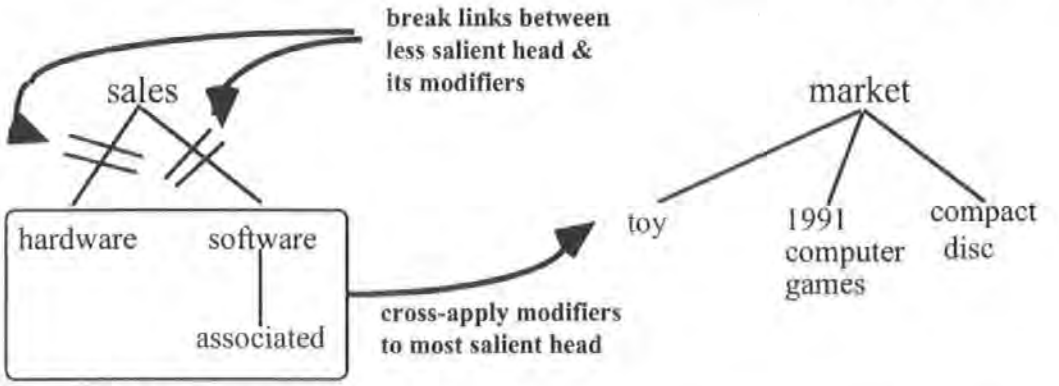
[43]The term L0 refers to words from the original text; L1 refers to definition words resulting from 1 Lookup process (and L2 to words in the definitions of L1 words).

information is not utilised at this stage, but is used to represent relative salience at a later stage of the implementation.

6) Selection of head noun and cross-application of modifiers.

If linkage has been found between two or more items which occur as salient nouns in the abstract, then there is justification for combining their modifiers and applying them all to the most salient of related nouns (the one occuring with the highest frequency in the original abstract). This stage identifies the appropriate noun and applies to it all the modifiers which relate either to that noun or to semantically related nouns. In the example, all the modifiers applying to each of 'sales' and 'market' in the original abstract can be represented by the following tree structures:-

The most frequent (and thus the most salient) of *sales* and *market* is found to be *market,* which becomes the head of the compound nominal being formed. The 'links' between the less salient noun (ie *sales* in the example) and its modifiers are broken, and the modifiers are cross-applied to join those which already apply to *market.*



This yields the expression:

toy 1991 computer games hardware associated software compact disc market.

6. Discussion

The above expression is the longest of those produced by the system at this stage. Although it is difficult to imagine a user coining such a query term, it does provide a surprising degree of insight into the aboutness of the example abstract. Considering the crudeness of this stage of the implementation, this is encouraging with regard to the possibilities for tuning the technique in later implementations.

It should be noted that the method as described above only pursues possible linkage through one level of lookup. In fact, investigations are currently under way into the effect on the linkage of extending the lookup to include words appearing at the L2 lookup level: in other words, to include words appearing in the definitions of defining words of parent terms. For the example abstract, very recent work at this extended level yields additional aboutness terms, including the following:

home toy, established toy, video games market, video games sales,
compact disc sales, toy sales, 1991 computer games sales,

all of which are novel yet highly representative of the aboutness of the abstract.

7. Conclusion.

Although the system is still without any constraints on the construction of the compound nominal aboutness expressions, its performance is encouraging. The technique described, which relies very heavily on the semantic information implicit within the dictionary definitions of terms, whilst minimising syntactic processing, produces some useful expressions. Whilst some of these are not particularly representative of content, many do express the aboutness of the input abstract with a fair degree of accuracy. Work is continuing on the formalising of the constraints necessary to improve the quality of the output expressions.

References.

Bauer, L. (1979) On the need for pragmatics in the study of nominal compounding. *Journal of Pragmatics* **3**, pp 45-50

Levi, J. (1978) *The syntax and semantics of complex nominals.* Academic Press, NY

Extracting the Essence from Text: a Computational Approach.

Jennifer Norris
Information Technology Research Institute
University of Brighton
Lewes Road, Brighton BN2 4AT
email: Jenny.Norris@itri.bton.ac.uk

Abstract

This paper describes the COMMIX system, which automatically extracts information on what a short piece of text (an abstract) is about, and generates that information in the highly compacted form of compound nominal expressions. The expressions generated are complex and may include novel terms, which do not appear themselves as items in the input text. From the practical (applications) point of view, the work is driven by the need for an additional level of representation which lies between an abstract and its 'key words' or indexing terms: in other words, a representation which is shorter and more concise than would appear in an abstract, yet more informative and representative of the actual aboutness than commonly occurs in indexing expressions and key terms. This additional layer of representation is referred to in this work as pertaining to the *essence* of a particular text. The driving force of this work, from the theoretical perspective is the exploration of the claim that there is sufficient semantic information contained within the standard dictionary glosses for individual words to enable the construction of useful and highly representative novel compound nominal expressions, without recourse to standard syntactic and statistical methods.

The methodology employed, and described herein, is domain-independent, and does not require the specification of templates with which the input text must comply. In these two respects, the methodology developed in this work avoids two of the most common problems associated with information extraction.

1 Introduction

The COMMIX system has been designed in part in response to an 'information gap' which currently exists in the representations of what a particular text is about. Current representations of the information conveyed by a text vary in their degrees of conciseness, compactness, succinctness and informativeness. We have the precis, the summary, the abstract, key- expressions and -words, and other indexing expressions. The first three of these types of representations occur in prose form, and may be said to represent the *gist*[44] (or the main expositional arguments) of the

[44]The term *gist* has been used (eg Scott, 1993; Rino, 1996) to refer to the salient information and arguments presented in a text. However, the distinction between this term and *aboutness* (of which *essence* is the most informative type) is worth noting at this point. The gist of a text refers to both what that text is about, and to the arguments contained within the exposition, and any conclusions that may be drawn from those arguments. As such, the gist cannot be represented as an entity, or a 'thing', and cannot therefore simply consist of a nominal expression, however complex. On the other hand, the essence of a text, or what it is (mainly) about, may be expressed as a highly informative nominal

original text. The latter two, however, constitute attempts to represent the *aboutness* [45] of the text (the topics, or 'things' that the text is about), and these are terms which occur in nominal form.

Yet there is as yet no good way of automatically representing the aboutness of a text in a highly informative way. There are many approaches which are based heavily on word frequencies (e.g. Jacobs & Rau, 1990; Paice 1990; Hodges et al, 1996), and these play a significant role, and contribute to a degree to the task of representing aboutness. They do, however, usually yield very general aboutness terms, as are prevalent in a typical index. Resulting expressions are generally single or two-word items, and give the overall subject area of the text. The use of template helps in the extraction of the required information, but this limits both the domain of the text and the type of information extracted.

These methodologies also fail to take into account the relatedness between two different words which may be at least partly synonymous. The result is that when full or partial synonyms are encountered they simply count as completely distinct tokens, rather than contributing to an increase in the importance factor attributed to related items. Any highly informative indexing (particularly in cases where terms which do not appear in the original text are used) still relies heavily on human involvement.

This paper describes a new method being developed for the extraction of information from a text at one level of representation, and its automatic conversion into a more concise, highly informative form. This work is still in progress, but some initial results are presented and discussed. The final section of the paper includes a discussion of the limitations of this work, along with some suggestions for possible future directions.

2 Standard Information Extraction

The production of a full Information Extraction (IE) system (which conforms to, for example, the requirements of the MUC participants) has never been part of the aim of this work. Most IE systems, such as PROTEUS (Grishman, 1996), aim to identify and extract all information pertaining to a set of prespecified events, within a limited domain, and store that information in prestructured templates. The methodology employed by such systems generally maximises the amount and detail of synytactic processing, utilising a fully parsed version of the text being processed, and relying on a domain specific knowledge base for its semantic processing.

In contrast, throughout the production of the COMMIX system, a very different approach has been taken. Rather than relying heavily on syntactic processing, detailed domain knowledge and predetermined templates, the aim behind this work has been to minimise the degree of reliance on syntactic and statistical measures, and

---

expression, of which the compound nominal is only one (albeit highly concise) variety. An expression of the essence of a text tells readers what the text is about in a highly informative way, and would rarely (if ever) comprise single-word expressions as commonly occur as indexing items.

[45]This term was used by Hutchins (1977) in a general way, but has appeared little since. It is not clear that his usage of the term requires the related expressions to be nominal, as is the usage in this thesis.

to avoid the use of templates altogether, whilst keeping the system domain-independent. The approach aims to place maximum emphasis on the full utilisation of the large amounts of semantic information which are included (implicitly) within the standard gloss definitions of terms, as they appear in a standard dictionary. As will be shown below, the methodology does include a degree of syntactic processing, but the reliance on syntax is minimised by adopting a number of assumptions which may be applied to the realm of the abstract.

In comparing this work to standard IE systems, it is important to remember that the COMMIX system does not aim to retain as much of the original information about a text as do typical information extraction systems, but rather aims to give a highly informative indication of its aboutness. As such, its true place lies between the realms of indexing systems on the one hand, and full IE systems on the other. What this system does specifically aim at, is the extraction of the essential aboutness of a text, and the representation of that essence as highly informative aboutness expressions, in very concise, compact nominal form.

3 How the COMMIX system tackles the problem

A standard lexicon, as represented by a standard dictionary, provides a huge amount of semantic information which can be exploited in the enterprise of expressing, in highly compacted terms, the aboutness of a text. The methodology developed for use in the COMMIX system relies crucially on this semantic information contained within the definitions of words appearing in an electronic dictionary.

It should be noted that the system in principle would utilise a standard dictionary, such as LDOCE. However, for reasons of practicality and economy the initial implementation of the COMMIX system uses the Wordnet database. Some of the consequences of this choice are discussed in section 6.2 below. It is worth mentioning at this point, however, that in order to get as close as possible to the dictionary glosses for items (as occur in a standard dictionary), the usage of Wordnet has been restricted to the immediate glosses (or synonym sets) given for individual words.

3.1 The Underlying Assumptions

Throughout this work the following assumptions have been assumed to hold true:

- There are concepts relating to complex referents, which an text is about, and which are developed throughout the course of that text, but which are not necessarily accorded linguistic 'labels' overtly within the text.

- Compound nominals offer a linguistic means of compacting large amounts of information into premodified nominal form, and facilitate the expression of such complex conceptual entities.

- There is a direct relationship between the salience of a term and its frequency of occurrence within a piece of text. To some extent the salience of an item is reflected in its position in the text, particularly when it occurs in the first sentence. The salience of a complex concept is assumed to be directly related to the combined strengths of the linkages leading to it in a semantic network.

- Any verbal information (except that associated with modal or auxilliary verbs) can be expressed and utilised in nominal form in the fomulation of compound nominal terms which express the aboutness of a piece of text.

3.2 General Approach

The general approach adopted in this work centres on the hypothesis that semantic relatedness can be established by matching words which appear in the definitions of distinct parent terms. Once semantic relatedness has been established, compound nominal terms representing aboutness concepts can be automatically generated by identifying the most salient nouns in the original abstract and applying to those nouns (or their hypernyms, or synonyms) some or all of the modifiers which apply to semantically related nouns, as well as noun forms of associated verbs. The system can then be used as a tool for investigating and specifying the constraints on the generation of meaningful compound nominal aboutness terms using this approach.

3.3 Establishing Semantic Relatedness

The assumptions specified above, in conjunction with the general approach, can be expressed as a set of axioms which may be applied to any abstract. Thus, given an abstract in which nouns N1 and N2 appear, with respective modifiers M1 and M1, the application of the above assumptions leads to the following assertions:

a) If N1 = N2 (= N), then given a piece of text in which M1N1 and M2N2 both feature, a compound nominal term M1M2N can be constructed and deemed at least partly representative of the aboutness of that text. In this case the relation N1:N2 is the highest possible (because they are the same noun[46]), and may be referred to as R1.

b) If N1 and N2 are different nouns, but N2 appears in the definition of N1, then they can be assumed to be semantically related, (to a weaker degree than in a)) and on that basis, there is justification for applying at least some of the modifiers of N2 to N1 in the construction of compound nominal aboutness expressions.

c) If N1 and N2 are different nouns, but there is a word, W, which is common to their definitions, then they can be assumed to be semantically related, although to a weaker degree than in a) and b). In this case there is justification for applying the modifiers of both nouns to each noun independently.

[46]Ambiguity of word sense tends not to be a problem in the genre of human-produced abstracts (being short, fairly concise texts), since the writer will typically use a synonym or paraphrase rather than create an ambiguity.

These axioms may be summarised as follows:

- N1 = N2 => strongly related, with strength R1
  -> M1M2N
- N1 /= N2 but N2 **MEMBER_OF** def (N1)
  => strength R2, where R2 < R1
  -> M1 (some of M2) N1
- N1 /= N2 & W **MEMBER_OF** def (N1) **INTERSECTION** def (N2)
  => strength R3 where R3 < R2
  -> M1M2N1 & M1M2N2

The degree of relatedness between N1 and N2 varies according to the number of words which are common to their definitions: the more terms which the definitions of two terms have in common, the greater is assumed to be their semantic relatedness. In short:

R **PROPORTIONAL_TO** no. words in def (N1) **INTERSECTION** def (N2)

3.4 Assumption of Semantic Relatedness

This section addresses the assumption that semantic relatedness can be established as described above in section 3.3. The first sub-section comprises theoretical justification based on the work of some other authors, whilst the second presents the results of some initial word pair relatedness findings, in terms of the numbers of words appearing in the intersection of sets of gloss words.

3.4.1 A theoretical justification

The basis of this justification comes from a different, but related fieldof linguistic analysis: that of word sense disambiguation. There are a number of authors in the field of lexical disambiguation whose work utilises the assumption that concepts which are related to one another will be represented in a dictionary or thesaurus by the same words. Some early examples of this approach are mentioned by Lesk (1986), such as Masterman (1957), who suggested looking up the thesaurus entries for different senses of a word, and counting the words which overlapped with those for adjacent words in the text. Lesk (ibid, p.3) also mentions unpublished work done by Lawrence Urdang in the 1960s, who used a dictionary rather than a thesaurus for the same purpose.

Lesk (ibid) describes his own implementation of dictionary disambiguation, using a machine-readable dictionary, but based on the same principle: that words which are related to each other will have some of the same words appearing in their definitions.

Rigau et al (1997) combine a variety of methods in their disambiguation system. Their "Heuristic 4: Word Matching" is used to count the total number of content words shared between a given hyponym and a candidate hypernym, and is once again based on this same principle. In their own words: " *This heuristic trusts that related concepts will be expressed using the same content words.*"

There are clearly differences between the word sense disambiguation work and the work described in this paper, in terms of their respective aims. Whereas the former utilises adjacent words and often incorporates cooccurrence data (eg Rigau et al, 1997), the COMMIX system identifies relatedness between words which do not necessarily appear together, even in the same sentence[47]. However, the basic assumption, that semantic relatedness between distinct items can be indicated (to some extent) by an overlap in the words appearing in their glosses, is the same.

In an effort to avoid simply 'taking it on trust', a small-scale word-pair analysis of overlapping words has been carried out, in relation to the abstract used in the development of the COMMIX system. The following sub-section presents the results of this small investigation.

3.4.2 A small-scale empirical justification

For this small-scale investigation into the overlap of gloss words from distinct words, the text used is the abstract which is discussed in detail in sections 4 and 5 of this paper. The full abstract appears in section 4. The overlap investigation was carried out on two sets of nouns: the first set comprised only those nouns which count as salient in the methodology employed by COMMIX in its basic ('pure') form (this notion being based solely on its having a frequency within the text of greater than one). The second set includes other nouns which are also considered salient, as indicated by the fact that they are already modified in the input text, even though they only occur once.

3.4.2.1 Method

The method used to investigate the word pair overlaps was as follows:
For each set of salient nouns (this is explained in detail in section 4, which details the methodology behind the whole system), all possible word pair combinations were identified. For both members of each word pair, the individual (WordNet) entries were looked up, and the number of words common to both entries were recorded.

3.4.2.2 Results

According to the methodology described in this paper, for the system in its most basic form, there are 7 salient nouns: games; toy; market; worth; million; pounds_sterling; sales. Allowing for the commutativity of pairing, this gives 21 possible word pair combinations.

For the extended notion of salience, there are 3 additional nouns: industry; entertainment; software. In combination with the nouns occurring in the first set, this yields a further 24 possible word pairs.

Tables 1 and 2 show the results of the search for overlap between entries for each of the word pair combinations, where n is the number of words which the entries have in common.

---

[47] It is important to remember that the COMMIX system and its underlying assumptions relate to the realm of the abstract. It is not a claim of this work that the methodology would be usefully applied to full, extended text (see Section 3.2).

Table 1. Nouns occurring more than once

Word Pair	n
games - toy	2
games - market	8
games - worth	0
games - million	1
games - pounds_sterling	0
games - sales	3
toy - market	0
toy - worth	1
toy - million	1
toy - pounds_sterling	0
toy - sales	4
market - worth	1
market - million	0
market - pounds_sterling	0
market - sales	3
worth - million	4
worth - pounds_sterling	0
worth - sales	2
million - pounds_sterling	0
million - sales	2
pounds_sterling - sales	0

Table 2. Word Pairs arising from additional nouns

Word Pair	n
industry - games	2
industry - toy	1
industry - market	11
industry - worth	0
industry - million	0
industry - pounds_sterling	0
industry - sales	4
entertainment - games	7
entertainment - toy	0
entertainment - market	1
entertainment - worth	0
entertainment - million	0
entertainment - £_sterling	0
entertainment - sales	1
software - games	1
software - toy	1
software - market	1
software - worth	1
software - million	1
software - £_sterling	0
software - sales	1
industry - entertainment	1
industry - software	0
entertainment - software	0

3.4.2.3 Indication

In such a small scale investigation it would be over-presumptive to form any conclusions, but the results do give an indication of the reliability of the assumption that overlap can be indicative of semantic relatedness.

In the sample of word pairs examined, there seem to be 3 main groupings: those with zero, 1 or 2 words in common, those with 7 - 11 words, and those somewhere in between. There are 3 examples which fall into the top group: namely *market - industry; entertainment - games;* and *games - market.* The first two of these pairs are strongly related semantically, and there are degrees of similarity in the third pair.

Occurring in the middle range group (those with between 3 and 6 overlapping words), there are 5 word pairs: *industry - sales; games - sales; toy - sales; market - sales;* and *worth - million.* Whilst we might expect to see more overap between, for example, *market* and *sales* (with an overlap of 3) than between *games* and *market* (overlap of 8), it is clear that the words in this mid-range are at least partially related.

The remainder of the word pairs (37 of them) exhibit low numbers of common words (see the tables above for all the examples). In general, these pairings involve words which are, indeed, not related, or are only very slightly related to one another.

In some of the examples, word pairs which are related (as judges by human experience) can end up with a low overlap, or even no overlap at all. This is unavoidable to some extent, but the problem is exaccerbated by the cultural difference between American and British usage of English (as discussed in section 6.2). Of particular importance in terms of out example abstract is the failure of WordNet to use the word *money* in its definition of *pounds_sterling* . This clearly affects all pairings in which *pounds_sterling* is involved, and we see, for example, no overlap between *pounds_sterling* and *sales*, which most people would judge as being fairly highly related. It would be expected that this problem could be overcome by using a more standard dictionary.

With this small sample size there would be little point in applying any test for statistical significance to the results, since a much larger sample would be required for meaningful results. Such an extended study is beyond the scope of the current work, but it is possible to get some indication of tendencies.
This indication is that highly related words do have a larger number of words in common to their definitions, and unrelated words do have lower numbers, with partially related words having rather more variable results.

It comes as no surprise that there are pairs of words that are unrelated, yet contain common words in their definitions. What is indicated by this small investigation is that when words are related (such as *market* and *industry*), the number of common words is much greater. Whereas unrelated words may have an overlap of 1 or 2, or even a few more words, the indication is that they do not have more than about 5 words in common.

The initial indication, then, is that the assumtion of semantic relatedness being proportional to overlap of definition words is justified, at least to some extent. Athough this has been taken as a specific assumption in this work (in the manner of many authors working on word sense disambiguation, as discused in section 3.4.1 above), it is still interesting to note this indication, which could perhaps serve as a pilot study for a more detailed analysis (which is, however, beyond the original scope of this work).

4 The COMMIX system: Method

The methodology is based on the assumptions and general approach described above. For the sake of clarity, the description of the COMMIX methodology for processing an abstract will be based on the following example abstract[48]:

[48]The abstracts used in this work are written by human professional abstractors. The original full text to which this abstract refers appeared in The Observer of Sunday 11th October 1992.

The video games industry is growing fast and will dominate the toy market and become an established part of home entertainment. The 1991 computer games market was worth 275 million pounds sterling growing to 500 million in 1992, half the toy market. Hardware sales will rise from 261 to 635 million pounds sterling in 1994. Associated software sales are forecast at 645 million pounds sterling in 1993. The compact disc market is worth 345 million pounds sterling. The main competitors in the market are Sega and Nintendo. Nintendo will spend 15 million pounds sterling on advertising over Oct-Dec 1992.

The processing of an abstract falls into the following stages:-

1) Sorting and labelling of the original abstract

This involves the deletion of all closed class words and the labelling of all open class words according to their syntactic word class (as listed in the Wordnet database). If an item is not listed in the database, it is retained and labelled '*n_l*'. Full labelling of the abstract involves numbering each item according to both sentence number and word number, indicating its syntactic class, and whether it is a name or number, and leaving a trace of items that have been deleted. Labelling results in a list of constituents of the form

[sentence number word number item syntactic class of item]
where 'item' is the item itself, or DEL if it has been deleted as a closed class word.

For the example abstract, the first 2 sentences thus labelled are as follows:

1 1 DEL det 1 2 video noun 1 3 games noun 1 4 industry noun 1 5 DEL aux 1 6 growing adj 1 7 fast noun 1 8 DEL conj 1 9 DEL modal 1 10 dominate verb 1 11 DEL det 1 12 toy noun 1 13 market noun 1 14 DEL conj 1 15 become verb 1 16 DEL det 1 17 established adj 1 18 part noun 1 19 DEL prep 1 20 home noun 1 21 entertainment noun 1 21 punc_s 2 22 DEL det 2 23 1991 number 2 24 computer noun 2 25 games noun 2 26 market noun 2 27 DEL aux 2 28 worth noun 2 29 275 number 2 30 million noun 2 31 pounds noun 2 32 sterling adj 2 33 growing adj 2 34 DEL prep 2 35 500 number 2 36 million noun 2 37 DEL prep 2 38 1992 number 2 38 DEL punc 2 39 half noun 2 40 DEL det 2 41 toy noun 2 42 market noun 2 42 . punc_s ... etc.
This detailed labelling of the abstract is not utilised to the full at this stage of the implementation, but will be required at a later stage when constraints (involving sentence number, the position and type of punctuation, the type and position of deleted items) are introduced. In the current implementation it is only the syntactic word class information which is utilised.

2) Initial processing of labelled abstract

This stage involves two processes: firstly, the identification of compound nominals which already exist in the abstract; secondly, the recording of the frequencies of all nouns which occur more than once (and are thus deemed to be 'salient' to some degree). The first stage yields the following in our example:

[video games industry] [growing fast] [toy market] [established part] [home entertainment] [1991 computer games market] [worth 275 million pounds sterling growing] [500 million] [toy market] [hardware sales] [635 million pounds sterling] [associated software sales] [645 million pounds sterling] [worth 345 million pounds sterling] [main competitors] [spend 15 million pounds sterling] [Dec 1992]

The result of the second stage is a list of the salient nouns (ie those which occur more than once), alongside their frequency of occurrence in the abstract. The result of this second stage in our example is the following list:

[games 2 toy 2 market 5 worth 2 million 6 pounds_sterling 5 sales 2]

2b) Conversion of verbs to nouns

Note that this stage does not appear in the current implementation, but is to be included at a later stage. It involves the conversion of verbs to nominal form. A simplistic, but nevertheless useful approach is taken to this problem: namely, to present verbs in the gerundive forms. In our abstract this would yield the changes *dominate -> dominating* (rather than the preferred *domination*) and *become -> becoming* (which could be eliminated as a non-crucial verb). Where cases of ambiguity of word class occur, (which include 'noun' as one of the possible interpretations), the default assumption is that the item is a noun. For this reason changes are not made to the ambiguous *rise, forecast* and *spend* which are all assumed to be (and labelled as) nouns.

3) Lookup salient nouns and Filter glosses

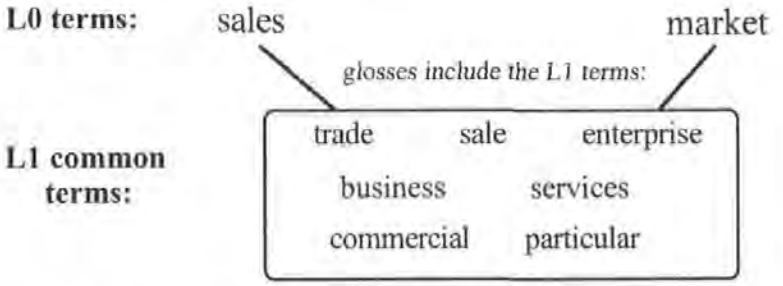
This is the point at which the defining gloss for each salient noun is looked up. The processing of each gloss involves the filtering of words which do not pertain to the definition (such as *senses* in *there are four senses of ...*), and the deletion of all closed class words, followed by the labelling of all open class words, as in stage 1. This stage takes as input, the list of salient nouns shown above and produces, for each one, a gloss from which all the closed class words have been removed. For the purposes of this example, the glosses for 'sales' and 'market' should suffice to show how relatedness (linkage) is established.

GLOSS FOR sales:
trade sales desperate boost sales commercial enterprise business enterprise business purchase sale goods services sale sale cut sale sales event selling specially reduced prices sale reduce inventory selling merchandising marketing exchange goods agreed sum money sale particular instance selling made three sales hour selling merchandising marketing exchange goods agreed sum money
GLOSS FOR market:
market customers particular product service class social class people sharing common attribute market securities industry securities markets aggregate market always rustates small investor industry people engaged particular kind commercial enterprise grocery store grocery market supermarket marketplace mart market marketplace commerical activity whereby good services exchanged competition there market activity behavior specific action pursuit avoided recreational activity senses market market deal market ransact deal business market sale produce sale trade deal merchandise engage trade commercialize market make commercial change alter cause change make different

The glosses are then compared against one another in the search for matching items, which will justify the formation of a link between their respective parents.

4) Matching the Terms

This stage involves searching for terms which are common to more than one gloss. As mentioned above, a link between two parent items is held to be justified if their respective glosses contain common terms. In the example, there are seven words which are common to the glosses of the parent (L0[49]) terms *sales* and *market*, and these are referred to as L1 terms, since they are from the first (1) lookup (L) process. This commonality is shown below:



The L1 common terms are held to justify the assumption of semantic linkage between the two L0 terms *sales* and *market*, and as such are the justification for applying all the modifiers of both L0 terms to the most salient of the two parent nouns.

5) Weighting the links

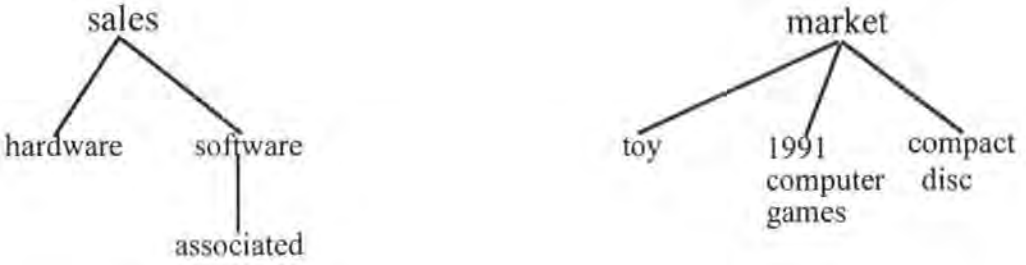
Once all the commonalities between the glosses of the salient nouns have been identified, the frequency of each matched term (i.e. link) throughout the glosses is recorded. This frequency is to be used subsequently in the establishment of a salience weighting value which reflects the strength of the justification for the linkage. This information is not utilised at this stage, but is used to represent relative salience at a later stage of the implementation.

6) Selection of head noun and cross-application of modifiers

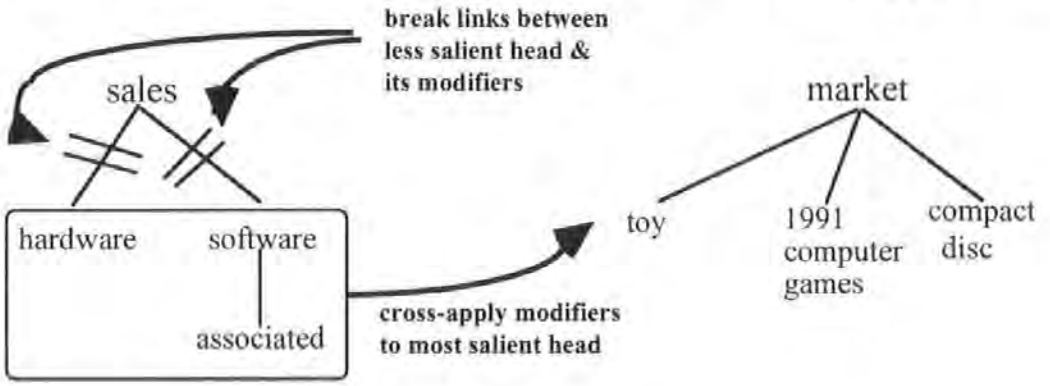
If linkage has been found between two or more items which occur as salient nouns in the abstract, then there is justification for combining their modifiers and applying them all to the most salient of related nouns (the one occuring with the highest frequency in the original abstract). This stage identifies the appropriate noun and applies to it all the modifiers which relate either to that noun or to semantically related nouns. In the example, all the modifiers applying to each of 'sales' and 'market' in the original abstract can be represented by the following tree structures:

---

[49]The term L0 refers to words from the original text; L1 refers to definition words resulting from 1 Lookup process (and L2 to words in the definitions of L1 words).



The most frequent (and thus the most salient) of *sales* and *market* is found to be *market*, which becomes the head of the compound nominal being formed. The 'links' between the less salient noun (ie *sales* in the example) and its modifiers are broken, and the modifiers are cross-applied to join those which already apply to *market*.



5 The generated expressions

The expressions generated by the system for any given abstract are aimed at being as informative as possible, whilst retaining the nominal character essential to these aboutness expressions. The system is flexible, and can produce a number of sub-expressions of the longest one, if required.

However, given that the quest for informativeness is primary, the most informative expression is always given first: in the case of our example abstract, this is:

toy 1991 computer games hardware associated software compact disc market.

The above expression is the longest of those produced by the system at this stage. Although it may be deemed by some as unacceptably long, concise, or even not well-formed, it does provide a surprising degree of insight into the aboutness of the example abstract. There are clearly possibilities for tuning the technique to match any future requirements.

Some of the other expressions which are generated by the system, then, include subsets of the longest one, such as:

toy 1991 computer_game s compact_disc market.

Linkage can also be pursued across lookup levels, so that, for example, we get the expression *video_games market* which is formed by the identification of an overlap

between the hypernym *market*, which occurs as a salient noun at the top level (that is, in the original abstract) and the noun *industry*, which appears in the gloss for *market*. This fairly crude technique is one method of generating novel output expressions, and the method is extendable to the indentification of overlap both across and within levels of lookup.

As would be expected with this methodology, some of the output expressions do not contribute at all (or barely) to an indication of the aboutness of the abstract. For the VGI abstract, such examples are:

500 million,
hardware 635 million associated software 645 worth spend 15 pounds_sterling,
hardware associated software pounds_sterling,

which inform us that the abstract is something to do with large numbers, money and probably computers.

It should be noted that the method as described above only pursues possible linkage through one level of lookup. In fact, the lookup process is in theory iterative, and investigations are currently under way into the effect on the linkage of extending the lookup to include words appearing at the L2 lookup level: in other words, to include words appearing in the definitions of defining words of parent terms. For the example abstract, recent work at this extended level yields additional aboutness terms, including the following:

home toy, established toy, video games market, video games sales,
compact disc sales, toy sales, 1991 computer games sales,

all of which are novel yet fairly representative of the aboutness of the abstract.

6 Discussion

There has as yet been no formal evaluation of the output, although this is planned for the near future. However, informal comments indicate that people do often get a fairly good idea of what the abstract is about from the expressions generated by this system.

There are clearly shortcomings in the COMMIX system: some of these are consequences of the methodology itself, whilst others could be easily rectified without changing the basic foundations of the system.

6.1 Some shortcomings

The pitfalls associated with the methodology generally arise from the simplicity of the approach. Throughout the development of this work every effort has been made to avoid the unnecessary and unprincipled introduction of additional knowledge, in order to get a clear picture of the performance of the system at its most uncluttered. The main such pitfall is the generation of, at worst, unrelated rubbish, and, at best, only mildly informative expressions. The problems are largely caused by the matching of very general words, and the results can be strongly skewed by the occurrence of conceptually high-level nouns, whose glosses contain very general

words themselves. Problems such as this are to be expected, but can be alleviated both by the incorporation of formal constraints to the system, and by its its integration with other approaches within a hybrid IE system.

The main shortcomings of the implementation at present, again stem from its simplicity in its basic form. As the results of the brief overlap analysis in section 3.4.2 showed, a notion of salience which is based solely on frequency of occurrence in the text, leads to a decrease in the amount (and quality) of linkage being identified. It is clear, then, that a polished version of COMMIX would use an extended notion of salience. The notion would be extended to represent the following two observations: firstly, that the initial sentence of a text, (particularly an abstract) is more likely than any other individual sentence to contain strongly representative indicators of content. Secondly, that the fact that a noun, as it occurs in the text, is already modified by something, is an indication that it is more salience than a single-occurrence noun which is not modified at all in the text.

The incorporation of constraints on the output is likely to improve the readability of the aboutness expressions generated, and eliminate some of the less informative ones. Again, such extensions are simple additions to the current 'basic bones' system.

6.2 Some Consequences of using WordNet

As has been mentioned, it was originally planned to utilise a standard machine-readable dictionary, such as LDOCE for this work. This choice (although uneconomical and impracticable at the time of development) would undoubtedly have led to the generation of some different aboutness expressions. At this 'uncluttered' stage of the work, it is likely that more linkage would be identified, given that the definitions tend to be more wordy than we see in WordNet. The use of WordNet can complicate the picture, in that definitions include multiple levels of hypernymy, and thus the grain size of the entries is rather variable. This appears to be a disadvantage when the system operates purely to level 1 lookup, but can be used to advantage when the lookup is extended to a deeper level.

There are other, more simple, but just as damaging, consequences of using WordNet. Being an academic tool, it has not been subjected to the rigours of accuracy checking that would be expected of a more commercial product. Thus, we find errors of spelling (e.g *commerical* for *commercial*), which are not problematic for human readers, but are a major problem for systems which require direct matching. Another element of this word-matching problem stems from the fact that, being an American tool, it fails to give the alternate (British) English spellings in cases where the two differ. In this regard, the other typical problem of different word meanings (e.g. *pants* versus *trousers*) also occurs.

6.3 The cost of compaction

Just one look at the longest expression generated for the VGI abstract (in section 5 above), in sufficient to exemplify this point. Readability is poor (relative to shorter expressions) and ambiguity of the relations between components of the expression is rife. However, the information content is very high, even though it is more difficult for the reader to assimilate at a cursory glance. There is clearly a play-off between

readability and informativeness, and this undoubtedly has bearing on the length of the compound nominal expressions that are generated by COMMIX.

7 Future Directions

There are two different ways in which the work might be used to advantage, and these can be categorised as 'academic' (or 'theoretical') and 'application-based' (or 'practical').

7.1 Academic use

The simplicity of COMMIX in its basic form means that it can lend itself very usefully as a tool for investigating a variety of linguistic phenomena.

The COMMIX system has been tested (by the author) on a number of abstracts with very different subject matter. In general, the output expressions generated by the system are much more representative of the aboutness of the text when the input abstract is informative (rather than indicative), and when it is non-biographical in nature. Biographical abstracts tend to simply give one fact after another, and there is often no relation *per se* between different parts of the text, other than that they relate to the same individual. Some recent investigations into the effects of name replacement and pronoun resolution have indicated that the latter is more likely to improve the quality of the output expressions.

Given that the system in its basic form does not rely on any constraints on the construction of the compound nominal aboutness expressions, it follows that there is scope for using it as a tool to test the effects of different (sets of) constraints on the performance. This could be of particularly use in the study of compound nominals as a linguistic phenomenon.

7.2 Practical Applications

There are numerous ways in which the approach to generating highly informative aboutness expressions could be incorporated into other information technologies. It would seem particularly useful as a module in a hybrid Information Extraction system, which could utilise this approach to filter the text either before, or in addition to other approaches which are more domain-dependent. Such a hybrid system could include template and statistical methodologies.

A further possibility would be the integration of this approach with a user query processing system, such as that developed by Sparck Jones and Tait in 1984. User queries could thus be translated into their most compact form, expressed as compound nominals, and used for either complete or partial matching against aboutness expressions of essence, which would, of course, be generated for each text occurring in the data set.

It must be remembered that this approach is primarily intended for use when the input text is an abstract, which is short, concise and less plagued by ambiguity than its associated full text. It would, of course, be possible to use the original full text (for which an abstract was written) in checking the output expressions, and such a combined method could be extremely useful for resolving conflict and assigning

weightings both to the expression generated and to the components of these expressions during the processing stages.

Bibliography

Bauer, L. (1979) On the need for pragmatics in the study of nominal compounding. *Journal of Pragmatics* **3,** pp 45-50

Grishman, R. (1996) The NYU System for MUC-6 or Where's the Syntax? *Proceedings of the Sixth Message Understanding Conference,* November, 1996, Morgan Kauffman.

Hodges, J., Yie, S., Reighart, R. & Boggess, L. (1996) An automated system that assists in the generation of document indexes. *Natuaral Language Engineering* 2 (2) , pp. 137-160.

Hutchins, W.J. (1977) On the Problem of 'Aboutness' in Document Analysis. *Journal of Informatics,* 1 (1), pp. 17-35.

Jacobs, P.S. & Rau, L.F. (1990) SCISOR: Extracting information from on-line news. *Communications of the ACM,* 33, (11), pp. 88-97.

Lesk, M. (1986) Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proc. 1986 SIGDOC Conference*, Toronto, Ontario, June 1986.

Levi, J. (1978) *The syntax and semantics of complex nominals.* Academic Press, NY

Masterman, M., Needham, R., Sparck Jones, K. & Mayoh, B. (1957) Agricola Terram Dimovit Aratro, ML92, Cambridge Language Research Unit, Cambridge, England. Reprinted by the Computer Laboratory, Cambridge University, April 1986.

Paice, C.D. (1990) Constructing literature abstracts by computer. Techniques and prospects. *Information Processing and Management,* 26 (1), pp. 171-186.

Rigau, G., Atserias, J. & Agirre, E. (1997) Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. *Proc. ACL 1997.*

Rino, L.H.M. (1996) A Discourse Model for Gist Preservation. Paper based on author's PhD Thesis: *Modelagem do Discurso para o Tratamento da Concisao e Preservacao da Ideia Central na Geracao de Textos.* Universidade de Sao Paulo, Brasil. April 1996.

Scott, D. & Hovy, E. (eds.) (April 1993). Burning Issues in Discourse. NATO Advanced Research Workshop, Maratea, Italy.

Sparck Jones, K. & Tait, J. (1984) Automatic Search Term Variant Generation. *Journal of Documentation,* 40, pp.50-66.

Bibliography

Agirre, E. and Rigau, G. (1995) A Proposal for Word Sense Disambiguation using Conceptual Distance. In *International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

Agirre, E. and Rigau, G. (1996) Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th International Conference on Computational Linguistics - COLING 96.* vol. 1, pp. 16-22.

Amsler, R. A. (1980) The Structure of the Merriam-Webster Pocket Dictionary. *Technical Report*, TR-164. University of Texas at Austin.

Aone, C., Okurowski, M. E., Gorlinsky, J. and Larsen, B. (1997) A Scalable Summarization System using Robust NLP. In Mani, I. and Maybury, M., Eds. (1997), pp. 66-73.

Arens, Y., Granacki, J. J. and Parker, A. C. (1987) Phrasal Analysis of Long noun Sequences. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, July 1987, pp. 59-64. Stanford University, Stanford, California.

Aretoulaki, M. (1996) *COSY-MATS: A Hybrid Connectionist-Symbolic Approach to the Pragmatic Analysis of Texts for their Automatic Summarisation.* PhD Thesis, Centre for Computational Linguistics, University of Manchester Institute of Science and Technology (UMIST), January, 1996.

Barzilay, R. and Elhadad, M. (1997) Using lexical chains for text summarization. In Mani, I. and Maybury, M., Eds., (1997), pp. 10-17.

Bauer, L. (1979) On the need for pragmatics in the study of nominal compounding. *Journal of Pragmatics*, 3, pp. 45-50.

Baxendale, P.B. (1958) Machine-Made Index for Technical Literature - An Experiment. *IBM Journal of Research and Development*, 2 (4), pp. 354-361.

Beardon, C. & Turner, K. (1991) An Analysis of the problems involved in understanding complex nominals. *Internal report*, Rediffusion Simulation Research Centre, University of Brighton.

Bikel, D., Miller, S., Schwartz, R., and Weischedel, R. (1997) Nymble: a High-Performance Learning Name-Finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics*, pp. 194-201.

Binot, J. L., and Jensen, K. (1987) A Semantic Expert Using an Online Standard Dictionary. *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 709-714. Milan.

Boguraev, B. and Kennedy, C. (1997a) Salience-based Content Characterisation of Text Documents. In Mani, I. and Maybury, M. Eds., (1997), pp. 2-9.

Boguraev, B. and Kennedy, C. (1997b) Technical Terminology for Domain Specification and Content Characterisation. In Pazienza, M. T., Ed. (1997), pp. 73-96.

Brachman, R. J. (1979) On the Epistemological Status of Semantic Networks. In Findler, N. V. (Ed.) *Associative Networks - Representation and Use of Knowledge by Computers.* Academic Press. New York.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G. and Anderson, A. H. (1997) The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23 (1), pp. 13-32.

Cavazza, M. (1998) Textual Semantics and Corpus-Specific Lexicons. In *Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, at 1st International Conference on Language Resources and Evaluation (LREC).* May 26th 1998, Granada, Spain.

Collins English Dictionary (3rd Edition, 1991). HarperCollins, England.

Comrie, B. and Thompson, S. A. (1985) Lexical nominalization. In Shopen, T. (Ed.) *Language Typology and syntactic description, vol. III: grammatical categories and the lexicon*, pp. 349-398. Cambridge University Press, Cambridge, England.

Coolican, H. (2nd Edn., 1994) *Research Methods and Statistics in Psychology.* Hodder and Stoughton. London.

Costello, F. and Keane, M. T. (1992) A Model-Based Theory of Conceptual Combination. In *Artificial Intelligence and Cognitive Science (AICS '92).*

Cowie, J. (1983) Automatic Analysis of Descriptive Texts. In *Proceedings of the Conference on Applied Natural Language Processing.*

Cumming, S. (1991) Nominalization in English and the organization of grammars. In *Workshop on Decision making throughout the generation process, International Joint Conference on Artificial Intelligence (IJCAI '91),* pp. 42-51.

DeJong, G. (1982) An Overview of the FRUMP System. In Lehnert, W. R. and Ringle, M. H. (Eds.) *Strategies for Natural Language Processing.* pp. 149-176. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Dixon Keller, J. and Lehman, F. K. (1991) Complex Concepts. *Cognitive Science,* 15 (2), pp. 271-291.

Downing, P. (1977) On the creation and use of English compound nouns. In *Language,* 53 (4), pp. 810-842.

Edmundson, H. P. (1969) New Methods in Automatic Extracting. In *Journal of the Association for Computing Machinery,* 16 (2), 264-285.

Endres-Niggemeyer, B. (1990a) A Procedural Model of an Abstractor at work. *International Forum of Information and Documentation (IFID)* 15, (4).

Endres-Niggemeyer, B., Waumans, W. and Yamashita, H. (1990b) A Cognitive Modelling Approach to Summary Writing. (Draft paper sent by author.)

Endres-Niggemeyer, B., Hobbs, J. and Sparck Jones, K., Eds., (1993) Summarizing Text for Intelligent Communication. *Dagstuhl-Seminar-Report;* 79, Dagstuhl, Germany.

Endres-Niggemeyer, B., Maier, E. and Sigel, A. (1995) How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert

Abstractor. *Information Processing and Management : Special Issue on Summarising Text.*

Enkvist, N. E. (1973) *Linguistic Stylistics.* Mouton & Co., The Hague, The Netherlands.

Evans, D. A. and Zhai, C. (1996) Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.* June 1996, pp. 17-24. University of California, Santa Cruz, Ca, USA.

Evans, D. A., Ginther-Webster, K., Hart, M., Lefferts, R. G., and Monarch, I. A. (1991) Automatic Indexing using selective NLP and first-order thesauri. In A. Lichnerowicz (Ed.) *Proceedings of the, RIAO '91 Conference: Intelligent text and Image handling,* pp. 624-644. Amsterdam, NL., Elsevier.

Fairthorne, R. A. (1969) Content analysis, specification and control. *Annual Review of Information Science and Technology,* 4, pp. 73-109

Farrow, J. F. (1991) A Cognitive Process Model of Document Indexing. *Journal of Documentation,* 47 (2), pp. 149-166.

Finin, T. (1980) *The Semantic Interpretation of Compound Nominals.* Coordinated Science Laboratory, University of Illinois, USA.

Fum, D., Guida, G. and Tasso, C. (1985) Evaluating Importance: A Step towards Text Summarization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-'85),* pp. 840-844.

Gay, L. S. and Croft, W. B. (1990) Interpreting Nominal Compounds for Information Retrieval. *Information Processing and Management,* 26 (1), pp. 21-38.

Gazdar, G., Klein, E., Pullum, G. K. and Sag, I. A. (1985) *Generalized Phrase Structure Grammar.* Basil Blackwell Publisher Ltd. Oxford, UK.

Gerrig, R. J. and Murphy, G. L. (1992) Contextual Influences on the Comprehension of Complex Concepts. *Language and Cognitive Processes,* 7 (3/4), pp. 205-230.

Gershman, A. (1977) Conceptual analysis of noun groups in English. In *Proceedings of the 5thInternational Joint Conference on Artificial Intelligence (IJCAI '77)*, pp. 132-138. Cambridge, MA.

Ghadessy, M., Ed. (1988) *Registers of Written English: Situational Factors and Linguistic Features.* Pinter.

Ghadessy, M., Ed. (1993) *Register Analysis: Theory and Practice.*. Pinter.

Ginsberg, A. (1993) A Unified Approach to Automatic Indexing and Information Retrieval. *IEEE Expert*, 8 (2), pp. 46-56.

Gladwin, P., Pulman, S., and Sparck Jones, K. (1991) Shallow Processing and Automatic Summarising: A First Study. *Technical Report No. 223,* University of Cambridge Computer Laboratory, Cambridge, England.

Grishman, R. (1996) The NYU System for MUC-6 or Where's the Syntax? *Proceedings of the Sixth Message Understanding Conference,* November, 1996, Morgan Kauffman.

Hahn, U. and Mani, I. (1998) Automatic text Summarization: Tutorial T6. *13th European Conference on Artificial Intelligence (ECAI-98).* Brighton, UK.

Hahn, U. and Reimer, U. (1998) Text Summarization Based on Terminological Logics. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-98),* pp. 165-169. Brighton, UK.

Halliday, M. A. K. (1988) On the language of physical science. In Ghadessy, M., Ed. (1988), pp. 162-178.

Hand, T. F. (1997) A Proposal for task-Based Evaluation of Text Summarization. In Mani, I. and Maybury, M., Eds. (1997), pp. 31-38. Madrid, Spain.

Hodges, J., Yie, S., Reighart, R. and Boggess, L. (1996) An automated system that assists in the generation of document indexes. *Natuaral Language Engineering* 2 (2) , pp. 137-160.

Hovy, E. and Lin, C. Y. (1997) Automated Text Summarization in SUMMARIST. In Mani, I. and Maybury, M., Eds. (1997), pp. 18-24.

Hutchins, W. J. (1977a) On the problem of 'aboutness' in document analysis. *Journal of Informatics*, 1(1), pp. 17-35.

Hutchins, W. J. (1977b) On the structure of scientific texts. *University of East Anglia Papers in Linguistics* 5, September 1977, pp. 18-39.

Hutchins, W. J. (1981) Information Retrieval and Text Analysis. In van Dijk, T. A., Ed. (1985) *Discourse and Communication*, pp. 106-125. De Gruyter. Berlin, Germany.

Hutchins, W. J. (1993). Introduction to "Text Summarization" workshop. In Endres-Niggemeyer, B. et al, Eds., (1993). pp. 14-19.

Isabelle, P. (1984) Anothe Look at Nominal Compounds. In*Proceedings of the International Conference on Computational Linguistics (COLING-84).* pp. 509-516.

Jacobs, P. S. and Rau, L. F. (1990) SCISOR: Extracting information from on-line news. *Communications of the ACM*, 33, (11), pp. 88-97.

Jacquemin, C., Klavans, J. L. and Tzoukermann, E. (1997) Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 24-31. Madrid, Spain.

Jensen, K. (1991) A Broad-Coverage Natural Language Analysis System. In Tomita, M. (Ed.) *Current Issues in Parsing Technology*. Kluwer Academic.

Jing, H., Barzilay, R., McKeown, K. and Elhadad, M. (1998) Summarization Evaluation Methods: Experiments and Analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*. Technical Report, AAAI, 1998, pp. 60-68. (Reference from Hahn, U. and Mani, I., 1998).

Jordan, M. P. (1991) The Linguistic Genre of Abstracts. In Della Volpe (Ed.) *The Seventeenth LACUS Forum*, pp. 507-527. Linguistics Association of Canada and the United States.

Jordan, M. P. (1993) Openings in Very Formal Technical Texts. *Technostyle*, 11 (1), pp. 1-28.

Jordan, M. P. (1994) Towards Plain Language: A Guide to Paraphrasing Complex Noun Phrases. *Journal of Technical Writing and Communication*, 24 (1), 77-96.

Justeson, J. S. and Katz, S. M. (1995) Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1 (1), pp. 9-27.

Kameyama, M. and Arima, I. (1994) Coping with Aboutness Complexity in Information Extraction from Spoken Dialogues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-94)*, pp. 1-4.

Kennedy, C. and Boguraev, B. (1996) Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*. pp. 113-118. Copenhagen, Denmark.

Kieras, (1980) ***

Klare, G. R. (1976) Judging readability. *Instructional Science*, 5 (1), pp. 55-61.

Krovetz, R. (1997) Homonymy and Polysemy in Information Retrieval. In *Proceedings of the Association for Computational linguistics (ACL-97)*,, pp. 72-79.

Kupiec, J., Pedersen, J. and Chen, F. (1995) A Trainable Document Summarizer. In *Proceedings of the ACM-SIGIR '95*. Seattle, WA.

Lauer, M. (1995) Corpus Statistics Meet with the Noun Compound: Some Empirical Results. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*,, pp. 47-54. MIT. Cambridge, MA, USA.

Lees, R. B. (1970) Problems in the grammatical analysis of English nominal compounds. In Bierwisch, M. and Heidolph, K. E. (Eds.) *Progress in Linguistics*, pp. 174-186. The Hague, NL. Mouton.

Lehnert, W. G. (1988). The Analysis of Nominal Compounds. In Eco, U., Santambrogio, M. and Violi, P., Eds. (1988) *Meaning and Mental Representations*, pp. 155-179. Indiana University Press. Bloomington, USA.

Leonard, R. (1984) *The Interpretation of English Noun Sequences on the Computer*. North Holland. Amsterdam.

Lesk, M. (1986) Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 1986 SIGDOC Conference,* June 1986.Toronto, Ontario.

Levett, W. J. M. (1989) Lexical Entries and Accessing Lemmas. In *Speaking: From Intention to Articulation,* Chapter 6, p. 181. MIT. London.

Levi, J. (1978) *The syntax and semantics of complex nominals.* Academic Press, New York.

Lewis, D. D. (1992) An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Copenhagen, Denmark.

Liddy, E. D. (1991) The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing and Management,* 27 (1), pp. 55-81.

Liddy, E. D., McVearry, K. A., Paik, W., Yu, E. and McKenna, M. (1993) Development, Implementation and Testing of a Discourse Model for Newspaper Texts. In *Proceedings of the ARPA Workshop on Human Language Technology,* Princeton, NJ, USA.

Lin, C. Y. (1995) Knowledge-based Automatic Topic Identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* pp. 308-310. Boston, MA.

Luhn, H. P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development,* 2 (2), pp. 159-165.

Luk, A. K. (1994) Lexical Matching - A Heuristic for Semantic Connection Search and Sense Disambiguation using On-line Dictionaries. In *Proceedings of the 1994 Conference on Computational Linguistics (COLING-'94).*

Luk, A. K. (1995) Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* MIT, pp. 181-188.

Lyons, J. (1968) *Introduction to Theoretical Linguistics*. Cambridge University Press. Cambridge, England.

Lyons, J. (1977) *Semantics. Vols I and II*. Cambridge University Press. Cambridge, England.

Maizell, R. E., Smith, J. F. and Singer, T. E. R. (1971) *Abstracting Scientific and Technical Literature: An Introductory Guide and Text for Scientists, Abstractors, and Management*. Wiley-Interscience. John Wiley. New York.

Mani, I. and Bloedorn, E. (1998) Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1, 1.

Mani, I., and Maybury, M., Eds. (1997) *Proceedings of the Association for Computational Linguistics (ACL/EACL) '97 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.

Marcus, M. P. (1980) *A Theory of Syntactic Recognition for Natural Language*. MIT Press. Cambridge, MA, USA.

Marsh, E. (1984) A Computational Analysis of Complex Noun Phrases in Navy Messages. In *Proceedings of the International Conference on Computational Linguistics (COLING-84)*. pp. 505-508.

Masterman, M., Needham, R., Sparck Jones, K. and Mayoh, B. (1957) Agricola Terram Dimovit Aratro, *ML92*, Cambridge Language Research Unit, Cambridge, England. Reprinted by the Computer Laboratory, Cambridge University, April 1986.

McDonald, D. B. (1982) *Understanding Noun Compounds*. PhD Thesis, Carnegie-Mellon University, Pittsberg, PA, USA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1990) Five Papers on WordNet. *CSL Report 43*, Cognitive Science Laboratory, Princeton University. Princeton, New Jersey.

Morris, J. and Hirst, G. (1991) Lexical cohesion computed by thesaural relations as an indicator of the structure of the text. *Computational Linguistics*, 17 (1), pp. 21-45.

Murphy, G. L. (1990) Noun Phrase Interpretation and Conceptual Combination. *Journal of Memory and Language*, 29, pp. 259-288.

Niwa, Y., Nishioka, S., Iwayama, M. and Takano, A. (1997) Topic Graph Generation for Query Navigation: Use of Frequency Classes for Topic Extraction. In *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97)*, pp. 95-100. Phuket, Thailand.

Norris, J. (1996) Compound Nominal Generation for Information Retrieval : The COMMIX System, In *Proceedings of the Workshop on Language Engineering for Document Analysis and Recognition* (*LEDAR*), AISB 1996 Workshop Series, 2nd April 1996, pp. 48-55. Sussex University, Brighton, UK.

Norris, J. (1997) Extracting the Essence from Text: a Computational Approach. In *Proceedings of the International Workshop on Lexically Driven Information Extraction*, July 16th 1997, pp. 63-80. Frascati, Italy.

Paice, C. D. (1981) The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In Oddy, R. N., Robertson, S. E., van Rijsbergen, C. J. and Williams, P. W. (Eds.) *Information Retrieval Research*, pp. 172-191. Butterworths.

Paice, C. D. (1990) Constructing literature abstracts by computer. Techniques and prospects. *Information Processing and Management*, 26 (1), pp. 171-186.

Park, H.Y., Han, Y.S. and Lee, K.H. (1996)A Probabilistic Approach to Compound Noun Indexing in Korean Texts. In *Proceedings of the International Conference on Computational Linguistics* (*COLING-96*), Vol 1, pp. 514-518. Copenhagen, Denmark

Pazienza, M. T., Ed. (1997) *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. International Summer School, SCIE-97, Frascati, Italy. Springer.

Penman (1988) *The Penman Primer, User Guide, and Reference Manual*. 1988. Unpublished documentation, USC Information Sciences Institute.

Procter, P. (Ed.) (1978) *Longman Dictionary Of Contemporary English.* Harlow, Essex, England: Longman Group.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J.A. (1985) *A Comprehensive Grammar of the English Language.* London: Longman.

Rau, L. F., Jacobs P. S. and Zernik, U. (1989) Information Extraction and Text Summarization using Linguistic Knowledge Acquisition. *Information Processing and Management,* 25 (4), pp. 419-428.

Rau, L. F., Jacobs, P. S., & Zernik, U. (1989) Information Extraction and Text Summarization using Linguistic Knowledge Acquisition. *Information Processing and Management,* 25 (4), pp. 419-428.

Reimer, U. and Hahn, U. (1988) Text Condensation as Knowledge-based abstraction. In *Proceedings of the 4th Conference on Artificial Intelligence Applications - CAIA '88,* pp. 338-344. Washington DC.

Resnik, P. (1993) *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. Thesis. University of Pennsylvania, December.

Resnik, P. (1995) Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of the Third Workshop on Very Large Corpora,* MIT.

Resnik, P. and Hearst, M. (1993) Structural Ambiguity and Conceptual Relations. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives,* June 1993, pp. 58-64. Ohio State University.

Resnik, P. and Yarowsky, D. (1997) A Perspective on Word Sense Disambiguation Methods and Their Evaluation. In *Proceedings of the Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Association for Computational Linguistics, Special Interest Group on the Lexicon,* pp. 79-86.

Richardson, R., Smeaton, A.F., and Murphy, J. (1994) Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words.. In *Working Paper CA-1294,* School of Computer Applications, Dublin City University, Dublin, Ireland.

Rigau, G., Atserias, J. and Agirre, E. (1997) Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. In *Proceedings of the Association for Computational Linguistics*, pp. 48-55.

Rino, L. H. M. (1996) A Discourse Model for Gist Preservation. Paper based on author's PhD Thesis: *Modelagem do Discurso para o Tratamento da Concisao e Preservacao da Ideia Central na Geracao de Textos.* Universidade de Sao Paulo, Brasil. April 1996.

Sager, J. C. (1990) *A Practical Course in Terminology Processing.* John Betjamins Publishing Company.

Sager, N. (1981) *Natural Language Information Processing.* Addison-Wesley, Reading, Massachusettes.

Salton, G. (1991) Developments in Automatic Text Retrieval. *Science,* 253, pp 974-980.

Salton, G., Allan, J., Buckley, C. and Singhal, A. (1994) Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. *Science*, 264, 1421-1426.

Salton, G., Singhal, A., Mitra, M. and Buckley, C (1997) Automatic Text Structuring and Summarization. *Information Processing and Management*, 33 (2), pp. 193-208.

Schank, R. C. (1973) Identification of conceptualisations underlying natural language. In Schank, R. C. and Colby, K. M. (Eds.) *Computer Models of Thought and Language*, pp. 187-247. Freeman, San Francisco.

Schank, R. C. and Abelson, R. P. (1977) *Scripts, Plans Goals and Understanding.* Lawrence Erlbaum. Hillsdale, New Jersey.

Schnattinger, K. and Hahn, U. (1998). Quality-Based Learning. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-98),* pp. 160-164. Brighton, UK.

Schriver, K. A. (1989) Evaluating Text Quality: The Continuum From Text-Focused to Reader-Focused Methods. *IEEE Transactions on Professional Communication*, 32 (4), pp. 238-255.

Scott, D. and Hovy, E., Eds. (1993). *Burning Issues in Discourse.* NATO Advanced Research Workshop. Maratea, Italy.

Soderland, S., Fisher, D., Aseltine, J. and Lehnert, W. (1995) CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-'95),* pp. 1314-1319. Montreal, Canada.

Sparck Jones, K. (1972) Keyword Classification for Information Retrieval. *Journal of Documentation,* 28, 11. Butterworths, London.

Sparck Jones, K. (1983) Compound Noun Interpretation Problems. *Technical Report No. 45,* , University of Cambridge Computer Laboratory, England.

Sparck Jones, K. (1993a) Discourse Modelling for Automatic Summarising. *Technical Report No. 290,* University of Cambridge Computer Laboratory, Cambridge, England.

Sparck Jones, K. (1993b) What might be in a Summary? In Knorz, Krause and Womer-Hacker (Eds.) *Information Retrieval 93: Von der Modellierung zur Anwendung,* pp. 9-26. Universitatsverlag Konstanz.

Sparck Jones, K. (1994) Reflections on TREC. *Information Processing and Management,* 1994.

Sparck Jones, K. and Tait, J. (1984) Automatic Search Term Variant Generation. *Journal of Documentation,* 40, pp. 50-66.

Sparck Jones, K. (1997) The way forward in information retrieval. In *Elsnet News,* July 1997, pp. 12-13.

Sussna, M. (1993) Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network, In *Proceedings of the Second International Conference on Information and Knowledge Management.* Arlington, Virginia.

Sussna, M. (1995) Information Retrieval using Semantic Distance in WordNet. Paper made available by personal communication with author.

Teufel, S. H. and Moens, M. (1997) Sentence Extraction as a Classification Task. In Mani, I. and Maybury, M., Eds. (1997), pp. 58-65.

Teufel, S. H. and Moens, M. (1998) Sentence Extraction and Rhetorical Classification for Flexible Abstracts. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization.* Spring 1998, Technical Report, pp. 16-25.

Vanderwende, L. (1993) SENS: The System for Evaluating Noun Sequences. In Jensen, K., Heidorn, G. and Richardson, S. (Eds.) *Natural Language Processing: The PLNLP Approach,* pp. 161-173. Kluwer Academic.

Wacholder, N. (1998) Simplex NPs Clustered by Head: A Method for Identifying Significant Topics within a Document. In *Proceedings of the Workshop on The Computational Treatment of Nominals: COLING-ACL '98,* pp. 70-79. Montreal, Canada.

Wacholder, N., Ravin, Y., and Choi, M. Disambiguation of Proper Names in Text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics,* pp. 202-208.

Whiteland, J. (1994) The SISTA Project. Seminar, ITRI, University of Brighton, March 1994.

Wilks, Y. A. (1987) Text Searching with Templates. *Technical Report ML 162,* Cambridge Language Research Unit.

Wilks, Y. (1997) Information Extraction as a Core Language Technology. In Patienze (1997), pp. 1-9.

Wilks, Y. A., Fass, D. C., Guo, C. M., McDonald, J. E., Plate, T. and Slator, B. M. (1987). A tractable machine dictionary as a resource for computational semantics. In Boguraev, B. K. and Briscoe, T., Eds. (1989), *Computational Lexicography for Natural Language Processing,* pp. 193-228. Longman. Harlow, England.

Wilks, Y. A., Fass, D. C., Guo, C. M., McDonald, J. E., Plate, T. and Slator, B. M. (1988). Machine tractable dictionaries as tools and resources for natural language processing. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88),* pp. 750-755.

Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T., and Slator, B. (1993) Providing Machine Tractable Dictionary Tools, in Pustejovsky, J. (Ed.), *Semantics and the Lexicon* , pp. 341-401.

Wilks, Y. A., Slator, B. M. and Guthrie, L. M. (1996) Electric Words: Dictionaries, Computers and Meanings. MIT Press. Cambridge, MA and London, England.

Williams, E. W. (1981) On the Notions of "Lexically Related" and "Head" of a Word. *Linguistic Inquiry*, 12, pp. 245-274.

Williams, J.M. (1990) *Style: Toward Clarity and Grace*. University of Chicago Press.

Yarowsky, D. (1992) Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In*Proceedings of the Fifteenth Conference on Computational Linguistics. COLING-92*, vol. II, pp. 454-460. Nantes, France.