Terms in Context

Jennifer Pearson

Studies in Corpus Linguistics 1

JOHN BENJAMINS PUBLISHING COMPANY

TERMS IN CONTEXT

SCL

Studies in Corpus Linguistics

Studies in Corpus Linguistics aims to provide insights into the way a corpus can be used, the type of findings that can be obtained, the possible applications of these findings as well as the theoretical changes that corpus work can bring into linguistics and language engineering. The main concern of SCL will be to present findings based on, or related to, the cumulative effect of naturally occuring language and on the interpretation of frequency and distributional data.

Volume 1

Jennifer Pearson

Terms in Context

Terms in Context

JENNIFER PEARSON

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM / PHILADELPHIA

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences — Permanence of Paper for Printed Library Materials, ANSI Z39.48–1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order* by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

Pearson, Jennifer
Terms in context / Jennifer Pearson
p. cm. -- (Studies in corpus linguistics, ISSN 1388-0373 ; v. 1)
Based on the author's thesis.
Includes bibliographical references and index.
1. Terms and phrases--Data processing. 2. Semantics. 3. Lexicography. I. Title. II. Series.
B305.18.D38P4 1998
401'.4--dc21 98-15154
ISBN 90 272 2269 X (Eur.) / 1 55619 342 4 (US) (Pb; alk. paper) CIP

John Benjamins Publishing Co. · P.O.Box 75577 · 1070 AN AMSTERDAM · The Netherlands
John Benjamins North America · P.O.Box 27519 · Philadelphia PA 19118-0519 · USA

For Alan and Daniel

Contents

Acknowledgements

I am grateful to many people for their support and encouragement. Elena Tognini-Bonelli, who suggested that I should embark on this project and provided moral support when it was most needed. John Sinclair, who supervised the dissertation on which this book is based; I am deeply indebted to him for his generosity of spirit, his time and his challenging comments. Wolfgang Teubert for his helpful comments on an earlier version of this book. Dublin City University for granting me sabbatical leave to complete this project. My colleagues in the School of Applied Language and Intercultural Studies, and especially Lynne Bowker and Dorothy Kenny, for their interest and support. My family and friends for their patience and understanding.

I am grateful to the following people at the University of Birmingham for their assistance in providing material support. the Cobuild Unit, and Tim Lane in particular, for setting up the GCSE corpus and arranging access; Oliver Mason of the Corpus Linguistics Group for computational support: Geoff Barnbrook of the Corpus Linguistics Group for his assistance in providing additional data; Tim Johns for providing domain-specific corpora. My thanks to Mike Scott for developing WordSmith, a powerful concordancer for the PC user.

I would like to thank the following for permission to reproduce extracts from their material: MacMillan Magazines Limited for permission to use extracts from *Nature*; the Cobuild Unit at the University of Birmingham for permission to reproduce extracts from the GCSE corpus.

0 Introduction

> It is now recognised that the only practical means of processing lexical data is by computer.
>
> (Sager 1990:129)

0.1 Background

As the title suggests, this book is about 'terms' in 'context'. In essence, it seeks to demonstrate that corpora can be used for semi-automatic terminography. Metalanguage patterns are a common feature of certain types of specialised text and frequently offer clues to the meanings of the terms to which they refer; this book will describe a methodology for retrieving and manipulating these metalanguage patterns so that they can be used in the formulation of terminological definitions. Corpora are already being used by some 'modern' terminologists as a basis for recognising and extracting terms and for retrieving contextual fragments. To date, however, corpora have not been used for specialised lexicography in the same way as they have been used for general language lexicography. A number of possible reasons spring to mind: lack of availability of appropriate corpora, a reluctance on the part of 'traditional' terminologists to rely on 'authentic' text, a conviction that terms are different from words and can only be defined by suitably qualified subject specialists.

It is true that, for a long time, it was difficult to get hold of what might be described as specialized corpora. However, the situation has improved in recent years with increasing availability of electronic text which means that it is now generally possible either to create one's own specialized corpus or to obtain permission to use an existing one. It is also true that terminologists have only started to use corpora in their work relatively recently and there are still many terminologists who have yet to be convinced of the advantages of using corpora. This is not unlike what happened in the early days of corpus linguistics when there was a clearly identifiable gap between traditional linguists who continued to rely on their own intuitions and corpus linguists who advocated that the text held the real truth. In terminology, there are two camps: the 'modern' terminologists and the 'traditional' terminologists where ''there is a major division between those who believe context to be relevant for the identification of usage and those who believe terms to be context indepen-

dent'' (Sager 1990:10). 'Traditional' terminologists tend to study terms in isolation from text and to ignore context, even when the terms have originally been sourced in text while 'modern' terminologists pay attention to usage, albeit mainly in the context of term recognition and the retrieval of appropriate contextual fragments. Although they use different methodologies, 'modern' and 'traditional' terminologists share a fairly similar worldview which is rooted in theoretical terminology. The theoretical approach conceived in the early part of this century by people such as Eugen Wüster, and subsequently applied and developed by standardizing bodies such as ISO (International Organization for Standardization) and individuals such as Rondeau and Dubuc perceives a clear distinction between words and terms. Terms are labels for concepts which are abstract entities isolated from text. In traditional terminology, the emphasis is on defining concepts, on isolating meaning prior to agreeing an appropriate label (i.e. term) for a concept. The term which is agreed may be a single word or a multiword unit. In the context of standardization, this label becomes the approved term. Where a term already exists, terminology is concerned with identifying the precise concept with which it is associated. The description of a concept is obtained by means of consultation with subject specialists. This description becomes the definition of the term. All terminologists, whether 'traditional' or 'modern', are concerned essentially with establishing knowledge structures for subject domains and they build these knowledge structures by comparing and contrasting related concepts, by examining vertical and horizontal links between concepts. The ordering and classification of knowledge are crucial to terminological studies. In 'modern' terminology practice, the emphasis is much more on usage with the use of ''real text as a primary source of data'' (Ahmad, Fulford, Rogers 1992:141) but the objective and the underlying principles remain the same. The notion that terms are clearly distinguishable from words is one which will paradoxically be both challenged and supported in this book.

0.2 Target Audience

It is hoped that this book, by adopting a new approach to terminography, or specialised lexicography, will be of interest to four distinct communities. First, it should be of interest to those in the traditional terminology community who have not previously considered adopting a corpus-based approach to their work or at least not on the scale proposed here. Second, it should be of interest to those in the 'modern' terminology community who use text primarily for the identification of terms and the retrieval of contextual examples. Third, it should be of interest to those in the corpus linguistic community who have hitherto used general language corpora for the purposes of lexicography and have not previously considered using

special purpose corpora for more specific lexicographic studies. Finally, it should be of interest to people in the ESP/LSP community who are interested in showing students how to use text as a means of ascertaining the meaning of terms.

0.3 General Outline

The first three chapters of this book are designed to be introductory chapters and will appeal to a greater or lesser degree to each of the above named communities depending on their familiarity with these areas. The first chapter aims to introduce non-terminologists to the principles of terminology, the notion of languages for specific purposes and sublanguage. As it challenges some of the traditional distinctions which terminologists make between terms and words, it should also be of some interest to terminologists. Chapter two, which provides an overview of corpus types and corpus compilation and text classification criteria is aimed at those who have little previous experience of working with corpora and specialised corpora in particular. Chapter three provides a general introduction to different approaches to general lexicography and should be of interest to those who have previously worked only in the area of specialised lexicography or terminography. The remaining chapters of this book describe how we devised and implemented our approach to corpus-based terminography.

0.4 Chapter Contents

Chapter one starts with a very brief overview of the origins and principles of the theoretical approach to terminology. It explains why it is difficult, in a corpus-based approach, to use the distinctions which terminologists, whether traditional or modern, make between 'words' and 'terms'. An attempt is made to devise a framework for distinguishing between words and terms which can be used in a computational environment. It seems that the most important factor in determining whether a particular lexical unit (e.g. a lexical unit such as *gate* which has a general language meaning which is quite distinct from its meaning in the domain of, for example, electronics) is to be interpreted as a word or a term is that of communicative setting. Words or phrases which can be described as *terms* in some communicative settings may be perceived simply as *words* in others. When some words or phrases which are used as terms by a particular community of specialists (e.g. the term *quantum* when used by physicists) are borrowed by a wider community, the link between the term and its meaning may become gradually more blurred until the term simply becomes a word, part of the general vocabulary used by the speech community as a whole.

Chapter two starts with an overview of different types of corpora and a description of the criteria used in corpus linguistics for classifying texts and compiling corpora. While the overview is designed primarily for readers who have little or no experience of corpus linguistics, it is also used as a springboard for devising text selection criteria for the compilation of specialized corpora. Very little literature is available on this subject, mainly because such corpora have hitherto been difficult to obtain and researchers have had to use whatever corpora they managed to obtain rather than devise stringent criteria prior to the corpus compilation process. We were in a slightly more fortunate position in that we had ready access to material available on the Internet and material held at the Cobuild unit at the University of Birmingham and were thus in a position to select the material which was most suitable. The criteria which were deemed to be most relevant are outlined and the three corpora which were selected for the investigation are described. These are the ITU (International Telecommunications Union) 4.7m word corpus available on CD-ROM from the University of Edinburgh, the 1m word GCSE corpus (General Certificate of Secondary Education) made available by the Cobuild Unit at the University of Birmingham, and the 230,000 word Nature corpus kindly provided by Tim Johns at the University of Birmingham. Each of the corpora is described in terms of size, type, function, authorship and intended readership. All three corpora are suitable candidates for the retrieval of metalanguage patterns but, as analysis in later chapters will reveal, the ITU and GCSE corpora are particularly suitable for this type of investigation.

While the main focus of this book is the retrieval from corpora of information about the meaning of terms, we are also interested in devising appropriate methods of expressing the information which has been retrieved. Should the definitions be formulated in conventional dictionary-ese, in ordinary prose? Chapter three describes some general lexicographic principles and provides an overview of how definitions are expressed in different types of dictionaries. Three different approaches to the construction of dictionary entries are described. These are what we term the conventional approach, the Cobuild approach and the Mel'cuk approach. The relative usefulness of these methods is discussed and it is suggested that the method adopted by Cobuild could prove to be as appropriate for technical terms as it is for general language words.

As we are interested in identifying metalanguage patterns which can be used as input for terminological definitions, we look, in chapter four, at some of the research which has been carried out on the role and expression of definition statements in text with particular reference to research into the teaching of English for special purposes to non-native speakers of English. Much of the research in this area has focused on teaching non-native speakers of English how to *formulate* definitions rather than on documenting how definitions are actually *expressed* in reality. The

investigation proves, nonetheless, to be useful for establishing what types of definitions are likely to be preferred for terms, and the examples provided by the authors, while not necessarily sourced in authentic texts, actually serve as useful input for the specifications which we devise in later chapters.

Chapter five explores the notion of defining as a performative action. We look at Austin's classification of performatives in general and the conditions which must apply for such performatives to be valid. We then focus on the defining performative in particular and make the case for distinguishing between different types of defining performatives in text. We distinguish between situations where authors are defining new concepts or modifying definitions of concepts which already exist (defining exercitives) and other situations where the definition of an existing concept is simply being repeated or expressed in another way but the essence of the definition remains the same (defining expositives). We suggest that defining expositives in particular are a feature of some of the types of communicative setting described in chapter one.

The identification and retrieval of terms from corpora are an important element of our investigation. This is an area which has already engaged the minds of many researchers and is well documented in the literature. Chapter six provides a very brief overview of some of the approaches adopted. For the purposes of our investigation, we are interested in particular in retrieving terms which co-occur with metalanguage statements which provide some information about the meaning or scope of a term. Consequently, we are not interested in producing an exhaustive list of all possible terms in the corpora. However, we do need to devise a mechanism for retrieving the terms which are of potential interest to us. What we have done is to use a fairly conventional pattern-matching approach which retrieves all term candidates which correspond to one of a set of term formation patterns which we have specified. In addition, we use a refinement mechanism which allows us to isolate those term candidates which are likely to co-occur with some form of metalanguage pattern. These terms are used as input for the syntactic patterns specified in the following chapters.

Chapter seven builds on the hypothesis formulated in chapter five, namely that authors writing within certain specified communicative settings are likely to provide explanations of some of the terms which they use. We demonstrate that certain syntactic patterns combined with certain other characteristics are actually metalanguage statements which function as what we have termed partial or complete defining expositives. This chapter focuses on the retrieval of defining expositives which correspond to what Trimble (1985) defines as formal and semi-formal definitions and what we have termed formal and semi-formal defining expositives. In a formal defining expositive, an author describes a term in terms of its superordinate and a distinguishing characteristic (i.e. X=Y+ distinguishing char-

acteristic). These are considered to be complete definition statements. In a semi-formal defining expositive, an author describes a term by its distinguishing characteristic alone, without specifying its superordinate. The former are considered to be complete defining expositives, the latter partial defining expositives. While we focus on the expression of expositives within the boundary of one sentence (simple defining expositives), we also examine instances of expositives which are expressed in more than one sentence (complex defining expositives). A set of conditions which can be used to retrieve complete and partial defining expositives in corpora is specified. Each of the conditions is explained and supported by evidence from the corpora.

Whereas chapter seven focused on the retrieval of full sentences which function as complete or partial defining expositives, chapter eight looks at smaller segments of text which also provide some information about the meaning of a term. Here, too, we have identified a number of patterns where the information which appears before and after the pattern is in some way equivalent. The equivalence relation may be one of substitution, paraphrasing or synonymy. We define what we mean by synonymy, paraphrasing and substitution and use these definitions to classify the patterns identified in the corpora.

Chapter nine explores the possibility of using terms rather than syntactic patterns as the nodes in the retrieval process. The concordances of a number of examples from the corpora are examined to show that it is possible to retrieve not only information about the meaning of a term but also information about related terms. By combining information which has been obtained using the sets of syntactic patterns specified in chapters seven and eight with information obtained using the term-centred approach, we start to build a terminological entry. The terminological entry contains a definition, information about related terms and, where available, information about usage.

In the concluding chapter of this book, we examine some of the possible implications of the work described here. We explore in particular the areas of corpus-based term retrieval, text evaluation and corpus design and terminography and the potential use of corpora in an LSP teaching environment.

1 Identifying differences between words and terms

> . . . the differentiation of terms from words is not straightforward, since the relationship between general language and sublanguages . . . is an interdependent one. (*Pointer final report* 1996, Section 4, p. 17)

1.1 Introduction

Language as a whole is a label used to describe all language and all language situations. It includes not only the language which we use to communicate in everyday situations but also the language which we use in 'special' situations, termed language for general purposes (LGP) and language for specific purposes (LSP) respectively. LSP is frequently called sublanguage by researchers in natural language processing. Terminologists, LSP and sublanguage researchers contend that what distinguishes LSP from LGP are restrictions on vocabulary and syntax. Terminologists hold that words become terms, i.e. acquire or have protected status when they are used in special subject domains. NLP researchers hold that the lexis and grammar used in certain subject domains or in certain text types are restricted. The manner in which *term* is defined may vary from one of these groups to another but the belief that terms are different from words remains constant. This chapter discusses some of the definitions of *term* which have been proposed. The chapter starts with a summary of the emergence of terminology as a discipline and an overview of the basic concepts underlying the traditional approach to terminology, as devised by Eugen Wüster who was one of the earliest proponents of this new approach to language description. The traditional approach to terminology is concerned primarily with fixing the relationship between terms and concepts in order to facilitate communication. Some of the problems associated with the traditional approach, in particular the question of what happens when terms are actually used in text rather than simply as labels for concepts in knowledge structures or classification systems will be discussed. We find that the traditional perception of *term* is somewhat idealised and difficult to apply in a computational environment. The following section outlines what have been described as more pragmatic definitions of *term*; these definitions are generally provided by people working in the field of LSP and sublanguage. We find that there are problems too with these definitions. The pragmatists tend to

focus more on distinguishing between different types of terms than on the notion of term itself. There is an assumption that terms are instantly recognisable and that the only real problem is how to distinguish between different types of term. Next, we explore the possibility of using the criterion of standardization to distinguish between terms and non-terms and find that this too is problematic. We then look to sublanguage descriptions to establish whether they can help us to ascertain what a term is but we find that they are of little practical use because they focus on very restricted subsets of language. In summary, we find that, in spite of extensive research in the field of terminology and in the field of sublanguages, there is no usable definition of term and no adequate communication model which allows us to identify when words are being used as terms.

While we accept that there are indeed differences between words and terms, we find that, without human intervention, it is not possible to use any of the proposed definitions of *term* as a means of distinguishing between terms and words. This is because terms very often look the same as words and frequently not only look the same as words but can also function as words, albeit in different circumstances.

In the absence of a usable definition of term, we approach the question from a different angle and look more closely at the circumstances in which terms are likely to be used. We suggest that it is futile to propose differences between words and terms without reference to the circumstances in which they are used. If we wish to compile corpora for terminology studies and if we wish to minimize human intervention, we need to establish a definition of *term* which will allow us to distinguish between words and terms in a computational environment. As we believe that the communicative setting will determine the likelihood of the presence of terms, a number of communicative settings are described. We contend that within these settings there is a tendency to use terms. Depending on one's point of view, i.e. whether terminologist, LSP or sublanguage specialist, one's perception of what constitutes a term within these communicative settings will differ. Some words will always be perceived as terms (standardized, non-standardized terms), others will occasionally be perceived as terms (non-standardized and subtechnical terms) and others again may never be perceived as terms but perhaps should be (subtechnical terms). We will suggest that, from a user's point of view, it may ultimately be more sensible to adopt an inclusive approach which does not distinguish between different categories of terms and to consider simply that all language used in certain communicative settings is potentially terminological, unless otherwise demonstrated.

1.2 Emergence of terminology as a discipline

A number of developments in the early part of this century led to an interest in the special usage of language. Rapid technological progress led to an explosion of new

concepts which needed to be named. The internationalisation of trade created a need for equivalent terminology in a range of languages. With the formulation and dissemination of new ideas, new terminology was being coined. Words were being selected from the general pool of language and assigned new, additional or more precise meanings. Others were being taken from older technical fields such as physics or mathematics and assigned different or related meanings in new disciplines. It was becoming clear that the speed of technological progress was such that it was no longer possible to control the naming of new concepts and there was a danger that the same concept might be named differently by different communities creating confusion and communication difficulties.

These developments led to a growing acknowledgment of the need for standardization in all areas, including language. In his introduction to the chapter on the terminological/standardized dictionary, Opitz (1983:163) quotes Wüster: ''So prolific and various has technical intellectual work now become, as compared with previous centuries, that natural and uncontrolled evolution of technical terms can no longer be relied on to ensure unambiguity and efficiency in the use of language.'' It was this concern about possible confusion in science and technology and a desire for greater linguistic clarity which led to the emergence of a principled approach to the naming and description of concepts. The first attempt to standardize technical terminology was carried out by the International Electrotechnical Commission (IEC) in 1906 which undertook to produce the Vocabulaire électro-technique international. This work appeared in 1938 and was designed as a reference document for people working in the electrotechnical field. A significant development came in 1931 with the publication of the doctoral dissertation of the engineer Eugen Wüster. His dissertation outlined a new approach to terminology. While the IEC, in compiling its vocabulary, had been concerned with the standardization of existing terminology, Wüster was interested in establishing principles for the creation of new terminology. He was concerned that the formation of new terms should be properly motivated. The thirties also saw the establishment of ISA (International Federation of Standardizing Associations) whose brief was to promote international trade by standardizing products and processes. ISA set up a technical committee to devise a set of principles for standardizing and presenting terminologies and its work was largely influenced by the approach adopted by Wüster. After the Second World War, a new organization, ISO (International Organization for Standardization) was created and it established a technical committee for terminology in 1951, known to this day as TC37. The first ISO recommendations for terminology were published in 1968.

1.3 What is terminology?

It is generally acknowledged that the word is polysemous and that it can refer to

three different entities which are not unrelated. Sager (1990:3) suggests that terminology can refer to:

1. the set of practices and methods used for the collection, description and presentation of terms;
2. a theory, i.e. the set of premises, arguments and conclusions required for explaining the relationships between concepts and terms which are fundamental for a coherent activity under 1;
3. a vocabulary of a special subject field.

Thus, terminology may be used to describe methods of collecting, disseminating and standardizing terms. This type of work is carried out by bodies concerned with making recommendations for the standardization of existing terminology and by those concerned with the collection and documentation of terminology, i.e. with the input to term banks, specialized dictionaries. Terminology may also be described as a theory; the word acquired this particular meaning as a result of the approach advocated by Eugen Wüster which will be discussed in Section 1.4 below. Finally, terminology may also be used to describe the vocabulary of a special subject field, the collection of words which one would normally associate with a particular discipline. These may be nouns, verbs, adjectives or adverbs which are considered to have a clearly defined meaning when used in the context for which they have been defined.

1.4 General theory of terminology

This section summarizes the general theory of terminology (often referred to as the traditional approach to terminology) as advocated by Wüster. Terminology gradually began to emerge as a separate linguistic discipline when people like Wüster argued that terms should be treated differently from general language words. Wüster suggested that work on terms differed from work on general language words in three respects. First, in contrast to lexicology where the lexical unit is the usual starting point, terminology work starts from the concept : ''Jede Terminologiearbeit geht von den Begriffen aus'' (Wüster 1979:1). The concept should be considered in isolation from its label or term. Concepts exist independently of terms and indeed independently of any particular language.

> Ein Begriff . . . ist das Gemeinsame, das Menschen an einer Mehrheit von Gegenständen feststellen und als Mittel des gedanklichen Ordnens (Begreifens) und darum auch zur Verständigung verwenden. Der Begriff ist also ein Denkelement.
>
> (Wüster 1979:7)

(Translation: A concept . . . consists of an aggregate of characteristics which we can cognize as being common to a number of individual objects and which we use as means for mental ordering and for communicataion. The concept is an element of thinking. (adapted from Felber 1984:103)

Concepts are mental constructs to which we assign labels. Each concept is the product of a mental process whereby objects and phenomena in the real world are first of all perceived or postulated. Once this has occurred such objects and phenomena take on an existence in the realm of our thoughts. This existence is an abstract one. Thinking depends on the manipulation of such abstractions which are bundles of properties (characteristics) assigned to objects, phenomena, events, etc., or classes of objects (phenomena, etc.). These abstractions are concepts.

The second distinction which Wüster makes is that terminologists are interested in vocabulary alone. They are not concerned with the theory of morphology or with syntax. This type of information will be provided by general language rules. As Wüster states: ''Nur die Benennung der Begriffe, der Wortschatz, ist den Terminologen wichtig'' (Wüster 1979:2) (Translation: Terminologists are interested only in the naming of concepts, vocabulary). This decision to leave questions of morphology and syntax aside is interesting, and confirms that Wüster perceived terms as being separate from words; different not only in terms of their meaning but in terms of their nature and use. They are a separate class which operate as labels and appear to work in much the same way as a system of proper names works in general language. There is a one-to-one correspondence between the term as label and the concept as mental construct and, ideally, a term refers uniquely to one and only one concept within a given subject field. As labels, terms are protected, set apart from language in use. Traditional terminologists such as Wüster were not concerned with examining terms in use; they were interested only in establishing what they represented. This becomes clearer in Wüster's discussion of the difference between what he terms the *Ist-Norm* and the *Soll-Norm*. ''In der Gemeinsprache gilt als Norm nur der tatsächliche Sprachgebrauch. Man kann ihn eindeutiger ''Ist-Norm'' nennen. (1979:2) (Translation: In general language, the sole prevailing norm is the norm of usage. We can call this the ''Ist-Norm''). Evidence for understanding the meaning of a word is gathered by examining language in use, and the outcome of the investigation is the Ist-Norm, the reflection of language as it is actually used.

Terminologists, on the other hand, according to Wüster, are concerned with imposing norms for the use of language. They are interested in devising the *Soll-Norm*, in dealing with language as it should be used, in fixing and standardizing meaning in order to avoid confusion. This led to the creation of standardized vocabularies. Wüster (1979:2) believed that the creation of standardized terminologies (Soll-Norm) would lead to these standardized terminologies becoming the Ist-Norm in

technical communication. Unfortunately, the fact that a standardized terminology exists does not guarantee that it will be used. Yet, the notion of fixed or standardized usage is central to Wüster's theory because these standardized terms were to be used as a means of representing the conceptual structures which underlie subject fields.

In summary, for Wüster, special subject domains comprise a series of concepts or mental constructs which are represented by terms. The relationship between terms and concepts is agreed and standardized. The relationship between concepts is represented by logical, ontological, and other relations which are used to construct hierarchical systems of concepts. The theory was developed in response to a need to standardize the terminology used by experts within closed subject domains. Wüster's objective was to standardize and fix the relationship between term and concept. Concepts are perceived as being 'pure', used only by closed communities who have agreed a set of principles for understanding. If we were to identify where Wüster stands in relation to the distinctions which are made between terms and words, Wüster would be located at one end of the scale where standardized terminology is to be found, where meaning is fixed and protected.

1.5 The 'traditional' definition of term

According to Rondeau (1984:19), the term is basically a linguistic sign in the Saussurian sense; it has a signifiant and a signifié. He gives the name *dénomination* to the label, and the name *notion* to the concept. Unlike Wüster, who used the word *term* to refer exclusively to the label, Rondeau uses the word *term* to describe the combination of *dénomination* and *notion*, i.e. the combination of label and concept.

Like Wüster, he argues that the terminologist must start with the concept and, once s/he has defined and described the concept, must then decide on which label is appropriate. The concept must be described in terms of its relations to other concepts in the same subject field. This echoes Wüster's desire for conceptually organized vocabularies. Rondeau claims that there is a distinction between words and terms but, apart from specifying that terms are used in special subject domains, he does not offer any verifiable distinctions which can be made between them.

Sager (1990:19) offers the following distinction between 'terms' and 'words':

> The lexicon of a special subject language reflects the organizational characteristics of the discipline by tending to provide as many lexical units as there are concepts conventionally established in the subspace and by restricting the reference of each such lexical unit to a well-defined region. Besides containing a large number of items which are endowed with the property of a special reference the lexicon of a special language also contains items of general reference which do not usually seem to be specific to any

> discipline or disciplines and whose referential properties are uniformly vague or generalized. The items which are characterized by special reference within a discipline are the 'terms' of that discipline, and collectively they form its 'terminology'; those which function in general reference over a variety of sublanguages are simply called 'words', and their totality the 'vocabulary'.

The definition is cited in full because it highlights some of the problems which arise when one tries to distinguish between terms and words. To take Sager's first assertion, i.e. that the lexicon of a special language tends to provide as many lexical units as there are concepts conventionally established in the subspace, this would seem to confirm what had been proposed by Wüster, namely that the lexicon reflects the conceptual structure of a subject field, that the reference of each lexical unit is restricted to the field in question and that the concepts are agreed (conventionally established). In principle, one can agree with Sager thus far. However, he goes on to say that the lexicon of a special language includes two classes of items, namely items with special reference and items with general reference. The items with special reference are terms; the items with general reference are words. This latter class consists of items which are not usually ''specific to any discipline or disciplines'' and their referential properties are ''uniformly vague or generalized''. He concludes therefore that they should be classed as words. He does not give any example of items of general reference but one can assume that if they can be said to belong to the lexicon of a subject field in the sense defined above, they must have some form of specific reference. Perhaps he is referring to what others (e.g. Yang 1986, cf. Section 1.6) describe as subtechnical terms, words which have special reference but which are used in more than one subject domain. These include words such as *factor, result, accuracy*. To claim that such words are precluded from being classified as terms is to distort the composition of the lexicon of a special subject field.

We would argue that all words or phrases which have special reference, regardless of the subject field to which they belong, and which may also form part of the lexicon of another subject field must be considered to be part of the terminology of that subject field. Thus, for example, as statistics terminology is now commonly used in computational linguistics, (e.g. for processing lexical density), we would argue that such statistical terms should also be included in the terminology of computational linguistics. Sager seems to suggest otherwise. When documenting the terminology of computational linguistics, one might choose to classify the statistics terms differently from those which one would associate exclusively with computational linguistics. They might be classified as a subfield within computational linguistics or they might even be classified as subtechnical or as general terms to indicate that they have the same reference in more than one subject domain. However, we do not accept that they should be ''simply called 'words'''. Sager is using words

as a catch-all category for all lexical items which do not fit neatly into his classification of terms. As already suggested, this problem of distinguishing between words and terms is one which recurs in the literature and it seems that there is a tendency, when a term does not fit neatly into a particular subject category, to dismiss it as not forming part of the core vocabulary of that particular subject field.

H. Felber (1983:8) defines three types of linguistic symbols:

> 1) the word, 2) the term, and 3) the thesaurus word… The *word* can have a multiplicity of nondefined meanings and shades of meanings or can be used for naming objects. The concrete meaning of a word is given by the context; in other words, it is dependent on context. The *term* is a linguistic symbol assigned to one or more concepts (defined meanings). The meaning of a *term* which is the concept, is dependent on the position of this concept in the system of concepts concerned. The *thesaurus word* is a word, for the most part a *term* or a name, that is used for indexing and retrieval of information in information systems.

Felber's distinctions also pose some problems. If words only acquire meaning through the context in which they are used and do not have a meaning on their own, as is suggested by Felber, one is tempted to ask whether they can meaningfully be described as linguistic symbols. His description might be more appropriate for function words. Contrary to what Felber suggests, there are many words which one might not consider to be terms but which do have defined meaning. Colours, for example, or abstract concepts such as happiness or love, concrete objects such as sheep, chair. It is unlikely that Felber would have classified such words as terms because this would mean that all words which can have defined meaning should be classified as terms.

When defining terms, Felber suggests that they are different from words because they have defined meanings but, as we have just noted, this is not a sufficient criterion for distinguishing between words and terms because words too can have a defined meaning. He goes on to say that the meaning (i.e. concept) underlying a term is derived from the position of the concept within the system of concepts. This is not unlike Wüster's notion that concepts within a given conceptual system can be defined in terms of the similarities and differences between them. The distinctions which Felber makes between words, terms and thesaurus word do little to clarify the differences which might exist between each of these categories.

ISO 1087 Vocabulary of Terminology (1990:5) offers the following definition of term:

> 5.3.1.2 **term**: Designation (5.3.1) of a defined concept (3.1) in a special language by a linguistic expression

> NOTE—A term may consists of one or more words (5.5.3.1) [i.e. simple term (5.5.5) or complex term (5.5.6)] or even contain symbols (5.3.1.1).

whereby 'designation' refers to ''any representation of a concept'' (1990:5) and 'concept' refers to ''a unit of thought constituted through abstraction on the basis of properties common to a set of objects'' (1990:1). The scope of this particular definition of term is very broad and can scarcely be described as adequate when it is compared with the definition of 'word' in the same ISO standard (1990:6):

> **word**: smallest linguistic unit conveying a specific meaning and capable of existing as a separate unit in a sentence
> NOTE—A written word is marked off by spaces or punctuation marks before and after.

1.5.1 Summary of traditional view of term

For traditional terminologists, the notion of term can apply to lexical items with special reference in a restricted subject field (Sager); it can be the label or linguistic symbol for a concept (ISO, Felber); it is the equivalent of de Saussure's linguistic sign, i.e. the combination of signifiant and signifié (Rondeau). Distinctions are made between technical terms which are used in a single subject field and general terms which are used in more than one subject field. Distinctions are also made between terms whereby the meaning (underlying concept) of terms is agreed, and therefore protected, and words where the meaning is not protected. It is easy to understand why traditional terminologists would wish to specify a one-to-one correspondence between concept and term because one-to-one correspondence reduces ambiguity and improves communication. It can facilitate the creation of conceptual hierarchies representing the knowledge structure of a subject field and the addition of new concepts to those hierarchies. It is useful for classification purposes, for the compilation of standardized terminologies. However, it is difficult to imagine how the definition of term as offered by those who subscribe to a theory of terminology can be applied in practice. For example, it would not be possible to use the criteria proposed by any of the authors discussed to decide on whether a lexical item in a text is being used as part of general vocabulary or as a term.

Wüster, for example, suggests that the Soll-Norm eventually becomes the Ist-Norm in specialized communication, i.e. that standardized terms become the terms which are used in text. The prescriptive (i.e. Soll-Norm) approach views the relationship between a term and its concept as being static once it has been fixed. How would it deal with terms which are misused, abused or re-used with a new meaning?

There is an assumption that terminology is used only by a closed expert commu-

nity, and that each subject field has its own discrete terminology. When a lexical item cannot be said to belong exclusively to one subject field, terminologists are not in agreement on how it should be treated. Furthermore, in the traditional approach, there appears to be a tendency to describe all special subject fields as separate entities. While this approach may be possible for the representation of the terminology of the exact sciences, it poses problems for other disciplines. With increasing interdisciplinarity, the demarcation lines between subject fields are becoming blurred; there is often considerable overlap between subject fields, Where subject fields overlap with each other and consequently have concepts/terms in common, should these terms be, as Sager suggests, simply called words?

To exclude general terms from consideration is to leave gaps in the conceptual hierarchy. The fact that a term such as *factor* for example has the same reference in a number of subject fields is not sufficient justification for excluding it from consideration.

How does the traditional terminology approach cope with the evolution of language? For example, ten years ago, the definition of CD-ROM would have specified that it is an electronic medium for the storage of written text whereas today it would be defined as a medium for the storage of video, graphics and sound as well as written text. The meaning of terms evolves as the need arises, regardless of what a standard prescribes.

Traditional terminologists adopt a prescriptive stance. It is difficult to see what this approach can contribute to our understanding of specialized texts beyond allowing us to identify standardized terms when they are used. Even then, we have no means of ascertaining whether the term is actually being used to refer to the concept to which it was originally assigned. The approach simply does not take account of language in use. It isolates terms, protects them and makes no allowances for variants or for language change, making it difficult to use in a computational environment.

1.6 Pragmatic definitions of term

In contrast to the traditional approach to terminology described in the previous section, this section will examine some more pragmatic approaches to the definition of term.

Hoffmann (1985:126-127) suggests that there are three different ways of approaching the question of what constitutes a term. There are those who take a very narrow view and suggest that subject-specific terminology alone should be given the status of terms and all other words should be considered as part of general vocabulary. There are those who suggest that all lexical units used in a particular LSP

can be considered to be terms and finally there are those (the most common of the three) who suggest that within a specialised vocabulary, there are three categories of terms: subject specific vocabulary, non subject-specific specialized vocabulary, and general vocabulary. For the latter group, the subject specific vocabulary consists of those words which are only used in one domain; they are monosemous. Non subject-specific specialized vocabulary includes words with special reference which are used in more than one domain, and general vocabulary includes words which do not have special reference in any particular domain but are perceived to be 'special' simply because they appear in the text. Hoffmann himself suggests that specialized texts contain three categories of words, the first two of which consist of terms: Fachwortschatz (subject-specific terms), allgemeinwissenschaftlicher Wortschatz (non subject-specific terms), and allgemeiner Wortschatz (general language words). Thus, he distinguishes between terms which have special reference in only one domain, terms which have special reference in more than one domain and ordinary words which are not terms at all. However, he acknowledges how difficult it is in practice to decide in any systematic way to which class a word belongs.

Trimble and Trimble (1978) do not define *term* itself but they distinguish between three categories of terms: highly technical terms, a 'bank' of technical terms and subtechnical terms. Of highly technical terms, they write that ''Little needs to be said about the highly technical terms. Each field has its own'' (1978:92). Those terms which are unique to a particular domain are considered to be highly technical. This suggests that Trimble and Trimble's highly technical terms are what Hoffmann calls subject specific terms. Of the 'bank' of technical terms, Trimble and Trimble (1978:92) write:

> It is true that there is also a bank of technical terms from which all disciplines can draw; but these as well as the specialized technical terms are usually learned by contact.

These appear to be the same as the non-subject-specific specialized vocabulary referred to by Hoffmann, i.e. words with special reference used in more than one domain. Of subtechnical terms Trimble and Trimble write that they are ''common words that have taken on special meanings in certain scientific and technical fields'' (1978:93). Examples cited include *control*, *operation*, *current*, *ground*, *sense*, *positive*, *contact*, *lead*, *folder*, *flux*. This category has become increasingly popular with terminologists but the criteria for category membership can differ. Trimble and Trimble suggest that subtechnical words are general language words that have taken on special meanings in certain fields. They do not specify whether the resultant terms are subject specific or non-subject specific. It may well be that some of these

in fact belong to the highly technical category and others to the 'bank' of technical terms and the only reason for creating the subtechnical category is the fact that these words are also used in general language.

Herbert divides terms into two categories, the first consisting of highly technical terms which ''usually have very specialized meanings'' (1965: v). These are likely to be the subject specific terms referred to by Trimble and Trimble. His second category consists of ''semi-scientific or semi-technical words which have a whole range of meanings and are frequently used idiomatically . . . work, plant, load, feed, force'' (1965: v). On the basis of the examples which Herbert provides, it seems that what he is suggesting is that there are general language words which, when used within special subject domains, may have different meanings from their general language meanings and/or may occasionally be used idiomatically. However, it is not clear from Herbert's description whether these words then have one meaning or more when they become terms. These may be the same as the subtechnical terms defined by Trimble and Trimble but it is not clear that this is what Herbert intended.

Godman and Payne also distinguish between two types of terms: technical terms and nontechnical terms. Technical terms are:

> those for which there is a congruity of concept between all scientists, whatever the language used . . . . In each case, the properties or characteristics can be enumerated to define the object in an unambiguous manner. (1981:24)

These, we assume, are the subject specific terms referred to by the previous authors, and Godman and Payne suggest that there is exact equivalence across languages.

Their nontechnical terms subdivide into:

> 1) terms of the general language: for example logical terms such as *coordinators, subordinators, determiners, quantifiers, adjuncts*, and 2) terms that can be described as a basic list for usage in science. The functions of the logical terms of the general language remain unaltered in scientific statements. (1981:28)

Godman and Payne are the only ones to suggest a general category for all words which do not fit into the category of technical terms. However, it is unfortunate that they should have chosen to call this category 'nontechnical terms' because it suggests that they are all terms when, in reality, only some words in this category are to be perceived as terms. Godman and Payne subdivide this latter category into two sub-categories, one of which consists of ordinary words and the other, of a basic list of terms for usage in science. With regard to the first subcategory, it is debatable how accurate it is to state that the functions of the words of the general language always remain unaltered when used in scientific statements. Godman and Payne

might find, for example, that words such as *if* and *or* have quite specific meaning in science statements. Godman and Payne's second category, the basic list, includes:

> terms that appear on first sight to belong to the general language but have, in fact, more limiting definition in their use in scientific statements. The terms are given a precise meaning and are thus ''purged of the ambiguity and vagueness of their meaning (Caws, 1964). Scientific information is provided by this limiting definition. (1981:28)

Terms in the basic list include the following: *study, assumption, inference, evidence, similarity*. These words are apparently given more specific meaning in scientific statements but the authors do not elaborate on how one can identify when they are functioning as terms and when they are functioning as ordinary words.

Yang (1986) advocates a distinction between subject-specific terms and what he calls subtechnical terms. The latter are terms which ''represent notions general to all, or most of, the subject fields'' (Yang 1986:98). The subtechnical terms which Yang identified in his corpus of scientific English include: *absolute, accuracy, electrical, fact, factor, result, feature*.

1.6.1 Shortcomings of the pragmatic approach

The distinctions made by the above authors were motivated initially by their need to define a language curriculum for non-native students who have to learn a language for special purposes. This would explain why the notion of *term* itself is never discussed; there is a tacit assumption that terms exist. It would also explain the insistence on distinguishing between different categories of terms. The categories range from subject-specific terms which students will readily identify as terms because they are unfamiliar to them, to general technical and/or subtechnical terms, ''the special meanings of words of this type are often learned with difficulty since in many cases the student must reject the common meaning'' (Trimble and Trimble 1978:92), to words of the ordinary language. Thus, we have Trimble and Trimble, and Godman and Payne, using familiarity and perhaps term origin as a criterion for distinguishing between the two categories. Words which are unknown in general vocabulary are deemed to be technical terms; words which are known in general vocabulary and are borrowed and assigned a special use in scientific statements are called subtechnical (Trimble and Trimble) or non-technical (Godman and Payne) terms. The criterion of familiarity is not a valid one because it would not be possible to measure it in any objective way. Nor does the fact that a term also has a general language meaning necessarily imply that it is any less specialized than a term which does not have a general language meaning.

Hoffmann, Yang and Herbert also identify two categories: subject specific terms

and non subject specific (Hoffmann), subtechnical (Yang) or semi-technical (Herbert) terms but divide the world differently from the previous authors. Their subject specific category is likely to be the same as that of Trimble and Trimble, and Godman and Payne. However, their second category is, we suspect, quite different and also better motivated. It consists of ''diejenigen lexikalischen Einheiten, die in mehreren bzw. sehr vielen Fachsprachen auftreten'' (Hoffmann 1985:127). These are terms which have the same reference in more than one subject domain. The examples provided would seem to suggest that whether the words are considered to be already known is unimportant; only the fact that they have the same reference in more than one domain appears to count. While this is certainly a more valid distinction than the distinction of familiarity, there are problems too with this distinction. For example, some of these subtechnical terms may not always function as subtechnical terms; they may not always represent notions which are general to all, or most subject fields. In other words, they may not always have the same special reference in different subject domains. We would suggest that at least some of the examples cited are polysemous terms with different referents in different subject domains and would like to exemplify this using the following examples.

Yang suggests that *absolute* and *factor* are subtechnical terms; this means that they have the same referent in more than one subject domain. We believe that this is not strictly accurate and would suggest that these terms are sometimes in fact no different from a term like *gate* which is used in a number of different domains but has a different referent in each domain. According to the definition of a subtechnical term, *absolute* and *factor* are supposed to have the same referent, regardless of the domain in which they are used. Yet, the dictionary shows us that *absolute* and *factor* do not always have the same special reference in different subject fields. In physics *absolute* means ''not relative to atmospheric pressure'' (*Collins English Dictionary* 1991); in law it means ''coming into effect immediately and not liable to be modified'' (*Collins English Dictionary* 1991). These are both clearly distinct terminological readings and the gap between the two readings is so great that we would suggest that *absolute* must be classified as subject-specific rather than as subtechnical. It just happens to be polysemous. Perhaps Yang recognized that there were also subject-specific readings but, if this was indeed the case, we still have no means of distinguishing between the subtechnical and subject-specific meanings. We would suggest that a similar problem arises with *factor. Factor* can be ''an element or cause that contributes to a result'' (*Collins English Dictionary* 1991). This particular reading does not give us any reason to consider it as a term but we can understand why, if it occurs sufficiently frequently in a number of domains, Yang might have been tempted to classify it as a subtechnical term. However, it also has a subject specific meaning; in mathematics, it can be ''one or two integers or polynomials whose product is a given integer or polynomial'' (*Collins English Dictionary* 1991). It

would appear that the first reading of *factor* shows that it is simply a general language word which happens to be used in a number of different domains and which may be used more precisely in special language domains. (Incidentally, it is worth noting that these authors never explain what they mean when they talk about more specific meaning). This would explain why it is classified as a subtechnical term but the second reading is clearly a subject-specific one, with *factor* having a specific meaning within mathematics alone. Thus, some subtechnical terms may also have a subject-specific reading. Yang does acknowledge that this can happen but does not specify how one can recognize when this occurs.

The second problem is that we have no means of knowing whether terms when classified as subtechnical terms are actually functioning as such. There may be occasions when words like *fundamental* and *correspondence* (from Godman and Payne's basic list) are not functioning as subtechnical terms at all but as part of general vocabulary. It would appear that any word which appears sufficiently frequently across a wide range of texts will be deemed to be a subtechnical term. In the corpora selected for investigation for this book, this would mean that words such as *question, report* and *meeting* would come under the heading of subtechnical term. Yang concluded that the terms he had selected were subtechnical terms because they appeared at or above a certain frequency in all or most of his texts. However, we would suggest that this does not tell him whether all of the occurrences are actually functioning as subtechnical terms.

1.6.2 Summary of discussion

When we looked at the more pragmatic definitions of term, we found that the broad notion of *term* is assumed to be understood and that most of the discussion focuses on distinguishing between different categories of term. We found that two broad classes of distinctions are made, the first using the criterion of known/unknown and the second distinguishing between subject-specific and non-subject-specific terms. The known/unknown distinction is rejected because it is difficult to see how 'knownness' and 'unknownness' can be measured objectively. The subject-specific/non-subject specific distinction is also rejected, not because it does not appear to be valid, but because it is too difficult to measure. We do not have any objective means of establishing whether a particular subtechnical term is indeed functioning as such or whether it is actually functioning as a subject-specific term or even as an ordinary word. We acknowledge that there may be a need to distinguish between different types of terms, especially, for example, for the purposes of special lexicography where lexicographers have to decide on a cut-off point for terms but feel that the distinctions offered here are not usable. The one question which is never addressed is how one might recognize a term, irrespective of the category to which it

might belong. In other words, none of the authors specifies how one might distinguish between terms and words in text.

1.7 Exploring other methods of distinguishing between words and terms

This section will investigate whether it is possible to use the criterion of standardization to distinguish between terms and words. We will also look at other criteria and conclude that it may be useful to explore the notion of communicative setting to help with the distinction.

1.7.1 Standardized terms

> Scientists and specialists in practical operations invent technical terms . . . . A technical term . . . and its meaning . . . fixed by an agreement of definition, which, in science, receives explicit formulation and strict adherence. (Bloomfield 1939:38)

One criterion which we chose to consider as a means of distinguishing between terms and words was the criterion of standardization. Terminology standardization is:

> acte par lequel un organisme officiel définit une notion et un terme pour la désigner de préférence à un autre ou à l'exclusion de tout autre, dans une ou plusieurs langues. (Boutin-Quesnel, Bélanger, Kerpan, Rousseau 1985:31)
> Translation: action by means of which an official body defines a concept and chooses a term for this concept in preference to all others, in one or more languages)

Terminology standardization involves official recognition and acceptance of a term and its meaning. It is achieved through close collaboration between terminologists and subject specialists. The standards which are created list terms which have a prescribed meaning and which are preferred over other terms which may have been used to designate the same concept in the past. The meaning of a standardized term is agreed and fixed by experts working within the domain. The procedure adopted is very much in the Wüsterian mode, its main objective being the naming of concepts and the establishment of conceptual hierarchies. As part of the standardizing procedure:

> the subject specialist is expected to establish the delimitation of the concept of the term to be standardized by locating it within a field and subfield, by presenting it in all its aspects and by placing it within the conceptual network in which it belongs. (Duquet-Picard 1983:95)

ISO R 704 Naming Principles lists 31 principles which are designed to help "unify and standardize concepts and terms or to create new ones" (1968:7). These include advice on language economy where conciseness is recommended, advice on the formulation of definitions, recommendations for preferring one term form over another, recommendations regarding synonyms. *ISO R919 Guide for the Preparation of Classified Vocabularies* (1969:6) provides "detailed guidance for authors of technical vocabularies and in particular of standardized vocabularies." The most important stages in the preparation of a standard are 1) defining the field of study which may be accomplished by consulting an existing subject classification system; 2) deciding on the form and layout of the vocabulary; 3) deciding on the number of concepts which are to be listed in a vocabulary (it should not exceed 1,000 concepts).

There are three types of source which should be used in preparing a vocabulary:

> 1) Terminological publications such as technical dictionaries and treatises devoted to problems of terminology; 2) Publications not specially devoted to terminology: handbooks and textbooks, technical encyclopedias, descriptive articles, commercial catalogues, catalogues of industrial fairs and exhibitions; 3) Classification tables, i.e. classified synopses of concepts pertaining to the field under consideration. (*ISO R 919 Guide for the Preparation of Classified Vocabularies* 1969:9).

Entries for terms will contain as a rule, a term number, the preferred term, agreed definition, the field or subfield in which the term is to be used, related terms and deprecated terms. They will not contain any indication of usage in terms of common collocates or grammatical restrictions. Deprecated terms are those terms which have been, or still are, used to refer to the same concept. By stipulating that they are now deprecated, the standardizing authority is attempting to prohibit further use of such terms.

Standardized terms are not always new terms in the sense that they do not suddenly come into existence because a standardizing body decrees it. Standardized terms are generally terms that have already been coined by users of the terminology. What the standardizing body does is give its seal of approval to one term and make recommendations for preferring that particular term over others which may have been used to describe the same concept in the past.

One might be tempted to suggest that if a term is standardized, it always qualifies as a term, giving us a tentative definition for term which might be as follows: a term is a word or phrase which has been assigned an agreed meaning and has been officially approved and published in a standard. Unfortunately, there are a number of reasons for rejecting this proposal.

First, while recommendations regarding the choice of source material exist, this does not mean that a standardized terminology prepared using this material will

necessarily include all of the terminology of a given domain. It will only include those terms for which a definition has already been agreed and fixed. There are many terms worthy of standardized status which do not appear in a standard simply because they have not yet been considered. It is also true that there are many others which have already been considered and rejected in favour of another term or because they are considered to be too general or too specialized.

Second, standards are not as widely disseminated as one might expect. Given that the role of standardizing bodies is a normative one and one which is designed to facilitate communication, one might expect standardized terminologies to be widely available and widely consulted. This would not appear to be the case. Standardized terminologies are expensive (e.g. the ISO glossary on cinematography costs £100) and are not widely available. Professionals working within a particular subject domain may not even be aware of the existence of a terminology standard for their domain. Standardizing bodies do not make any real attempt to disseminate the contents of their glossaries. For example, it would be reasonable to expect specialized dictionaries, which tend to be consulted more frequently than standards and which often specify in the introduction that standards have been consulted for the compilation of the dictionary, to flag those entries which have indeed been standardized. This does not happen and so the user has no means of knowing which term is preferable to another and may even innocently choose to select a term which would be considered to be deprecated by the standardizing bodies.

Third, users may still knowingly choose to use deprecated terms instead of standardized terms (e.g. *marketing* rather than the standardized *mercatique* in marketing terminology in French, or *memory* rather than *store* in computer terminology) because they are the ones which continue to be used by their community.

While, in many ways, the category of standardized term should be the least controversial of the different term categories which will be discussed here, it is not possible to decide whether or not a word or phrase is to be described as a term using the standardization criterion. This is because, as Ahmad, Fulford and Rogers (1992:43) suggest, standardized terminology is idealized. It does not always reflect language as it is used. To include only standardized terms would mean the inclusion of some terms which are never used, in spite of their standardized status, and the exclusion of others which are not standardized but are used. We conclude therefore that it is not a useful distinction for our purposes.

1.7.2 Non-standardized terms

As we have noted, many terms which one would like to categorize as such are not included in standards, either because they have not yet been considered, or because

they have been considered but rejected for being either too general or perhaps deprecated. We would suggest that, as far as the users of these terms are concerned, the meanings of these terms are just as fixed as the meaning of standardized terms; the fact that they do not have standardized status is actually of little relevance because, as already noted, many users simply do not know which terms have been standardized. Notwithstanding their status, we deem non-standardized terms to be the same as standardized terms; when they are assigned a specific meaning within a particular subject domain by people working within the field and when they are used within certain communicative settings, they are deemed to refer to that specific meaning. These terms, like standardized terms, have either been coined especially for the subject field to which they belong, have been borrowed from another subject domain or have been borrowed from the pool of general language. *Byte*, *virus* and *mouse* in computing are each examples of these three categories.

Perhaps we can widen the scope of our definition of term to include non-standardized terms. This would give us a definition which states that a word or phrase may be deemed to be a term iff the meaning of that term (or each meaning, in the case of polysemous terms) applies to one domain alone. We propose to call this a subject-specific term. Thus, a term may be polysemous in different domains but each of its meanings must refer to only one domain. Take as an example, the term *gate* which in *electronics* designates:

> a logic circuit having one or more input terminals and one output terminal, the output being switched between two voltage levels determined by the combination on input signals. (*Collins English Dictionary* 1991)

and in *rowing* designates:

> a hinged clasp to prevent the oar from jumping out of a rowlock. (*Collins English Dictionary* 1991)

Although this term appears to be used in more than one domain, it actually has a different referent in each of the domains which means that it can still be considered to be a subject specific term. However, there are problems with this definition of term too, not least the fact that it does not account for terms which have the same referent in more than one domain and which are still perceived by their users to be terms. These will include terms which are common to more than one strand of a discipline (e.g. basic science, basic engineering terms) and terms which may be used in more than one discipline (e.g. mathematical terms). They are probably what other authors have described as subtechnical terms. It is therefore clearly not accurate to state that a term must have reference in only one domain in order to qualify as a

term. The definition needs to be widened even further. If we were to include these subtechnical terms in our definition of term, we would find ourselves with a definition which states: a term is any word or phrase used to designate a concept in a subject field. This is interesting because it resembles quite closely the definition of *term* provided in the *Shorter Oxford Dictionary*:

> **term**: a word or phrase used in a definite or precise sense in some particular subject or discipline; a technical expression; any word or group of words expressing a notion, or conception, or used in a particular context.

This definition is so broad that, in many respects, it is not very helpful at all. It does not tell us how we are to recognize terms. Nor does it tell us how we can recognize whether words are being used in a definite or precise sense. However, it does tell us that they may be classified as belonging to a particular subject or discipline and that they will be used in a particular context. We now find ourselves in much the same position as others described previously who distinguished between different types of terms. We happen to have come up with a slightly different way of defining what terms are but we still have no objective means of recognizing them in text. The next question for us is to establish how to do this, and we suggest that the issue of communicative setting will be crucial and that it is only by examining context that one will be able to determine whether language is behaving 'terminologically' or normally. We further suggest that all other efforts to define what a term is and to examine what distinctions, if any, exist between different types of terms are irrelevant if context has not been considered.

1.7.3 Exploring the notion of communicative setting

People behave and speak differently in different situations. The way in which they refer to objects depends 1) on the context or situation in which they find themselves, and 2) on the type of knowledge which they bring to the particular situation. Thus, physicists speaking among themselves of physics are likely to use many words or phrases (i.e. terms) which the ordinary speaker of a language will not understand. They use a technical language, a language which has its own restricted vocabulary with terms which are clearly defined and understood by the physics community, but not necessarily accessible to outsiders. People have no difficulty in accepting that the language used is technical and that the words have specific meaning. What of a situation where doctors are discussing the causes, symptoms and treatment of full-blown AIDS? Are they, too, using technical language? The answer should probably be in the affirmative because doctors, like physicists, will use terms which for them have explicit meaning. However, a definite affirmative is less likely from the per-

spective of an outsider because many terms describing the symptoms and treatment of AIDS have seeped into general language as a result of widespread media coverage of the subject and people use AIDS related terms in ordinary language as if they really understood the precise meaning of these terms. What of the language used by presenters on gardening or cookery programmes on television? They also use terms which for them have specific meaning. Is the language used in these programmes likely to be perceived as being particularly specialized? If not, why not, and, if so, on what basis? Do we perceive words like *baste* (cookery or sewing), or *tack* (sewing or sailing) to be terms?

We believe that there is a direct correlation between the number of people who are familiar with a particular special vocabulary and the perception of that vocabulary as being specialized. The fewer the number of participants in a subject domain, the more the domain, and its vocabulary, are likely to be perceived as specialized. We assume that a word like *cryogenics* (physics) is more likely to be judged a term than, for example, *basting* (cooking or sewing) and *tacking* (sewing or sailing). *Cryogenics* will be deemed a term because people do not recognize it or do not know what it means. Also, it looks technical. *Basting* and *tacking*, on the other hand, look like ordinary words and will therefore feel familiar to most people and some may even know precisely what they mean. Perhaps then, what allows some people to determine whether or not a word or phrase should be considered to be a term is its relative infrequency in general language and/or the communicative setting in which it is used. We would suggest, however, that the first assumption is invalid and that the second one has never been adequately defined.

With regard to the notion of relative infrequency, we would accept that, while some terms may be infrequent words which are never used in everyday language and may therefore be identifiable as terms, there are others which may have a general meaning in everyday language and a quite specific meaning when used in special communicative settings. For example, people will not describe a phrase such as *part-time work* as being a term because they recognize it and believe it to be part of their general vocabulary. It is a phrase which may be used to describe any employment which is not undertaken on a full-time basis (students supplementing their grants, parents supplementing the household income etc.). Yet, if the same phrase is used within the context of employment law, it will be clearly defined and the definition may include stipulations about the number of hours worked per week, employees' rights, level of entitlements etc. In the context of employment law, it becomes a term. It is assigned explicit meaning and when used within the appropriate context, is deemed to have that meaning. It is even less likely that people would consider *clean room* to be a term unless, again, it were used in a particular communicative setting, in this case computing. Relative infrequency is therefore not a very useful criterion for distinguishing between terms and words.

With regard to the notion of communicative setting, we suggest that this may be the most important factor in allowing us to decide whether words are being used as terms or words. It is an area which has been neglected by terminologists because the assumption is that people know instinctively which communicative settings are likely to show a high occurrence of terms. Some researchers in NLP have argued that clear syntactic and lexical differences exist between general language and special language situations. In the next section we propose to examine these distinctions to ascertain whether they can be used for identifying situations where language is used terminologically.

1.8 Sublanguages

Depending on the branch of linguistics in which one is involved one will speak either of LSP, sublanguages, scientific languages, specialized languages. The term *sublanguage* appears to be used primarily by researchers concerned with the computational tractability of natural language. This section will look at some of the arguments which have been made in favour of a distinction between general language and sublanguage and some of the definitions of sublanguage which have been proposed. Our intention is to establish whether sublanguage descriptions can be used to help us define the types of texts in which we can expect to find terms.

Harris (1968) who is credited with the introduction of the concept, defines sublanguage as follows:

> Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it. (1968:152)

Thus, a sublanguage exhibits some form of closure, i.e. a finite set of words and grammatical constructions is used. While proposing that a sublanguage constitutes a subset of the sentences of a language, Harris argues that the same is not true for the grammar of a sublanguage:

> Thus the sublanguage grammar contains rules which the language violates and the language grammar contains rules which the sublanguage never meets. It follows that while the sentences of such science object-languages are included in the language as a whole, the grammar of these sublanguages intersects (rather than is included in) the grammar of the language as a whole. (1968:155)

He later confirms this hypothesis. While ‘‘the sentences of the sublanguage are a

subset of the sentences of, say, English, the grammar of the sublanguage is not a subgrammar of English'' (1988:39). In other words, the sublanguage grammar may have additional rules which would be considered deviant in general language. This point is made fairly consistently by other sublanguage researchers too.

Of particular interest for us is Harris' thesis that:

> When the word combinations of a language are described most efficiently, we obtain a strong correlation between differences in structure and differences in information. This correlation is stronger yet in sublanguages…Indeed, a major interest in analysing the language of science is not so much that such formal or quasi-formal systems exist, as that they can be used to characterize the information in the given sentence. (1988:40)

To demonstrate his hypothesis, Harris wrote a grammar which described the language of a series of research papers on immunology by listing ''how words occurred with each other in sentences of the articles, and collecting words with similar combinability into classes'' (1988:42). He found that there were fifteen classes which divided into classes of nouns and verbs (operators) and that there were ten main sentence types. The analysis focused on the verbs and the classes of nouns which could co-occur with them. The use of symbols allowed Harris to avoid synonyms, to disregard the internal composition of a class member, ''to omit grammatical requirements of the whole language that are irrelevant to the particular science (e.g. tense, plurals)'' (1988:49). Word classes were allowed to contain whole phrases such as *appears in, is found in* and did not require further analysis.

> The science language is then a body of canonical formulas representing the science statements after synonymy and the paraphrastic reductions have been undone. Its grammar states the class symbols (here, capitals), the class members (here, subscripts), the modifiers (superscripts), with the constraints on each and with the combination of them that constitute sentence types. (1988:53)

To test Harris' hypothesis, we attempted to use the same approach to represent sentences in the ITU corpus, one of the corpora used for the investigation in this book (cf. Section 2.11.1 for description of ITU corpus). It proved difficult to apply the Harris style grammar for a number of reasons. The first of these was sentence structure. None of the sentences which Harris analysed has a subordinate clause, whereas the opposite is true of the ITU corpus where few of the sentences consist of a single clause. Application proved difficult for another reason, the range of vocabulary in the ITU corpus. Even in a small set of just eight sentences, 7 classes of nouns, three classes of verbs and 3 other classes of words (prepositions, descriptors, connectors) were identified. While it is not possible to know from such

a small set of sentences exactly how many word classes or indeed types one might find in the entire corpus, it is certainly likely to be much larger than what Harris found (i.e. a total of fifteen classes for his corpus). A third problem which we encountered was in trying to specify, as Harris had done, which classes of nouns could occur with which classes of verbs. It simply was not possible to make any generalizations about relations between noun and verb classes. Our practical experience led us to conclude that this approach was not feasible for the size of our corpus and that the Harris definition of sublanguage was not appropriate for the type of texts with which we were working. We concluded therefore that it was likely to be feasible for only a limited number of texts and text types such as those described by Harris.

Lehrberger points out that we use labels such as the language of biophysics, or the language of pharmacology to describe subsets of language:

> as though there were certain well defined languages used by specialists in various fields. But a glance at technical or scientific writing reveals that the language used is basically a language such as English or French . . . If we can recognize that a text is 'in English' and yet feel that it is distinct enough to be described as being 'in the language of X' (physics, aeronautics, electronics etc.) then we may be justified in saying that the language of X is a 'sublanguage' of English. (1982:82)

Thus, the languages of physics and electronics, for example, are not different languages in the sense that French and English are but he believes that there is some justification for suggesting that they form a subset of the language in which they are used. In a later publication, Lehrberger states ''*sublanguages* are not determined a priori but emerge gradually through the use of language in various fields by specialists in those fields'' (1986:20). Lehrberger challenges Harris' definition of sublanguage arguing that the obvious relation between subsystem and system in mathematics is not as clear-cut in the case of the relation between sublanguage and natural language. To support his argument, he examines some of the assumptions which are made about the relation between sublanguage and natural language, one of the most common being that ''a sublanguage of a natural language L is part of L'' (1986:20). He believes that our definition of sublanguage depends on whether we consider it as being an independent system or part of, but on the fringe of, natural language.

He points to the ambiguity in linguistics about the concept underlying the term natural language and prefers to make a distinction between language as a whole, standard language and sublanguages. Standard language is the language defined by others as LGP or language for general purposes, the language used for everyday communication. Sublanguages differ from standard language because the lexis and

semantics are more restricted than in standard language, and the syntax may deviate in some respects from the syntax of standard language, thus lending support to Harris' contention that sublanguage grammars may be deviant from the grammar of general language. Language as a whole (L) is the term which Lehrberger uses to embrace standard language and all possible sublanguages. L "subsumes many varieties of speech and writing, including an indefinable number of sublanguages . . . With such an interpretation of L, a grammar of L is not likely to be available" (1986:22). This means that L is a label used to describe all possible varieties of language and that the grammar of some of these varieties may differ slightly or significantly from the grammar for general language.

The second assumption of Harris' which he challenges is the notion that "a sublanguage is identified with a particular semantic domain" (1986:20). The problem which Lehrberger sees with this assumption is that it can be difficult to decide to which semantic domain a text belongs. Frequently, specialized texts contain material from different semantic domains. Consequently, this criterion may not be particularly useful for defining a sublanguage. Our analysis of the ITU corpus would certainly bear this out. While telecommunications terminology is the most common terminology in the ITU corpus, there is also a large number of what should be described as general administrative terms. Furthermore, we would argue that even when a text can be easily assigned to a particular semantic domain, it will not necessarily make the task of describing the sublanguage any easier because of the variety of text types that one will find in any one semantic domain. This is confirmed by Lehrberger:

> Generally speaking, text purpose affects text structure in fairly predictable ways. It is important to bear in mind when dealing with subject-matter sublanguages that given a set of texts from the same subject-matter field, structural homogeneity is not to be expected if text purpose is not the same throughout. (1986:30)

Text purpose is a factor which Lehrberger considers to be important for determining restricted or deviant use of language. He argues here that even when two texts deal with the same topic, they may exhibit different lexical and syntactic patterns depending on the purpose of the texts.

Lehrberger lists six factors which help to characterize a sublanguage: (i) limited subject matter, (ii) lexical, syntactic and semantic restrictions, (iii) "deviant" rules of grammar, (iv) high frequency of certain constructions, (v) text structure and (vi) use of special symbols (1986:22). To support his argument, Lehrberger (1982) examined a corpus of aircraft maintenance manuals for factors such as restrictions, reductions and frequently occurring forms.

Restrictions include lexical restrictions whereby the number of types which will

appear in any given sublanguage is likely to be highly restricted and the number of parts of speech which can be assigned to any particular word will be restricted. Certain words, e.g. personal pronouns such as *I, we, he, she* will not appear in most sublanguage texts. Others will be subject field dependent. Others again may appear in a number of different sublanguages (e.g. common technical words). Restrictions also include syntactic restrictions. The type of syntactic structures used in a given sublanguage may depend on the text type. Thus, in aircraft maintenance manuals, the reader is unlikely to encounter interrogative sentences, use of the past tense, passive voice. Finally, Lehrberger discusses semantic restrictions which result in a reduction in ambiguity. The first of these restrictions is categorial whereby a word which may occur in more than one category in general language will only occur in one category in a sublanguage. The second of these restrictions relates to the number and type of semantic features which are assigned to words.

> Many nouns which designate either concrete or abstract objects in the language as a whole are used only concretely in this sublanguage . . . .The same is true of words that may be used for either human or non-human objects . . . .Verbs are likewise restricted in the kinds of subjects and objects they can take. (1986:23)

The most common forms of reduction in the texts analysed by Lehrberger are omission of the definite article and of the copula but these omissions are not systematic. Frequently occurring forms in the corpus which he analysed include the unusually frequent use of the imperative, the number of adjectives which never occur in predicative position and the ''presence of many long strings of nouns, or nouns and adjectives, within nominal groups'' (1982:110).

In an article originally published in 1972, N. Sager defines sublanguage as follows:

> The discourse in a science subfield has a more restricted grammar and far less ambiguity than has the language as a whole. We have found that the research papers in a given science subfield display such regularities of occurrence over and above those of the language as a whole that it is possible to write a grammar of the language used in the subfield, and that this specialized grammar closely reflects the informational structure of discourse in the subfield We use the term sublanguage for that part of the whole language which can be described by such a specialized grammar. (1982:9)

and, in a separate article with Hirschman in the same publication, defines it as follows:

> We define sublanguage here as the particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical speciality or

> science subfield), in which the authors of the documents share a common vocabulary and common habits of word usage. As a result, the documents display recurrent patterns of word co-occurrence that characterize discourse in this area and justify the term sublanguage. (1982:27)

Her argument in the earlier publication (as Harris also argued) is that the grammar used in certain science subfields reflects the informational structure of the discourse, and sublanguage is the term to be used for the subset of a language which can be described by a specialized grammar. In reality, however, much of the language used in these subfields does not fit the specialized grammars, and researchers generally have to clean or edit the language so that it fits the sublanguage grammars. For example, in the Linguistic String Project which was directed by Sager, some hand-editing had to be carried out in order to obtain a correct parse.

> If all else fails, the text can be edited by a human reader, adding a subject before an ambiguous verb for example, or an article before a noun. This is the least prized option at the Linguistic String Project, but an interactive system should be possible where the person entering data will be asked to paraphrase if no successful parse is obtained. (MacLeod, C., Chen, S., Clifford, J.M. 1987:172)

In the second definition, Sager and Hirschman suggest that the language used in a restricted subject area displays recurrent word patterns which can be exploited in order to retrieve the informational content of the text. Here, the emphasis is on a circumscribed subject area, authors with common vocabulary and habits of word usage and recurrent patterns of word co-occurrence. We think that a definition which does not take text function and target readership into account will run into difficulty because authors write for a purpose and for a readership and they tailor their language accordingly.

Kittredge believes that we do not have ''an empirically adequate definition of the term'' (1982:110):

> the closure property proposed by Z. Harris is not in itself sufficient [for allowing us to decide] what the limits are for a given sublanguage, and whether closely related varieties of language should be considered parts of the same sublanguage or as constituting separate systems. (1982:110)

He supports Hirschman and Sager's recommendation for including shared habits of word usage in the definition of sublanguage. While Hirschman and Sager also refer to a shared community of speakers as part of their definition, Kittredge is not sure that this can always be well-defined, particularly in the case of semantic domains

where ''access to the texts is relatively free (e.g. stock market reports, recipes, newspaper columns on playing bridge, weather bulletins etc.)'' (1982:110).

Kittredge is much more interested in text types (pharmacology reports, weather bulletins, recipes) than in entire subject fields because different text types dealing with the same circumscribed subject area may have quite different grammars. He outlines in detail how a number of text types can be said to use a restricted grammar which frequently deviates from standard grammar. He is doubtful about suggestions that sublanguages have a closed lexicon stating that ''a precise measure of size is possible only to the extent that the sublanguage is lexically closed, and it appears that few sublanguages are'' (1982:124) but more hopeful about suggestions that there may be restrictions on the number of word classes used.

1.8.1 Summary of discussion

Sublanguage research was motivated by the realization that natural language processing was unable to cope with all of natural language. It seemed reasonable to look at smaller subsets of language, especially as other linguists such as Swales (1971) and Widdowson (1979) had been suggesting for some time that there were subsets of general language which had a restricted syntax and vocabulary. Initially, natural language researchers focused on entire semantic domains but they very quickly shifted their attention to text types when they discovered that the range of language used in any given semantic domain was likely to be as broad as the range used in everyday communication in terms of the grammatical patterns used and was therefore unlikely to be any more tractable. The focus of attention shifted to text types because, as Kittredge states ''Functionally homogeneous texts referring to a single semantic domain normally make use of only a small subpart of the language's lexicon'' (1981:446). Even then, there were problems because, as Kittredge says ''the notion of sublanguage, . . . is essentially an abstract construct. One rarely finds a sublanguage which is totally sealed off from the rest of the language'' (1981:464). Kittredge believes that seepage is inevitable and that, for example, scientific articles will contain digressions, regional weather alerts will be phrased differently from normal weather bulletins and that no sublanguage system will be able to cope with these 'unexpected' structures. This author has had some experience of working with the 'sublanguage' of telecommunications in the Eurotra Machine Translation project where it proved to be impossible to process the texts as originally written. To cite just two examples: subordinate clauses had to be deleted or rewritten and, as the system was unable to cope with anaphoric reference all such references had to be edited. The project ended up working with a text which looked quite different from the original and was very much a simplified version. The Eurotra project was not alone in carrying out this type of pre-processing or in re-

stricting the types of sentences which were to undergo processing in order to improve the quality of the output. There are others such as Rank Xerox and IBM who have investigated the possibility of specifying controlled language input in order to improve the quality of the output from NLP systems. While we are sceptical about many of the claims made e.g. about the existence of a separate language such as the language of science, we believe that it may indeed be true that certain text types use a restricted set of word classes and a restricted grammar. This would apply, for example, to weather bulletins (e.g. the Taum-Météo MT system), sewing patterns (O'Brien 1993), recipes. We are not sure, however, that any of the sublanguage descriptions can be used to help us in identifying those texts which are likely to have a high frequency of terms. While they may identify some of them, we suspect that there are many other text types which are likely to have a high frequency of terms but would not qualify as sublanguages. Sublanguage description is more concerned with identifying those subsets of language which can be computed by machine than with providing descriptions of all subsets of language which exhibit some form of restriction and which might have proved useful for our purposes. This is phrased much more elegantly by Kittredge who says that, in many cases, these

> [sublanguage] systems have not been based on a thorough linguistic study of the sublanguage in question. In general, the attention of theoretical and descriptive linguistics to restricted language has lagged behind the attempts of computational linguists to make domain-based language processing systems work. (1981:446-447)

It seems that sublanguage descriptions do not provide any answers to our question about how and where we are likely to find terms. We must therefore devise our own method for doing this and, as we suggested in Section 1.7.3, we believe that if we can adequately define the communicative settings in which terms are likely to occur, we are much more likely to succeed in distinguishing between terms and words.

1.9 Classifying communicative settings

In this section, we will try to describe some of the types of communicative settings where we believe terminology is used and we will also try to ascertain whether there is any need to make distinctions between different classes of terms. While examples of text types will be provided, they are not intended to be exhaustive because there are probably many additional types of publication which fit the communicative settings described here.

All of the definitions of *term* provided hitherto in this chapter portray terms and words as being quite separate and different either on syntactic and/or semantic grounds. The differences are said to be semantic because terms are uniquely defined for a particular subject domain; they are syntactic because of term formation patterns. We have already rejected the latter criterion as a criterion on its own but would like to consider the former in a little more detail. The notion of special subject domain recurs constantly in terminology literature because membership of a subject field is an essential characteristic of termhood. What terminologists fail to do is put the notion of subject domain in context and explain exactly what they mean beyond using vague terms such as 'scientific' or 'technical' discourse. It will be suggested here that terms can only be considered as terms when they are used in certain contexts and that all of the discussion about whether or not a term is really a term is irrelevant if the discussion is not rooted in reality.

1.9.1 Expert-expert communication

When experts in any given field communicate about their subject, they tend, as we have already noted in Section 1.7.3, to use a highly specialized jargon. It is assumed that author and reader share a common language and that when certain words or phrases are used, each understands what is meant. This language differs from general language in that specific meanings have been assigned to the language used, and these have been defined prior to the communication act by an external authority. The external authority may be a standardizing body; it may be simply what is commonly held to be true about the domain; it may be a specialized dictionary dealing with the domain in question. Thus, depending on the field in which the experts are working, *big bang, quantum* and *chaos*, for example, will have unique and explicit meanings when used by experts within that field. A non-expert, when using these terms, will use them much more loosely and even incorrectly, as is often the case with the latter two examples. What makes these and other terms specialized in a discussion between experts is the communicative setting. The speakers agree to understand and use the terminology as originally defined and, in general, they will only explain the terminology which they are using when they are redefining an existing concept or if they are coining a new term. Writer and reader, or speaker and hearer are assumed to have the same or very similar level of expertise. This expert-expert communicative setting applies to publications in learned journals, academic books, research reports, legal documents such as laws and contracts and any other written documents where the author is writing about his/her area of expertise and addressing readers who are understood to have a similar level of expertise. While the communicative setting requires that the communication be written by experts for experts, no such restriction exists on the subject matter covered as long as it is

within the realms of their expertise. Thus, learned books may be books about any subject, ranging from quantum physics to organic horticulture. The status of the terminology which the experts use may differ in that there may be more standardized terminology in one field than another but this does not prevent other (non-standardized) terms from having equivalent status in the minds of the speakers. This particular communicative context is likely to be the one with the highest density of terms.

1.9.2 Expert to initiates

Frequently, experts working within a subject domain are called upon to communicate with others in their field who, while they have some knowledge of the field, do not have the same level of expertise. They may be students of a particular discipline, as in the case of advanced students in third level institutions. They may be people working within the same area but with a different training background, e.g. engineers and technicians, medical specialists and general practitioners. Here again, terms are likely to be used. While these experts will use the same terminology as they would use when communicating with their peers, they are likely to explain some terms which they believe to be unknown or inadequately understood by their readers.

What distinguishes this type of communicative setting from the expert-expert context is the difference in the level of expertise of the writer and reader. Consequently, term density is likely to be lower as the communication will be interspersed with explanations which may, when necessary, include the use of more general vocabulary, i.e. non-terms. As the function of the communication in this context is to assist the reader in improving their knowledge of the domain, explanations will be quite detailed and specific. This type of communicative setting arises in subject-specific textbooks which are aimed at people who already have some experience in the particular discipline. As with expert-expert communication, there is no restriction on the subject area as it is suggested that the function of the discourse will be similar regardless of the subject field.

1.9.3 Relative expert to the uninitiated

What we term the uninitiated are adults with a general education who are not necessarily involved, either professionally or through their leisure interests, in a particular subject field. Texts written for this audience are likely to have a much lower term density than in either of the previous two communicative settings. The only assumption made by authors is that people have a good knowledge of the language in which the communication is written. As no prior subject-specific knowledge is assumed,

authors may even choose to use a general language word to describe a concept rather than risk alienating their readers by using the more appropriate specialized term. This approach is very common in popular science journals such as the *New Scientist* or in special interest columns in newspapers, such as 'Computimes' in the *Irish Times* or 'Online' in the *Guardian*. When terms are used, the author either explains them or indicates that s/he thinks that the concept is already known to the reader. However, it is not absolutely necessary that author and reader have the same understanding of those terms which are assumed to be known. For example, in this age of computer technology, we have all started to use terms such as *log on, back up* and *modem* with relative ease but how many of us actually understand these in the precise way that experts understand them? What distinguishes this particular communicative setting from the two previous ones is that there is no need for author and reader to achieve the same level of understanding of the terms used as long as the broad thrust of the message is understood. Consequently, it is suggested here that this particular communicative setting is not conducive to terms being used in a rigorous manner or being perceived as such.

1.9.4 Teacher-pupil communication

The term 'pupil' is used to describe people who have no prior knowledge of a particular subject field but are required to acquire it for educational or professional purposes. It includes groups as diverse as secondary school pupils learning about science and academics learning about the Internet. What distinguishes this group from the audience described in the previous category is that they need to learn about a particular subject. The publications which they are likely to use for this purpose include introductory textbooks, handbooks and instruction manuals. Authors writing for this audience will use the appropriate terminology but will assume a much lower level of expertise than in the second category. Explanations and definitions will be provided more frequently and they will be expressed either in general language or in simplified technical language. While it is important that concepts are grasped, pupils are not required to reach the same level of understanding as those in the second category. Nonetheless, when terms are used, they are assigned specific reference within the particular subject domain and readers are required to understand them in this way.

1.9.5 Summary of discussion

Terminology is used in each of the four communicative settings described but the way in which it is used is not the same in all cases. As ISO states ''technical terms should have the same meaning for everyone who uses them. This goal can be

achieved only if there is general agreement on the meaning of these terms'' (*ISO/R 860* 1968:5). We believe that it is very important that technical terms ''should have the same meaning for everyone who uses them'' and that the communicative setting is one means of determining whether the conditions are conducive to this. The authors in the first, second and fourth categories described above will endeavour to use terminology in a precise way because the communicative setting requires it. In expert-expert communication (setting 1), there is an assumption that the terminology used is known and understood by the readers except in the case of recently coined terms which may be explained by the author. In communication between experts and initiates (setting 2), many basic concepts are known and understood and readers are expected, through their reading, to acquire and understand additional terminology in order to broaden their understanding of a subject field. In teacher-pupil communication (setting 4), basic concepts are explained with a view to introducing readers to a new or relatively unfamiliar subject field and, as with the previous category, the purpose is didactic and pupils are expected to understand the terminology as explained. In communication between relative experts and the uninitiated (setting 3) terminology is used in a much more popularized way than in any of the other categories. Authors in this setting are providing a general overview, merely a taste of a domain without the intention of building on this overview subsequently. They will tend on the whole to use analogies rather than definitions for explanations and will frequently use general language words instead of terms in order to avoid alienating their readers. The environment is less rigidly defined than for the other settings. The jargon used in this setting should not be considered to be terminology in the sense in which we are now defining it. There is too much scope for vagueness and misunderstanding within this communicative setting for it to warrant consideration as a source of terminology. What we are suggesting here is that the terminology used in settings 1, 2 and 4 is likely to be used in a precise way but that the terminology used in setting 3 is used in a less rigid manner and more as part of a general communicative situation. Consequently, we conclude that settings 1, 2 and 4 are reliable sources for potential term candidates, and that when terms are used within these contexts, we can assume that the people using the terms accept the stipulated and agreed meaning associated with these terms.

1.10 Conclusion

In this chapter, we have tried to look at the concept of *term* in the context of real text situations rather than as an abstract label for a concept in a classification system. The main objective was to establish when words acquire and lose terminological status. We found that there is a tendency to distinguish between different types

of terms with the distinguishing criteria ranging from known/unknown to subject-specific/non-subject-specific terms. We found these too vague to be usable or measurable and tried instead to use other criteria for recognizing terms. We investigated the possibility of making distinctions between standardized terms and general language words, standardized and non-standardized terms, subject specific terms and terms with the same reference in more than one domain. We began to wonder whether there was any real need for all of these distinctions because it did not advance us any further in our search for an objective means of recognizing terms. Did it really matter whether a term was subject-specific (i.e. special reference in one domain) or general (special reference in more than one domain)? Was it not more important to be able to stipulate when words were being used as terms and to retrieve all of the terms used in a particular domain than to be able to distinguish between subject-specific and general terms? This prompted us to investigate how we might do this. We thought that we might be able to use sublanguage descriptions provided by NLP researchers. We had envisaged that their claims regarding the lexical closure of sublanguages might be useful but found that sublanguage researchers are more concerned with identifying computationally tractable segments of language than with describing the entire discourse of a domain, or broad communicative settings.

It seemed that the main problem with all of the distinctions was that there was an underlying assumption that terms could be recognized intuitively. We believe we demonstrated that this was not the case and that there was a need to describe the situations in which language might be used terminologically. We therefore tried to define communicative settings in which we were likely to find words behaving as terms. We identified four sets of communicative settings in which words might be used as terms. We found that the first, second and last communicative settings were more likely to contain words which really were functioning as terms than the third communicative setting. In all other contexts, it is not possible to state with certainty that words which look like terms are actually being used as terms. Of course, not all of the words used in the specified contexts will be terms, and we will still have to distinguish between words and terms in these contexts. To do this, we use term formation and other selection criteria which are outlined in Chapter 6. From a practical point of view, we suggest that there is no need to distinguish between different types of terms because users will be more interested in distinguishing between term and word status than in knowing what type of term it is, i.e. whether or not a particular term is standardized and whether it has special reference in only one domain or in a number of different domains.

2 Corpora, corpus design and corpus selection

2.1 Introduction

The first chapter of this book dealt with the subject of terminology, its background and the differences which terminologists perceive between terms and words. It was in many respects an introductory chapter designed to familiarise non-terminologists with the basic notions of terminology. In an analogous manner, much of this chapter is designed to introduce people who are not familiar with corpus linguistics to the terminology of *corpus* linguistics. Thus, the first section of this chapter will define the word *corpus* and look at different categories of corpus which have been proposed by corpus linguists. Different approaches to corpus studies will also be described briefly.

When a decision is taken to compile a corpus, a number of general issues need to be addressed prior to compilation of the corpus and these are discussed in the next section of this chapter. Texts which have been selected for inclusion in a corpus are generally classified according to a number of different criteria. These criteria may be what corpus linguists term internal or external criteria. While we will focus on external criteria because these are the most widely documented, we will also discuss some internal criteria. Thus, the second section of this chapter will provide an overview of some of the design and classification issues which arise during the corpus compilation process.

As the corpora under investigation in this book are what we have chosen to call 'special purpose corpora' the latter half of this chapter will be devoted to this type of corpus. To begin with, we will look at the corpus design and text selection criteria used by other researchers working with such corpora. As we find that there is no generally applicable set of design criteria for the compilation of special purpose corpora, we attempt to devise a set of criteria for our own work. Some of these considerations are general design considerations which, we hope, may prove useful for others in the future; others reflect the purpose for which our corpora are intended, i.e. the identification and retrieval of metalanguage statements about terms from text. The final section of this chapter describes how we approached the text selection process and offers a brief description of each of the three corpora or col-

lections of texts which were selected and which form the basis of our analysis in chapters six to nine of this book.

2.2 What is a Corpus?

While the concept of *corpus* appears to be self-evident to corpus linguists, it is a concept which is often misunderstood by others. We propose therefore to look at a number of definitions proposed by different corpus linguists with a view to defining the essential characteristics of *corpus*.

Sinclair (1994a:2) defines *corpus* as ''a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language''. It is interesting that in an earlier publication, he had defined *corpus* as ''a collection of naturally-occurring language text, chosen to characterize a state or variety of a language'' (1991:171).Instead of using the term *text*, he chooses now to use the term 'pieces of language' to describe the components of a corpus; this is because the term *text* can be misleading; it could be interpreted as meaning complete texts whereas the pieces of language selected for a corpus are not always complete texts. The pieces of language selected for inclusion in a corpus are selected according to explicit linguistic criteria; this means that the selection is not arbitrary, and texts must fulfill certain conditions in order to be included. The selected texts are chosen to be used as a sample of the language; they are therefore to be perceived as being representative of the language or some subset of the language, depending on the selection criteria which have been used.

Atkins, Clear and Ostler define corpus as ''a subset of an ETL[1] built according to explicit design criteria for a specific purpose, e.g. the Corpus Révolutionnaire (Bibliothèque Beaubourg, Paris), the Cobuild Corpus, the Longman/Lancaster corpus, the Oxford Pilot Corpus'' (1992:1). In this definition, there is an assumption that the material which is to be selected for inclusion in a corpus is already available in electronic form. We are not convinced that all corpus compilers find themselves in such a fortunate position, particularly those involved in compiling spoken corpora. It is not necessarily true that the compilers of corpora always know in advance what they are going to do with their corpus apart from having a fairly general purpose in mind such as linguistic analysis which might prove to be an umbrella term for a whole range of more specific purposes. Consequently, the notion of a specific purpose is perhaps not essential to the definition of corpus. It might have been more useful to

[1] **Electronic text library**: a collection of electronic texts in standardized format with certain conventions relating to content, etc., but without rigorous selectional constraints.

specify, as Sinclair did, that the corpus is to be used as a sample of the language.

Francis defines corpus as ''a collection of texts assumed to be representative of a given language, dialect, or other subset of language, to be used for linguistic analysis'' (1992:7). Francis' definition, expressed in 1982, would now be considered to be too vague because it is not sufficient to state that texts are 'assumed' to be representative. If representativeness is considered to be an important criterion, then the means of achieving it should be explicit rather than assumed. Like Atkins et al., Francis specifies the purpose for which a corpus will be used (i.e. linguistic analysis). Corpus linguistics has evolved considerably in recent years and is now used for purposes other than linguistic analysis (e.g. as a testbed for natural language processing systems) which means that his definition would require some revision.

McEnery and Wilson define corpus as follows:

> (1) (loosely) any body of text; (2) (most commonly) a body of machine-readable text; (3) (more strictly) a finite collection of machine readable text, sampled to be maximally representative of a language or variety. (1996:177)

These definitions are interesting in that they confirm that 'corpus' is not yet fully defined by the linguistic community. We would suggest that the first and second readings proposed here are too general to be useful but that the third one is closest to what we would consider to be an adequate definition. It incorporates the notions of collection, sampling and representativeness, all of which are important to the description of a corpus.

On the basis of the definitions provided above, there appears to be a consensus that a corpus is an artefact; it is selected, chosen or assembled according to explicit criteria. It is stored in electronic form. It consists of pieces of naturally occurring language. In this context, we understand naturally occurring to mean that the pieces of language have not been tampered with or edited. The corpus may, however, be annotated during or after the compilation process; grammatical tags or SGML markups (e.g. indicating text origin, authorship) may be added to facilitate information retrieval. A corpus may be used as a ''sample of the language'' (Sinclair) or because it is ''representative of a given language'' (Francis). A corpus may be a collection of transcribed spoken and/or written pieces of language, contrary to what the use of the word text might suggest.

2.3 Types of corpora

As corpus linguistics is a relatively new field of enquiry, many new terms have been coined. In addition to corpus we read of subcorpora, components of corpora, special

corpora and specialized corpora, monitor corpora and reference corpora, archives and general corpora, full text corpora, sample corpora, parallel corpora and comparable corpora. Some of these terms have been assigned more than one meaning, others are not yet fully defined. What we propose to do in this section is simply to provide an overview of the meanings ascribed by the various users in the literature and to try and clarify some inconsistencies where they arise.

2.3.1 General reference corpora and Monitor corpora

According to Sinclair, a general reference corpus is:

> not a collection of material from different specialist areas - technical, dialectal, juvenile etc. It is a collection of material which is broadly homogeneous, but which is gathered from a variety of sources so that the individuality of a source is obscured unless the researcher isolates a particular text. (1991:17)

The function of a general reference corpus is:

> to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauruses and other language reference materials. (Sinclair 1994a:12-13)

Within the hierarchy of corpus types, a general reference corpus appears to be the superordinate in the hierarchy, even though it is not representative of all varieties of a language. It is broadly homogenous and designed to be representative of all ''relevant varieties'' of the language and the ''characteristic vocabulary'' of a language. English appears to lead the field in terms of the size of reference corpora available. The Bank of English, over 200 million words, and the British National Corpus, over 100 million words, are described as general reference corpora. Other European countries, participating in the EU funded Parole project, are at present engaged in compiling general reference corpora for their own official languages. In France, for example, INaLF in Nancy is responsible for compiling Frantext, a corpus of the French language, while in Germany, the IDS in Mannheim has been very active in creating resources for the German language.

A monitor corpus is one where texts are ''scanned on a continuing basis, 'filtered' to extract data for a database' but not permanently archived'' (Atkins et al 1992:5). It is:

> a dynamic rather than a static phenomenon, consisting of very large amounts of elec-

> tronically-held text . . . A certain proportion of the data will be stored at any one time, but the bulk will necessarily be discarded after processing. The object will be to 'monitor' such data, from various points of view, in order to record facts about the changing nature of the language. (Sinclair 1987:21)

Developments in computing in the eighties led to the creation of very large corpora (e.g. Bank of English) and made it possible to envisage the design of a monitor corpus which would allow linguists to monitor changes in language use. It was originally envisaged that the monitor corpus would remain the same size with 'old' material being relegated to the archives as new material was added but the concept has evolved in the meantime, with the introduction of the notion of 'rate of flow' whereby all of the material in the corpus is changing constantly and new material simply flows through the corpus at a predetermined rate. However, while the material changes, the composition of the corpus remains ''comparable to its previous and future states'' (Sinclair 1994a:15). Sinclair suggests that the rate of flow for each genre may be adjusted if new sources of data become available and when new procedures enable scarce material such as spoken material to become more plentiful.

2.3.2 Subcorpora, components of corpora, specialized corpora and special corpora

There appears to be some difference of opinion about the scope of the term subcorpus. Atkins et al. define subcorpus as ''a subset of a corpus, either a static component of a complex corpus or a dynamic selection from a corpus during online analysis'' (1992:1). If we have understood Atkins et al. correctly, a subcorpus may be a subset of any type of corpus, whether it is a sample corpus (cf. below), a full text corpus, a monitor corpus or a general reference corpus. The definition does not specify whether a subcorpus must contain the same number of genres as the corpus from which it is drawn, thereby making it a small-scale version of the original corpus, or whether the subset can consist of, for example, just one genre, in which case it is not a small-scale version of the original corpus. Sinclair, who states that corpora can be divided into subcorpora, and that corpora and subcorpora can be divided into components, defines a subcorpus as having ''all the properties of a corpus but happens to be part of a larger corpus'' (1994a:4). Thus, a subcorpus must have all the properties of a larger corpus. We understand this to mean that it is representative of the larger corpus. A component, on the other hand. according to Sinclair, illustrates a particular type of language and is selected ''according to a set of linguistic criteria that serve to characterize its linguistic homogeneity'' (Sinclair 1994a:4). It differs from a subcorpus in that it is not intended to be representative of the corpus from which it is drawn and is therefore not necessarily an adequate sample of a language.

Sinclair uses the term specialized corpora to describe a series of smaller corpora which were designed ''with various purposes in mind'' (1987:16). The first of these was the TEFL corpus, completed in early 1983, and from the description of this particular corpus in Sinclair (1987), it is possible to conclude that the terms specialized corpora and special corpora (defined below) are to be considered as synonyms, with the latter being now the preferred term. In a later publication, Sinclair (1994a) no longer lists specialized corpora in his framework for classification of corpus types and one is inclined to conclude that specialized corpora are now subsumed under the heading special corpus but this is not absolutely clear. On the other hand, it is clearly not possible to classify them under the heading of subcorpus because they are not designed to have all of the properties of a larger corpus. For the moment, therefore, it is best to include them under the heading special corpora, defined by Sinclair as follows:

> those which do not contribute to a description of the ordinary language, either because they contain a high proportion of unusual features, or their origins are not reliable as records of people behaving normally. (1994:7)

Examples of special corpora given by Sinclair (1994a:7) are corpora of the language of children, the language of geriatrics, the language of non-native speakers and the language of very specialized areas of communication. What Sinclair means by the origins of some corpora not being ''reliable as records of people behaving normally'' is that they may have been compiled in artificial situations, e.g. task-based conversations in recording laboratories. Special corpora are to be considered as separate entities from general reference corpora because they contain ''a high proportion of unusual features'' (Sinclair 1994a:7). They are not considered as components in the same way as sublanguages are but the distinction between the two is not clear. Sublanguages can also contain a high proportion of features which are not usual in 'ordinary' communication situations and which some linguists might describe as 'deviant' from the norm, i.e. ''not reliable as records of people behaving normally'' (Sinclair 1994:7). Furthermore, the examples of specialized communication which Sinclair cites for inclusion under the heading of special corpora (i.e. knitting patterns, the heraldic blazon) seem to us to display the same types of differences which one has come to expect in sublanguages. While Sinclair does not specify what the unusual features in special corpora are, it is assumed here that they might include the use of structures which would be considered ungrammatical in 'normal language'. For example, very young children have a tendency to create sentences without verbs. If this is indeed the type of difference which is envisaged, we see no reason not to include sublanguage under the heading special corpus and we would argue that there is no need to create a separate category for it. On the

other hand, there appears to be some justification for excluding corpora such as those constructed in artificial conditions from this category and for creating a separate category for them.

2.3.3 Sample corpora and full text corpora

Early corpora such as the Brown and LOB corpora are now described as sample corpora because these corpora consist of ''a large number (500) of fairly short extracts (2,000 words), giving a total of around one million words'' (Sinclair 1991:23). These were originally described simply as corpora but, with developments in computing and the concomitant changes in corpus size and composition, it became possible to include complete and unabridged texts. Consequently, a distinction is now made between a corpus which comprises extracts (e.g. a sample or samples corpus) and a corpus which contains unabridged texts (a full text corpus).

2.3.4 Parallel and comparable corpora

In the last few years, there has been a growing interest in the use of bi- and multilingual corpora for contrastive studies. Teubert (1996:245) uses the term comparable corpora to describe ''corpora in two or more languages with the same or similar composition''. McEnery and Wilson (1996:57) describe comparable corpora as ''collections of individual monolingual corpora'' which use ''the same or similar sampling procedures and categories for each language but contain completely different texts in several languages''. Peters, Picchi and Biagini (1996:69) also use the term comparable corpora to describe ''sets of texts from pairs or multiples of languages which can be contrasted and compared because of their common features''. A parallel corpus, on the other hand, is a ''bi- or multilingual corpus that contains one set of texts in two or more languages'' (Teubert 1996:245). According to Teubert, a parallel corpus may contain 1) original texts written in language A and their translations into B and C, 2) an equal amount of texts originally written in languages A and B and their respective translations, or 3) only translations of texts into languages A, B and C where the texts were originally written in language Z. Although McEnery and Wilson do not provide any detail on the possible composition of parallel corpora, they too state that they ''actually hold the same texts in more than one language'' (1996:58). Barlow uses the term parallel corpus to refer to ''texts that are translations of each other'' (1996:49), as do Peters et al. who define parallel corpora as ''sets of translationally equivalent texts, in which generally one text is the source text and the other(s) are translations'' (1996:69). McEnery and Wilson point out that other researchers have assigned different meanings to terms such as parallel and comparable corpora:

> some corpus linguists…..refer to what we have termed 'parallel corpora' as *translation corpora* and use the term 'parallel corpora' instead to refer to the other kind of multilingual corpus which does not contain the same texts in different languages. (1996:60)

Given that multilingual corpora have only recently begun to receive the attention of corpus linguists, it is not surprising that there should still be some disagreement about the terminology which is emerging.

2.3.5 Special purpose corpora

In addition to the types of corpora already described, we believe that there is another type of corpus which does not correspond directly to any of those described previously. This is what we choose to call a special purpose corpus, a corpus whose composition is determined by the precise purpose for which it is to be used. While a special purpose corpus may be derived from a general reference corpus or from a monitor corpus it will not constitute a subcorpus in the sense defined by Sinclair because it will not have all of the properties of a larger corpus. Restrictions relating to genre, author, period or other criteria may be imposed depending on the purpose for which the corpus is intended. Nor will it constitute a special corpus because, in special corpora there is an a priori expectation that the language used will deviate from the norm. This is not the case with the language of special purpose corpora. There may be lexical deviations in the sense that some words are used in a precise and specialized way but this hardly constitutes a contravention of normal rules because these terms are not used incorrectly, as they might be in the language of geriatrics, aphasics or children. As the corpora on which we are working do not fit into any of the categories ascribed by others, we have deliberately chosen to coin this new category, i.e. the special purpose corpus. We plan to use this term whenever the specific purpose for which the corpus is to be used (e.g. retrieval of definition statements, analysis of gender-related issues) is the reason for creating or selecting the corpus.

2.4 Approaches to corpus studies

Tognini-Bonelli (1994) distinguishes three different approaches to corpus studies, namely corpus-based, corpus-driven and data-based. The purpose for which the corpus is being used may determine the approach one adopts. The corpus-driven approach:

> constitutes a methodology that uses a corpus beyond the selection of examples to sup-

> port linguistic argument or to validate a theoretical statement….This type of approach is likely to bring the analyst to make new and often unforeseen statements about the language. (1994:1)

The linguist approaches the corpus with an open mind, hoping to validate a hypothesis but expecting to discover new insights in order to refine the hypothesis. The corpus is ''not used just as a repository of examples to back pre-defined theories'' (1994:1). The linguist is equally interested in the exceptions to the theories being validated and uses these as a means of refining the statements which s/he is making about language. The corpus-driven approach is being increasingly adopted by researchers as a basis for refining linguistic theories.

In the corpus-based approach, the corpus ''is used mainly to expound on, or exemplify, existing theories'' (Tognini-Bonelli, 1994:1). The corpus remains primarily a repository ''used to validate existing categories or different applications, to test a tagger or a parser, for example'' (Tognini-Bonelli, 1994:1). The linguist who adopts the corpus-based approach is, to a large extent, not unlike the 'armchair linguist' who invents examples to illustrate a grammatical point, the only difference being that s/he can now use real examples. The intention is less to learn from the corpus than to use it as a means of confirming what one already knows.

The data-driven approach is generally used in the context of language teaching where concordances are used to allow ''students to form their own hypotheses about the regularities and rules of the language''. Its objective is to ''improve the learning rather than to make statements of a more general nature'' (Tognini-Bonelli, 1994:2).

2.5 Corpus users

Atkins et al. (1992) provide a comprehensive overview of potential and existing users of corpora. They believe that corpus users

> can be divided into three types: those interested in the language of the texts, those interested in the content of texts and those interested in the texts themselves as a convenient body of test material for electronic media. (1992:13-15)

These users are described below according to the type of approach which they are likely to adopt.

The corpus-driven approach is likely to be used by lexicographers, terminographers and computational linguists when they are seeking to discover new facts about a language. The Cobuild dictionary is a product of the corpus-driven approach

to lexicography. The meanings of words are identified by means of an analysis of their usage in text. Terminographers may use the corpus-driven approach to identify potential terms in a corpus. Daille (1994) used this approach in her research into term formation patterns in the field of telecommunications. Computational linguists ''separate into two camps, the 'self-organising' and the 'knowledge based'" (Atkins et al. 1992:14). The self-organisers use the corpus-driven approach and attempt ''to use the statistical regularities to be found in mass text as a key to analysing and processing it'' (Atkins et al. 1992:14) while the knowledge-based people tend to use the corpus-based approach. The self-organisers use the corpus ''to train and extend the model'' (McNaught 1993:228). They use the corpus to refine the grammar which they have designed. Sinclair (personal communication 1995) suggests that people who use the corpus-driven approach to reformulate linguistic hypotheses should be described as corpus linguists. The fact that they use corpora as the basis for their hypothesis would certainly justify the creation of this new category of linguist and would allow us to group people who might otherwise be classified separately as grammarians, lexicographers, NLP researchers etc.

Users of the corpus-based approach include theoretical linguists, knowledge based computational linguists and media specialists. Theoretical linguists will use corpora to provide ''a check on the evidence of their own, or their informants' intuitions'' (Atkins et al. 1992:14). The knowledge based computational linguist ''makes extensive use of linguistic knowledge often of a highly theory-based nature'' (McNaught 1993:227). These linguists tend to use corpora as a testbed to assess the linguistic coverage of their model. The same applies to media specialists (i.e. developers of information retrieval, machine translation, or speech processing systems). Although the corpus-based approach is still used by many researchers, particularly if they have previously been using purely theoretical models for their hypotheses, the trend towards the corpus-driven approach is growing, as people realize that the corpus can tell them more than they could ever imagine if they were to rely mainly on their own intuition.

The main users of the data-driven approach are applied linguists who are interested in developing new methods of foreign language teaching.

2.6 Compilation of corpora: design considerations

Prior to compiling a corpus, compilers will have addressed a number of issues which, according to Atkins et al., will include criteria such as size, range of language varieties, the time period covered and ''whether to include writing and speech and the approximate level of encoding detail to be recorded in electronic form'' (1992:2).

Perhaps one of the more important issues which compilers have to address is how big their corpus is going to be. In Sinclair's view (1991:18) a corpus should be ''as large as possible, and should keep on growing''. In Sinclair's (1994a) list of characteristics and default values which corpora are assumed to have, the default value for size is large. Where the value is other than large, the corpus is likely to be a special corpus. If one intends to carry out linguistic studies on language as a whole, it is understandable that one would wish to build as large a corpus as possible. However, if one wishes to carry out linguistic studies on a subset of the language, size may be less important but it will still be important for the corpus to be representative of the subset in question and, consequently, the larger it is, the more representative it is likely to be. Biber (1993) supports this notion that size may not be a major consideration because the adequacy of a corpus depends on the application for which it is intended. What is perhaps more important than the issue of size is the question of representativeness. However, as Biber (1993:256) points out, it is not possible to determine what will be an adequate size at the outset:

> The bottom-line in corpus design, however, is that the parameters of a fully representative corpus cannot be determined at the outset. Rather, corpus work proceeds in a cyclical fashion that can be schematically represented as follows:
>
> Pilot empirical investigation and theoretical analysis'' → Corpus design → Compile portion of corpus → Empirical investigation
>
> ↑ ↵

It is only by proceeding in this cyclical fashion that one can establish whether a corpus is adequate and representative.

A second issue which must be addressed prior to compilation of a corpus is whether the corpus is to contain written and/or spoken transcriptions. As Sinclair states: ''Perhaps the most far-reaching decision is whether the corpus will contain only written texts, or only spoken transcriptions, or both'' (1991:15). As this book is concerned with written text alone, we do not intend to dwell on the problems associated with the collection of authentic spoken material.

Another important issue which may have to be addressed prior to compilation is what period the corpus should cover. Definition of the particular period(s) covered by the corpus may be determined by the purpose for which the corpus is intended. For a general reference corpus which may also be used for etymological and historical studies, the corpus will need to contain material covering many hundreds of years, with each text dated appropriately, while a corpus which is being used for terminological studies (e.g. in the field of computer science) may require that the material be less than ten years old.

2.7 Classification of texts: external and internal criteria

Compilers of corpora require some means of classifying the texts which they have chosen in order to facilitate the retrieval of information from the corpus and the generation of smaller corpora from the main corpus for the purpose of specific corpus studies. Many corpus linguists distinguish between two categories of criteria for the classification of texts in corpora. These categories are 1) external criteria which concern the participants, the communicative function, the occasion and the social setting and 2) internal criteria which concern the recurrence of language patterns within the piece of language (Sinclair 1994a). The distinction is essentially between non-linguistic (i.e. external) criteria and linguistic (i.e. internal) criteria. Atkins et al. make a similar distinction and stress the importance of this distinction for constructing a corpus for linguistic analysis:

> The internal criteria are those which are essentially linguistic . . . . External criteria are those which are essentially *non-linguistic* . . . . A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation. A corpus selected entirely on external criteria would be liable to miss significant variation among texts since its categories are not motivated by textual (but by contextual factors). (1992:5)

Although the criteria will vary from one researcher to another, the distinction remains essentially the same. Increasingly, corpus linguists are advocating an approach to text selection and classification which combines external and internal criteria. However, while there is general agreement on the types of external criteria which are relevant, the definition and relative relevance of many of these vary considerably and are still the subject of much debate. Below, we outline some of the external and internal criteria which are used for text classification purposes.

2.7.1 External criteria

The broad categories of external criteria include genre, mode, origin and aims of the texts and each of these is discussed briefly here. The description which follows focuses on classification criteria for written language alone and draws heavily on the EAGLES Text Typology produced by Sinclair and Ball (1995).

The genre category allows for distinctions to be made between different types of written publications such as books which may subdivide further into fiction and non-fiction, newspapers, magazines, ephemera correspondence, ''typed'' material which includes all types of reports and documentation, and manuscript material

which consists of handwritten texts. Each of these categories may be further subdivided if necessary. There is no single universal system for classifying genre and no set of universally agreed specifications for each particular genre. Consequently, each corpus project tends to have its own method of classifying genre.

Mode is used to describe in what form a text was originally produced i.e. whether it is a transcription of the ''spoken'' word or whether it was originally in written form. Sinclair and Ball (1995) recommend the addition of a third category ''electronic'' to cater for texts transmitted in electronic media because the language used may be different from that used in ''the older established modes'' (1995:7). Electronic texts would include e-mail, discussions in newsgroups etc.

Origin allows compilers to indicate who has been involved in the production of a text. These may include the author, editor, publisher, rights holder, translator and adapter. Compilers may choose to include further information about the originator(s) such as their age, sex, language background and nationality.

The aims of the text include considerations about the target audience and the intended outcome of the text. Audience may include details about audience size and constituency, the latter subdividing into general public, informed lay people, professional people, specialists, students and trainees. It may be considered useful to specify the relationship between the author and reader, whether distant, neutral or personal. The intended outcome is the purpose for which a text is written and includes the following categories: information, discussion, recommendation, recreation which includes fiction and non-fiction, instruction which includes academic works, textbooks and practical books.

2.7.2 Internal criteria

In the past, internal criteria received much less attention than external criteria but they have started to attract greater interest in recent years, largely because of innovations proposed by researchers such as Biber and Phillips. However, they are still difficult to apply because of the lack of appropriate tools. Sinclair and Ball (1995:15) suggest that ''two central parameters of the classification of texts are better described using internal, or text-linguistic, rather than external, or sociocultural criteria''. These two parameters are topic and style. Previously, both topic and style had been assigned on the basis of external criteria but Sinclair and Ball make a very cogent argument for assigning these using internal criteria.

2.7.2.1 Topic

Topic, as Sinclair and Ball state, ''is one of the central controversial areas of text typology'' (1995:3). It is also considered to be a very important criterion in the

classification of texts in corpora. As Sinclair and Ball point out, there are almost as many means of classifying topic as there are corpora. Each corpus views the world differently, with its texts classified according to a system devised by the compilers of the corpus or using existing classification systems such as Dewey or UDC. Topic may be identified by looking at what a particular text is about (e.g. on the basis of its title, table of contents in the case of a book) and classifying the text accordingly. However, to classify texts in this way is to ignore the fact that texts may deal with more topics than the one specified in the title or indeed in the table of contents. Sinclair and Ball reject this approach to text classification, suggesting that it is grossly oversimplified, and advocate the following instead:

> No existing external classification seems to be satisfactory. We recommend that it is classified principally as an internal matter, to do with things like the vocabulary choices in a text, rather than an external matter, where the Universe is endlessly chopped up into subcategories . . . . In the classification of topic, the internal evidence is primary. (Sinclair and Ball 1995:3)

They recommend that the internal evidence, ''such as the vocabulary clustering, is developed first of all, and the external evidence is added at a stage of greater detail'' (1995:3). They offer a detailed description of Phillips work on the 'aboutness' of text. Phillips (1983) suggested that the topic of a text could be identified by examining the lexical structure of a text and identifying keywords used frequently in the text. This type of approach is already being used in some abstracting and information retrieval techniques. However, Sinclair and Ball concede that ''It is likely to be some years before automatic methods of topic assignment are devised, tested on a sufficient variety of data in many languages, and agreed by a body such as EAGLES'' (1995:21). Consequently, they recommend that, in the interim, the corpus community should agree to use only a very broad indication of topic ''to avoid wasted and needless effort'' (1995:21). They suggest the following list of topics which ''should be varied and extended to suit the researchers' priorities'': the life of the mind, culture, the physical world, living things, society, manufacture, communications (1995:21). It remains to be seen whether the corpus community considers this broad approach to be useful and whether it is prepared to adopt this recommendation which is a major departure from the previous practice of using quite fine-grained classifications. In our opinion, the topics proposed appear to be almost too broad to be useful; if the reasons for such a coarse-grained approach are lack of consensus on topic classification systems and methods of assigning topic, we would suggest that a better interim solution might be not to assign any topic at all.

2.7.2.2 Style

> Style is a notorious term, because it is used in so many different ways by researchers from several disciplines, and has popular meanings as well. It is used here to mean the way texts are differentiated other than by topic. (Sinclair and Ball 1995:22)

Hitherto, the corpus community has used categories such as formal, informal or colloquial to classify text style but, as Sinclair and Ball point out, ''there are no institutionalised schemata'' (1995:22) for these categories. One person's formal may be another's informal and what may be considered formal in speech might be considered to be informal in written text. If no a priori conditions exist for distinguishing between these categories, the results are likely to be at best confusing, if not unhelpful. There is a tendency to assign a style category on the basis of genre and text purpose. Thus, a report is likely to be classified as formal and a discussion may be classified as informal or formal. How does one decide which category is appropriate? Some might argue that the context will be of some assistance but, ultimately, the decision rests on intuition rather than explicit criteria. Biber (1993) has demonstrated that genre is not necessarily a useful means of classifying style as different genres may have similar styles, i.e. certain text types within one genre may share similar linguistic structures with text types in another genre. He advocates using internal linguistic criteria as a means of determining style or text type. In a study described in *Variation across speech and writing* (1988), he identified a number of linguistic features of text which he then used in a cluster analysis which allowed him to classify texts according to type. He concedes that this type of analysis can only take place after a corpus has been compiled and that ''in defining the population for a corpus, register/genre distinctions take precedence over text type distinctions'' (1993:244-245). This is because registers are based on external criteria which can be identified before a corpus has been compiled whereas ''there is no a priori way to identify linguistically defined types'' (1993:245). It would seem, therefore, that style will continue to be judged initially by external criteria until such time as some *a priori* means for identifying linguistically defined types is found.

2.8 Observations

In the corpus compilation process, the emphasis tends, in general, to be on external criteria, both for the classification of texts and for the design of corpora. This is not surprising as internal criteria can only be defined once general decisions have been reached about the nature and purpose of a corpus. The EAGLES recommendations

for further research into automatic classification using internal criteria are likely to lead to a gradual shift towards greater emphasis on the use of internal criteria.

2.9 Overview of design considerations in the compilation of special purpose corpora

In much of the literature available on the subject of special purpose corpora, authors simply document the corpus compilation process and offer little discussion of the design criteria used. Given that it is still quite difficult to source texts for special purpose corpora, authors would perhaps not have been justified in devising a lengthy list of criteria which they might then have been unable to apply. Relatively little has been written about the design of special purpose corpora in general; Roe (1977), Yang (1986), Fang (1991), Flowerdew (1993), James et al. (1994) and Gledhill (1996) are among the few authors to our knowledge to have written about this subject.

2.9.1 Corpus size

A corpus consisting of ca. one million words is the size usually selected for special purpose corpora, with the justification for this varying from intuition:

> for an investigation into the text of a restricted and relatively clearly defined subject area, a corpus of around a million words, but consisting of a small number of large samples would be appropriate. (Roe 1977:21)

to:

> Following the practice of major corpora…. we decided that our corpus should contain one million words of running texts . . . one million words represent a reasonably large proportion of the finite subset of the language under study. (Fang 1991:74)

Fang's decision on corpus size was based on the size of corpora used in other projects; he chose to follow the practice of Brown and LOB and opted to collect one million words of running texts for his corpus. Flowerdew (1993) believes that a smaller corpus can suffice ''where a course is designed for a particular specific purpose a much smaller corpus of language, drawn from the given specific purpose area, is more appropriate'' (1993:232).

The corpus compiled by Gledhill, for example, consists of only 500,000 words

but it is ''a corpus that is highly specific to one discourse community, one genre and one topic'' (Gledhill 1996:110). This notion that a special purpose corpus does not need to be as large as other more general corpora is echoed by other compilers of special purpose corpora (e.g. James et al. 1994). In reality, however, the rationale for using a smaller corpus tends to be related less to a decision to keep it small than to the availability of material and copyright considerations. Compilers of such corpora hope, rather than know for certain, that the phenomena which they are investigating will appear with sufficient frequency in their smaller corpora to give them adequate results. It is likely that the best practice, even for the compilation of a special purpose corpus, is to aim to compile as large a corpus as possible and to ensure that, as Sinclair suggests, the corpus keeps on growing.

2.9.2 Topic

In publications (e.g. Roe, Fang, Flowerdew, James et al., Gledhill) describing the corpus compilation process, topic generally appears at the top of the list of design criteria. Each of the authors decided to focus on a particular subject field, ranging from the broad subject of science, in the case of Roe, to computer science, in the case of Fang and James et al. and research on cancer treatment in the case of Gledhill. With the exception of Gledhill who selected texts by asking researchers in the area of cancer treatment to ''submit their own articles and also to recommend journals and even specific papers which they would consider relevant to their own research''(1996:110), all of the other authors cited made their own text selection. The texts selected for the chosen topic were deemed to be appropriate if they formed part of the third level syllabus for the topic in question. If the texts had been recommended by academics working in the subject in question, they were assumed to deal with that particular field. No investigation was made by any of the authors beforehand to establish whether the texts did indeed deal with the chosen topics although Roe and James et al. demonstrate through frequency counts that the vocabulary of their texts reflects the vocabulary of the field. This may mean that in special purpose corpora, the title and/or table of contents alone will indicate what the topic is and there may be no need for the type of vocabulary clustering analysis proposed by Sinclair and Ball (1995).

2.9.3 Genre

With the exception of Gledhill who focused on research articles, research in this area has focused mainly on academic textbooks (Roe, Fang, James et al., Flowerdew) and lectures in third level institutions (Flowerdew).

2.10 Proposals for design criteria for the design of special purpose corpora

We had originally assumed that corpus design and text classification criteria for special purpose corpora would be well documented in the literature but, as we have noted, this type of information is still fairly scarce. Consequently, we found it necessary to devise a set of criteria for selecting texts which would be suitable for the type of analysis which we were proposing to undertake. Readers will note that we have borrowed heavily from the list of text attributes provided by Atkins, Clear and Ostler (1992) in their discussion of corpus design criteria. What we have done is to select those attributes which we believe to be essential for our work and defined each of the attributes in terms of our special purpose. These attributes have been refined in the course of our research, testimony to Biber's assertion that ''corpus work proceeds in a cyclical fashion'' (1993:256). We arrived at our set of attributes in the following manner.

We initially selected a range of different texts which we thought might be useful for our purposes (cf. Section 2.11 for more detailed discussion of the text collection process). We then proceeded with a pilot linguistic analysis of all of the texts. We found that some of the texts were unsuitable. They were unsuitable because the language used was often quite informal and did not have the rigour required for the type of metalanguage statements which we were seeking. When such texts did contain metalanguage statements, the statements were frequently modified by the use of hedges or modals. Interestingly, the texts which we ultimately rejected as being unsuitable for the retrieval of definition statements had also proved to be unsuitable as reliable sources of terms. Gradually, our initial ideas about which text attributes might prove useful evolved and crystallized into a set of criteria for corpus design and text selection. It is important to stress that it was only through carrying out a pilot linguistic analysis of the texts that we arrived at the set of attributes described below. If further texts which meet the criteria outlined below are to be considered for inclusion, a linguistic analysis will also be necessary in order to assess their suitability.

2.10.1 Size

In some respects, the issue of corpus size is problematic when it comes to the compilation of special purpose corpora and in other respects it is irrelevant. On the one hand, if the outside world (i.e. general corpus linguists) deems that a corpus is too small to be representative, a linguist working with special purpose corpora runs the risk of having his/her work ignored. This of course raises the issue of representativeness. If a one million word corpus is deemed to be a representative subset of

the subject area and text type under investigation (cf. Section 2.9.1), perhaps it is sufficient but the question of how one determines the size of a representative subset is another unresolved issue. On the other hand, the issue of size may be irrelevant, and by that we do not mean unimportant, in the sense that corpus size may be dictated by the amount of material which is already available in electronic form or which has to be converted to electronic form.

We are not convinced that there is any justification for setting an upper limit on the size of a corpus and we have therefore set no upper limit on the size of our special purpose corpus. Any text which meets the suitability criteria should be eligible for inclusion in the corpus. Researchers may, at some stage, have to specify an upper limit but this will only be because there are hardware and software constraints which oblige them to do so.

2.10.2 Written text

As our own work is concerned exclusively with the analysis of written texts, all texts in the corpus must be drawn from written sources. All texts must be full texts. In other words, if books are being included, and the entire book has been written by the same author(s), the corpus must contain the full book. Similarly, if research articles have been selected, the entire research article must be included. The reason for specifying that texts must be full texts is that we are interested in retrieving as much definition information as possible. When a term is introduced for the first time in a text, it is occasionally explained. It could be argued that such explanations are more likely to be provided in the early sections of a publication and that therefore it should be sufficient to ensure that when samples are selected, they should include the introductory chapters. We have found, however, that explanations are very often provided throughout a publication. For example, in an introductory textbook, simple terms may be explained in the early chapters and more complex terms introduced and explained in the later chapters. If we select the entire publication for inclusion in the corpus, we may find that we have explanations for most of the more important terms in a given conceptual framework. If we were to select only the introductory chapters, we would miss out on explanations of more complex terms. If we were to select only the later chapters, we might find that we were unable to understand some of the terms used in the explanations. Yet it is very likely that these very terms will already have been explained in earlier sections of the publication.

2.10.3 Published

All texts must have been published. We are using Biber's (1993) operational definition of published texts for our selection purposes, namely:

(1) they are printed in multiple copies for distribution;
(2) they are copyright registered or recorded by a major indexing service. (1993:245)

Published texts is an umbrella term which may include books, reports, standards, manuals. The reason why we have chosen to stipulate that texts must have been published is that we believe that the status of 'published' combined with certain other factors such as the status of the author will validate the reliability of the material as a potential source of definition information.

2.10.4 Text origin

The text may be a 'single' text, i.e. the product of one individual, or a 'joint' text, i.e. the product of a collaborative venture where separate sections are not attributed to different individuals. Thus, the text may have been produced by an institution, as would be the case, for example, with standards.

2.10.5 Constitution

The text may be single or composite in the senses defined by Atkins et al., namely, ''one integral text by one author is single; a newspaper journal, collection of essays, or textbook made up of many distinct small texts which could each be classified individually is composite'' (1992:7). We specified under 'Written Text' above that, in the case of entire books written by the same author(s), the entire book was deemed to be a full text and had to be included. In the case of books or other eligible publications where different sections are written by different people and each of these sections could be classified individually, each of these sections is deemed to be a full text, and the entire publication is a composite text.

2.10.6 Author

The author(s) must be an acknowledged individual or institution. By 'institution', we mean a body such as a standards institute, an academy or institute of experts. By 'acknowledged', we mean that the authors must be recognized by their peers as having the level of expertise required to write about the particular subject. In other words, they must be qualified to speak about the subject. Their educational and/or professional background must be in the discipline about which they are writing. This factor, coupled with the fact that a text must be published in order to qualify for inclusion should ensure that only 'acknowledged' authors will be eligible for consideration.

2.10.7 Factuality

The texts must be factual. They must purport to represent what is known to exist, or believed to exist. We realize that this one may be difficult to measure but believe that by combining this criterion with others such as author, intended outcome, audience and setting it should be possible to isolate the factual from the non-factual.

2.10.8 Technicality

This attribute is ''based on the degree of specialist/technical knowledge of the author and target readership/audience'' (Atkins et al. 1992:8). The text may be technical (written by specialists for specialists) or semi-technical (written by specialists for a specific target audience).

2.10.9 Audience

The audience is the target reader, the person for whom the author is writing. The audience may 1) have the same level of expertise as the author or 2) have a lower level of expertise than the author but have an interest or need to learn more about a subject. The audience may be second or third level students of the particular discipline, or they may be professionals working in the discipline.

2.10.10 Intended outcome

The intended outcome of the texts must be informative, didactic or stipulative. A didactic text, as the name suggests, may be a text used in the teaching of a subject. A stipulative text may be a standard or regulatory text which prescribes and defines the terms used in a particular subject domain.

2.10.11 Setting

The setting is institutional or academic/educational. This means that the texts are destined for use in a corporate or institutional context, or in a broad academic/educational context. For our purposes, the setting must correspond to one of the three communicative settings selected in chapter one, i.e. communication a) between experts, b) between experts and initiates and c) between teachers and pupils. As special interest columns in newspapers and magazines of general interest correspond to the setting of communication between relative experts and the uninitiated they do not qualify for consideration.

2.10.12 Topic

The text must purport to be about a specific subject domain but, from our analysis, it would appear that no particular subject field need be specified. It should be possible to extract the same type of information (i.e. metalanguage statements) from texts dealing with very different topics, as long as all of the other criteria have been met. However, if we are interested in isolating statements about terms from any one domain, we will need to be able to identify those texts within the corpus which refer to the domain in question. This might involve assigning a topic label to each of the texts included in the corpus, thus enabling us to focus on one domain at a time. The idea that topic may not be a relevant criterion is in contrast with the approaches adopted by previous researchers of special purpose corpora (i.e. Fang, Roe) who identified a topic and then proceeded to work within the confines of that topic.

2.11 The search for suitable texts

When we started to look for texts to include in our special purpose corpus, we were looking for specialised texts. We were using the term specialised texts to describe texts with a high density of terms. As our search continued, we realized that term density was in fact much less important than other factors. Hence the gradual emergence of a set of criteria which allowed us to select only those texts which were likely to contain metalanguage statements which could be used as input for the formulation of definitions. These are the criteria described in Section 2.10 above.

Prior to commencing our search for suitable texts, we had hoped that we might be able to obtain permission to use existing special purpose corpora. Given the growing popularity of corpus linguistics and the fact that many linguistic research and/or information retrieval research groups seemed to be using corpora as a resource and/or testbed for their theories and applications, we believed that gaining access to a suitable corpus would be relatively straightforward. Already in the 1980's, linguists investigating the nature of sublanguages and developing tools to cope with these sublanguages, had demonstrated the basis of their theories using 'real' text as input. These researchers included, in particular, Naomi Sager (1982) working with the Linguistic String Project at New York University who used hospital discharge summaries and clinic visit reports in several areas of medicine as input for her research; Lynette Hirschman (1982) working with the R&D Division of SDC, an American company, who used a corpus of pharmacology articles on digitalis and a corpus of follow-up radiology reports on 19 cancer patients. Researchers at the University of Montreal were using weather bulletins as input for the develop-

ment of their machine translation system, TAUM-METEO. Information retrieval researchers were using rather larger corpora (e.g. financial reports in the Wall Street Journal). By the early nineties, many researchers were convinced of the need for large corpora of real text and, consequently, several general reference corpora have since been compiled both in Europe and elsewhere. Some of these corpora run into millions of words. The larger corpora tend to consist of a combination of spoken and written material drawn from a range of sources.

As the Cobuild unit at the University of Birmingham agreed to allow us access to their corpus, we decided to start by searching the Cobuild corpus for suitable material. We identified one publication in the corpus, the *New Scientist*, which was initially of some interest to us. The *New Scientist* is a fairly specialised periodical aimed at the lay reader. As this was the only collections of texts in the corpus which corresponded to our, still very general, requirements, we began to search further afield for access to a specialized corpus.

We were aware that the European Commission was about to finance a number of corpus-based projects, and was already co-funding the production of a multilingual CD-ROM. We contacted them to enquire about the possibility of access and were told that we would have to await publication of the CD-ROM. The CD-ROM became available twelve months later. The CD-ROM, known as the *European Corpus Initiatve Multilingual Corpus* contained a range of collections of texts in different languages. These included a 4.7 million word collection of texts published by the International Telecommunications Union (ITU) and available in three languages, English, French and Spanish. This particular collection of texts also appeared to correspond to our requirements. As we felt that our analysis would be more useful if we used a wider range of texts than those covered by just two collections of texts (i.e. the *New Scientist* collection and the ITU collection), we decided to continue our search.

We again contacted the Commission twelve months after our initial contact in the hope that it might be possible to obtain, for research purposes, some of the material which had become available since our previous contact. Not so. While the Commission was publicly committed to compiling large corpora and making them freely available, it was not, perhaps understandably, in a position to make provisional results available pending the release of the finished corpora. We then contacted some compilers of existing special purpose corpora but copyright restrictions prevented them from making the material available to other researchers.

Consequently, in early 1994, we decided to search for texts ourselves. We did this in two ways. We spent many hours trawling the Internet and just as much time in the library looking for suitable material. Our notions of what we wanted were still quite vague. We were not interested in fiction; we were looking for complete texts preferably dealing with technology, or a technical or scientific subject. We subse-

quently narrowed down our search on the Internet and began to search for texts on information technology; we identified some publications which we thought might be suitable and downloaded the following texts in order to carry out a small scale linguistic analysis: *Guidebook to Network Resource Tools* (25,000 words), *Zen and the Art of the Internet: A Beginner's Guide to the Internet* (38,000 words). We also narrowed down our search in the library and eventually decided that it would be interesting to look at how journalists present a specialised topic such as information technology in dedicated special interest columns which appear weekly in newspapers such as the *Guardian* (English daily)and *Irish Times* (Irish daily). We had access to two further collections of texts: the GCSE corpus held by the Cobuild unit at the University of Birmingham and a collection of articles on plant biology and transportation kindly made available by Tim Johns of the University of Birmingham.

We started to look more closely at the collections of texts which we now had at our disposal and very gradually began to develop a set of criteria which would allow us to select only those which seemed likely to be most useful for our purposes. Thus, we rejected special interest articles in newspapers; while these collections of texts met some of our criteria (e.g. intended outcome, published, author) they failed on the grounds of technicality, audience and setting. We also decided against investigating further the *New Scientist* and the publications downloaded from the Internet, not because they were unsuitable but because three other stronger candidates had emerged. These were the ITU, GCSE and NATURE collections of texts described in detail below. The reason why we chose these three collections of texts was they not only met all of the criteria which we specified in Section 2.10 above but each of them corresponded to a different communicative setting. These settings were communication between experts, communication by experts to initiates and teacher-pupil communication We thought it would be interesting to investigate the use of metalanguage patterns in these three different settings.

2.11.1 The ITU corpus

The ITU corpus contains the International Telecommunications Union CCITT Handbook "The Blue Book". It was made available on CD-ROM in 1993 by the European Corpus Initiative (ECI) and was provided to the ECI by the International Telecommunications Union in Geneva, Switzerland. The original text is copyright International Telecommunications Union 1988. The corpus consists of 4.7 million words. In terms of communicative setting, it corresponds to our second category, namely communication between experts and initiates. The text was produced for ITU members who, it is assumed, already have a certain level of expertise in the field of telecommunications.

2.11.2 The GCSE corpus

The GCSE corpus was compiled by the Cobuild unit at the University of Birmingham. It consists of one million words and comprises a number of textbooks on the GCSE syllabus. It proved difficult to obtain precise information about the corpus mainly because there does not appear to be any published material about its design. We are very grateful to Tim Lane of Cobuild who did his utmost to assist us in our search for more precise information. We managed to ascertain that the corpus consists of thirteen textbooks on subjects such as history, geography, biology, chemistry, sociology and politics. The titles of the books, and author information when available, are listed at the end of this chapter. We do not have any information about the length of each of the textbooks and, in some cases, we have no information about the author or editor of the books. However, the corpus as a whole meets the criteria specified for the compilation of our special purpose corpus. In particular, the setting corresponds to our fourth communicative setting, namely teacher-pupil communication.

2.11.3 The Nature corpus

The Nature corpus was kindly made available to us by Tim Johns at the University of Birmingham. It consists of a collection of articles from the journal *Nature* which were published in 1989. The corpus consists of 230,000 words which makes it the smallest of the three corpora. It meets all of the criteria specified and, in particular, the setting corresponds to our first communicative setting, namely expert-expert communication.

The three corpora described here vary considerably in size consisting of 4.7 million words, 1 million words and just 230,000 words respectively. However, as our analysis in later chapters will show, this discrepancy in size does not have a significant bearing on our results. Each of the corpora corresponds to a different communicative setting and it is the nature of the communicative setting rather than the size of the corpus which will determine how much definitional information authors will provide and how they will express such information.

2.12 Conclusion

In this chapter, we have provided an overview of the various types of corpora which have been identified by researchers in the field. An additional category, namely the special purpose corpus, has been proposed. The different approaches that are used in corpus research have been summarized, as have the users of each of these ap-

proaches. With a view to devising a set of design attributes for our own purposes, we investigated which corpus design and text classification criteria had been used by compilers of general reference corpora. We also investigated how special purpose corpus compilers had compiled their corpora. As very little documentation was available on this subject, we devised our own set of attributes for selecting the types of texts which would suit our special purpose corpus. In applying these attributes, we noted that many texts which were likely to be suitable as sources for identifying terms functioning as terms rather than as part of general vocabulary were also likely to be suitable sources for the types of definition statements which we were hoping to retrieve. Using the attributes which we had selected, we chose, from the texts at our disposal, three collections of texts or corpora for further analysis. These three corpora will form the basis of our analysis in the final four chapters of this book.

List of Books in the GCSE corpus

Bushell, J., Nicholson, P. Biology Alive.
Cadogan, A. Green, N. *Biology*.
Hill, G. *Chemistry Counts*.
Leake, A. *Action Economics, A coursebook for GCSE*.
Book 3, The Changing World, A Sense of Place (no information on author, editor)
Hart, C. (Ed.) *Worldwise Issues in Geography*.
Making Modern Britain, British Social and Economic History from the 18th Century to the Present Day (no information on author, editor)
A Handbook of Modern History, World History since 1870 (no information on author, editor)
Gibbons, S.R. *A Handbook of Modern History*.
People and Politics in Britain (no information on author, editor)
Dobson, K. *Co-ordinated Science, GCSE Introductory Book*
Bishop, K. *Science for Life*.
Andrews, J. *Understanding Sociology*.

3 Dictionaries and defining strategies

> Dictionaries do not exist to define, but to help people grasp meanings.
> (Bolinger 1965:572)

> It is a commonplace that if dictionary definitions could be read as stating necessary conditions, any dictionary definition describing a dog as a four-legged animal would make a three-legged dog a logical impossibility. This is not of course an argument for weakening dictionary definitions; it is an argument for reading them as explanations stating what is normally the case rather than what is necessarily the case.
> (Hanks 1987:118)

3.1 Introduction

The language dictionary is the reference source that people are most likely to use when they wish to find out what a word means. There are many different types of language dictionary designed for different types of user. Broadly speaking these divide into general language dictionaries in one or more languages and specialised language dictionaries in one or more languages. Users will select the dictionary which corresponds to their particular needs. In this chapter we propose to provide a brief overview of different types of language dictionaries and a description of the sort of information which dictionary entries in these dictionaries are likely to contain. We examine the process of producing definitions and look at three different approaches to the production of definitions for general language dictionaries: the 'traditional' approach, the Cobuild approach and the approach proposed by Mel'cuk and others. We compare and assess the 'traditional' and Cobuild approaches to the formulation of dictionary definitions. In the final section of this chapter, we examine ISO recommendations for the formulation of definitions for terms and make some tentative recommendations for good defining practice. It may seem strange to some readers that we should choose to look at principles for general language lexicography in a book where the main focus is on terminology. However, we have good reason to do so. While terminography and general language lexicography generally operate as two separate disciplines, there are principles which are applied in each of these disciplines which could usefully be adopted in the other. In the case of

definitions, general language lexicographers could, for example, benefit from the very strict approach adopted by terminologists particularly in relation to the naming of superordinates, and terminologists may have something to learn from certain general lexicographic principles in relation to the phrasing of definitions. What we are attempting to do is to devise a method of formulating terminological definitions which will be useful and comprehensible to specific groups of target users.

3.2 Language dictionaries

We distinguish between four different types of language dictionary: the monolingual general language dictionary, the bilingual general language dictionary, the monolingual specialized dictionary and bi- and multilingual specialized dictionaries. Each of these is described in terms of its general purpose and the content of the dictionary entries.

3.2.1 Monolingual general language dictionaries

Here, we are concerned exclusively with conventional language dictionaries which deal with the meaning and usage of words. Dictionaries of, for example, etymology, place names, idioms are excluded from this discussion. The monolingual general language dictionary is perhaps the best known and most commonly consulted of the dictionary types. As Zgusta states:

> The rationale on which this category is founded is the circumstance that these dictionaries are concerned mainly with the general language (as opposed to the different restricted ones), i.e. with the standard national language as generally used. (Zgusta 1971:210)

In terms of purpose, monolingual general language dictionaries can be broadly divided into native speakers' dictionaries and learners' dictionaries. Native speakers' dictionaries tend to provide more comprehensive coverage of a language than learners' dictionaries where the emphasis is on covering the most common words of a language. Entries in monolingual general language dictionaries are generally organized alphabetically, and the headwords of a particular entry may consist of single words or multiword units. The entry may be subdivided to accommodate one or more reading distinctions which may relate to the headword alone or to phrases containing the headword. The entry may include etymological information, a phonetic description, an indication of the grammatical category, a definition of each of the readings in the entry, examples illustrating usage. Users consult these dictionar-

ies for the purpose of ascertaining the meaning of a word and/or its pronunciation and usage or for confirmation of what they already know about a word. Depending on the size and scope of the dictionary, coverage will range from the words most commonly used in a language (e.g. *Collins Cobuild English Language Dictionary*) to comprehensive coverage of a particular language (e.g. *Oxford English Dictionary* (OED)); the former is a learners' dictionary and the latter a native speakers' dictionary. In addition to providing definitions of what are commonly described as general language words, all general language dictionaries will include definitions of technical terms. A learners' dictionary such as the *Collins Cobuild English Language Dictionary* will include some entries which define technical terms while a comprehensive native speakers' dictionary such as the OED will include a large number. Landau (1989:33) estimates that ''over 40 per cent of the entries in an unabridged dictionary are scientific or technical and that in college and desk-sized dictionaries the percentage is 25 to 35 per cent''. However, in general language dictionaries, the manner in which technical terms are defined will differ from the manner in which they are defined in monolingual specialized dictionaries. There are two reasons for this, expressed very clearly by Sinclair in the introduction to the first edition of the *Collins Cobuild English Language Dictionary*:

> The meanings given are the meanings that are actually used in our ordinary texts and not necessarily what a specialist would say. (1987: xix)

and

> Hence we have explained the technical words according to the way we use them in ordinary English, and we have kept the explanations as simple as possible. (1987: xx).

Definitions of technical terms in general language dictionaries are expressed simply and not necessarily in the manner in which a subject expert would express them. As noted in chapter one, the terminological status of a lexical item will depend on the communicative setting in which it is used. It would seem from the above that when technical terms are defined in general language dictionaries, the definition refers to what they mean in general language rather than to what they might mean to a subject expert within a specialized communicative setting.

3.2.2 Bilingual general language dictionaries

As Zgusta (1971:294) states: ''The basic purpose of a bilingual dictionary is to co-ordinate with the lexical units of one language those lexical units of another language which are equivalent in their lexical meaning''. The bilingual general

language dictionary is designed for use by people wishing to identify an equivalent for a word or phrase in another language. This type of dictionary is frequently bi-directional (e.g. English-French, French-English) and is generally organized alphabetically. The headword usually consists of a single word, and the entry may be subdivided to accommodate multiple readings of the head word or variants of the head word in the form of phrases. The phonetic description and grammatical category of the headword are generally provided, as are examples of usage in the target language. However, bilingual dictionaries do not usually provide definitions. It is assumed, in our experience erroneously, that the meaning of the word is already known to the user and that, if it is not known, the user will consult a monolingual general language dictionary for this purpose. It is not our intention to dwell further on this issue here except to say that, by failing to provide definitions, bilingual dictionaries can cause great frustration and result in, occasionally very amusing, mistranslations by their users. Meyer suggests that the weaknesses of bilingual dictionaries are linked to four related problems which are present to varying degrees in such dictionaries. These are: ''1) an attempt to serve L1-L2 and L2-L1 users simultaneously; 2) a lack of comprehensiveness; 3) inadequate formalization of lexicographic principles; and 4) insufficient illustration of matching complexities between SL and TL items'' (1990:177).

A recent development in the compilation of bilingual dictionaries which is worthy of mention here is the Bridge series of dictionaries, devised by the Cobuild Unit at the University of Birmingham. A Bridge dictionary is a cross between a monolingual dictionary and a bilingual dictionary. The headword and examples of usage are in English and are essentially the same as in the Cobuild monolingual dictionary (cf. Section 3.3.2) but the definition is translated into the native language of the user in order to facilitate comprehension. The dictionary is appropriately named because it bridges the gap between the bilingual and the monolingual dictionary but should be considered as complementing rather than replacing either of these.

3.2.3 Monolingual specialized dictionaries

> the 'technical' dictionary is compiled on the basis of criteria provided by particular target groups and their professional or special-interest needs. Neither its content nor its structure and methods are exclusively, or even predominantly, determined by purely linguistic considerations. (Opitz 1983:163)

Monolingual specialized dictionaries subdivide into general specialized dictionaries and subject-specific specialized dictionaries and, as Opitz correctly points out, they are not usually governed by linguistic considerations. Their main purpose is to define technical terms, either of a broad range of subject fields in the case of general

specialized dictionaries (e.g. *Longman Dictionary of Scientific Usage* 1979) or of a circumscribed subject field in the case of subject-specific specialized dictionaries (e.g. *Astronomy: A Dictionary of Space and the Universe* 1977). The organization of both types of specialized dictionary and the content of their entries are broadly similar. The headword of each entry consists of a single word or multiword unit. Each entry contains a definition which may contain cross-references to other entries. A general specialized dictionary may contain multiple readings of a single headword where the headword has different meanings in more than one domain. In the event of multiple readings, each definition specifies to which domain it refers. A subject-specific specialized dictionary will only deal with instances of polysemy if the headword is polysemous within the domain described by the dictionary. Entries are unlikely to contain the phonetic description of the headword or any examples of usage. Many specialized dictionaries do not provide the grammatical category of the entry or any information relating to whether the entry is transitive, in the case of verbs, or whether it can be used in the plural, in the case of nouns. The purpose of these dictionaries is to clarify meaning rather than to specify usage. The main drawback of these dictionaries is that they do not contain grammatical or phraseological information which means that the dictionary users are given no advice regarding, for example, collocation restrictions on technical terms. Ideally, a specialized dictionary should contain not only all of the information which is already being provided in such dictionaries but also grammatical and phraseological information which would show the user how to use the term correctly.

3.2.4 Bi- and multilingual specialized dictionaries

As with monolingual specialized dictionaries, bi- and multilingual specialized dictionaries subdivide into general bi- and multilingual specialized dictionaries and subject-specific bi- and multilingual specialized dictionaries. They tend to be organized alphabetically. In bilingual specialized dictionaries, the headword of the entry generally consists of a single word, below which are listed multi-word variants of the term. Equivalents for each of these in the other language are provided, and frequently more than one equivalent is provided, leaving the user to guess which one is appropriate. Phraseological information is occasionally provided, giving the user some idea of common collocates for the term. The grammatical category is specified but phonetic descriptions are not. Nor is it usual to provide definitions. It is assumed that the user already knows and understands the term in his/her own language. We would suggest, however, that where the user is consulting the dictionary to find the equivalent of a foreign language term in his/her mother tongue, a definition is crucial for identifying the appropriate equivalent.

Multilingual specialized dictionaries (e.g. the Elsevier series of technical dictio-

naries, glossaries compiled by the Commission of the European Communities) are even less informative than bilingual specialized dictionaries. The dictionary pages are generally laid out in vertical columns with a different column assigned to each language. Each entry consists of a head word or phrase, with, along the same line across the page, its equivalent in each of the languages. Only a very limited number of multilingual specialized dictionaries provide definitions, again generally only in one language. Zgusta, in his discussion of multilingual dictionaries, suggests that:

> the only domain in which multilingual, more-than-bilingual dictionaries have a justification is the field of technical terminology. The meaning of technical terms is usually much more precisely defined than that of a general word, so that semantic equivalence can be established more accurately. . . . But even in this field, the difficulties are great and the false friends . . . more numerous than one would casually assume. (Zgusta 1971:214)

We would refute Zgusta's suggestion that the use of multilingual dictionaries is justified for technical terminology and we would even suggest that he himself is not convinced by the arguments which he puts forward. His suggestion that technical terms are ''usually much more precisely defined'' is already an indication that terms are not always precisely defined which means, in our view, that there is even greater justification for providing definitions. We would suggest that it is always useful to provide a definition, even in situations when terms are precisely defined and a direct one-to-one correspondence exists between terms in two languages. Second, his suggestion that equivalence can be established ''more accurately'' and not simply accurately confirms that the situation is a lot more complex than ''one would casually assume'', as he finally concedes.

3.3 Lexicographic methods

Here, we propose to provide a brief overview of three different approaches to the production of dictionary entries for general language dictionaries. The first is what we call the conventional approach (Section 3.3.1) which had its origins in the methods devised by Johnson in the eighteenth century and which led subsequently to the publication of the *Oxford English Dictionary* (OED) in 1928. A brief historical overview of the origins of the OED is provided, and some of the principles which distinguish it from the other approaches are described. The second approach (3.3.2) is the corpus-based approach which was pioneered by researchers at the Cobuild unit in Birmingham. The third (3.3.3) is the explanatory combinatorial approach

devised by Mel'cuk and others in the eighties. This last approach has never been implemented on a large scale but warrants some discussion, nonetheless.

3.3.1 Early methods

Johnson's dictionary of the English language, published in 1755, represented the start of a new era in lexicography. ''No English dictionary had yet appeared as ambitious, particularly for establishing definitions and usage'' (Reddick 1990:25). It was the first dictionary to attempt to cover the standard language. Previous dictionaries such as Cawdrey's A Table Alphabetical (1604) had concentrated on defining what were known as hard-words, words which had come into English from other languages, principally Latin. Reddick (1990) cites Bailey and Chambers as two important influences on Johnson; the former was particularly interested in the etymology of words, the latter produced *An Universal Dictionary of Arts and Sciences*. Johnson's Dictionary was later to be edited, supplemented and expanded to become the *Oxford English Dictionary* as we know it today, the first complete edition of which was published in 1928. In 1857, one hundred years after publication of Johnson's Dictionary, the Philological Society announced the appointment of a committee ''to collect unregistered words in English'' (Compact OED, 1933: iv). They were to collect all those words which had been omitted from Johnson's and Richardson's (1819: first instalment of *Encyclopaedia Metropolitana*) dictionaries; it was intended that the output would serve as a supplement to Johnson's and Richardson's Dictionaries. In fact, the committee's efforts led, in 1858, to a proposal for a New Dictionary of the English Language to be prepared under the authority of the Philological Society. In 1859, details of the undertaking were published in the 'Proposal for the Publication of a new English Dictionary by the Philological Society'; it contained a number of basic lexicographic principles, two of which, cited in the Preface to the *Compact Oxford English Dictionary*:

> I. The first requirement of every lexicon is that it should contain every word occurring in the literature of the language it professes to illustrate.
> IV. In the treatment of individual words, the historical principle will be uniformly adopted.

Johnson had begun his task of compiling a dictionary by reading the English writers which he deemed to be suitable sources for a dictionary of the English language. According to Reddick, Johnson's first step ''was to mark passages in printed books to use as examples of usage'' (1990:33); these were marked by a short vertical line at the beginning and at the end of the passage. These passages were:

> to convey more than a simple illustration of good usage: I therefore extracted from philosophers principles of science; from historians remarkable facts; from chymists complete processes; from divines striking exhortations; and from poets beautiful descriptions. (Reddick 1990:33–34, original Johnson reference not provided).

They were to be used not only for the selection of a word but also for the purposes of exemplification and as input for definitions. In essence, he was doing what Zgusta later describes as the ''excerption of texts'' (Zgusta 1971:225). The same principle was also adopted for the preparation of the New English Dictionary. The dictionary ''aims at exhibiting the history and signification of the English words now in use, or known to have been in use, since the middle of the twelfth century'' (1933:*x*).

The intended coverage of the *Oxford English Dictionary* was:

> to deal with all the common words of speech and literature, and with all words which approach these in character; the limits being extended farther in the domain of science and philosophy, which naturally passes into that of literature, than in that of slang or cant, which touches the colloquial. In scientific and technical terminology, the aim has been to include *all words English in form*, except those of which an explanation would be unintelligible to any but the specialist; and such words, not English in form, as either are in general use, like *Hippopotamus, Geranium, Aluminium, Focus, Stratum, Bronchitis*, or belong the more familiar language. (Preface 1933)

Dictionary entries in general language dictionaries have traditionally been modelled on the entries in the OED which consist of I. the Identification (main form of the entry with its usual spelling), pronunciation, grammatical designation; II. the Morphology (which includes etymological information); III. the Signification which is the sense or meaning of the entry. Where an entry has more than one sense, ''that sense is placed first which was actually the earliest in the language: the others follow in the order in which they appear to have arisen'' (1933:*xi*). It was later established that the sense order was not always correct in the first edition of the Dictionary because the compilers may not have had all of the necessary historical facts at their disposal. The policy remains unchanged today, and different reading distinctions in the OED are still listed according to their date of origin.

The last part of the dictionary entry in the OED consists of quotations ''showing the age of the word generally, and of its various senses particularly;'' (1933:*x*). These are always quotations from published works. ''It is to be distinctly borne in mind that the quotations are not merely examples of the fully developed use of the word or special sense under which they are cited; they have also to illustrate its origin, its gradual separation from allied words'' (1933:*x*).

This was rather a tall order for lexicographers as the quotation was to show not

only the current use of the entry but also the origin and evolution of its sense. Many contemporary general language dictionaries are still compiled using the traditional excerption method and the principles underlying the defining strategies remain the same.

3.3.2 The Cobuild approach

The Cobuild approach differs from the conventional approach described above in many respects but primarily in terms of how the information for dictionary entries is sourced and the manner in which the definition is expressed. While the Cobuild dictionary is designed for use by the language learner and its coverage is therefore more restricted than that of larger general language dictionaries such as the OED, the approach adopted for compiling dictionary entries is one which could also be used for larger general language dictionaries.

> The techniques used to compile it are new and use advanced computer technology. For the user the kind of information is different, the quality of information is different and the presentation of the information is different. (Sinclair 1987:xv)

Quite apart from the fact that computers were used in the compilation process, the nature, quality and presentation of the dictionary entries made this dictionary different from all previous dictionaries. This ''was the first dictionary to present a comprehensive account of English vocabulary derived from direct observation of the way the language is being used'' (Sinclair 1995: xii). The data which was used to source the material for the dictionary consisted of a representative group of English texts stored on computer. These texts included inter alia books, magazines, conversations, pamphlets, radio and television broadcasts. Whereas, in the past, lexicographers did not have any objective means of establishing whether a particular lexical item warranted inclusion in, for example, a learners' dictionary, Cobuild lexicographers were able, with the use of computer technology, to calculate the frequency of occurrence of lexical items in their corpus and to make motivated decisions about whether or not to include a given lexical item in the dictionary. Much of the information such as definitions, grammatical fields labels, synonyms etc. which one normally finds in dictionaries still had to be produced manually using the evidence of the computer (Clear 1987:41) but access to computer output greatly facilitated the task. Cobuild lexicographers were able, for example, to view a printout of a full page of concordances of any one lexical item at a time making it much easier to formulate definitions, select examples, identify co-occurrence restrictions than if they had had to consult large numbers of dictionary slips in order to make their selection.

> So the computer was to analyse the data in ways which would make clearer the structure of English, the lexicographic team would use the computer output to supplement the traditional lexicographical tools and the results of the lexicographers' work would be fed straight back into the computer in the form of a structured database. (Clear 1987:41)

Examples in the Cobuild dictionary are selected from the corpus and are therefore representative of language as it is used. Unlike many learner dictionaries where the examples are invented and are often definitions in another form, the Cobuild principle is that examples should be examples of actual usage which complement the definition but are not themselves definitions.

3.3.3 The explanatory combinatorial dictionary

The Explanatory combinatorial dictionary (ECD) is in fact a lexicographic methodology for a general language dictionary rather than a dictionary which actually exists. It is based on the Meaning-Text Theory, proposed by Mel'cuk and Zholkovsky in 1970. From what we can gather from the literature, the methodology has been tested by means of the creation of a few hundred dictionary entries but no complete dictionary has been compiled using this methodology. The reasons for this may become clearer in our discussion. Mel'cuk's proposal for a dictionary gives priority to linguistic and lexicographic theory rather than to the target user. The objective of this new type of dictionary was to represent the lexicon of a language in a standard and uniform way. The underlying theory rests on two principles. The first:

> affirme que l'acte de parole présuppose trois composantes: l'information qui est communiquée, ou le sens; les formes qui sont perçues, ou le texte; et la correspondance entre un ensemble infini de sens et un ensemble infini de textes qui constitute la langue à proprement parler. Une langue est donc perçue comme un ensemble de règles qui établissent la correspondance entre un ensemble infini de sens et un ensemble infini de textes ou vice versa. (Mel'cuk 1984:xiv)
> (Translation: asserts that the speech act presupposes three components: the information which is communicated, i.e. meaning; the forms used, i.e. the text; and the relationship between an infinite number of meanings and an infinite number of texts which make up a language. A language is therefore perceived as a body of rules which establish the relationship between an infinite number of meanings and an infinite number of texts or vice versa).

The second:

> affirme la caractéristique fonctionnelle ou cybernétique du modèle Sens - Texte où

> seules les entrées et les sorties sont directement observables. (Mel'cuk 1984:xiv) (Translation: asserts the functional or cybernetic characteristic of the Meaning-Text model where only the inputs and outputs can be observed).

Mel'cuk perceives the dictionary as a system of lexical relations where all of the relations of any one entry to any other entry in the lexicon must be specified, thereby creating a network. The resultant network functions as a large-scale internal pointer system, with systematic cross-referencing which allows the user to know that certain entries point to others and also to know why, and how they relate to those other entries (Frawley 1980/1981). What distinguishes the ECD from other general language dictionaries is the use of a precisely defined set of lexical functions to describe the relations between entries. Mel'cuk devised some 60 lexical functions for this purpose. The lexical functions used in an ECD:

> constitute an important part of the Meaning-Text Theory (MTT) of language, for they systematically describe certain semantic and collocational relations between lexemes. At the two deepest levels of representation within the MTT - the Semantic Representation and the Deep Syntactic Representation - the name of a function together with a keyword may be used to signify a set of either phraseological combinations related to the keyword or those words which can replace the keyword under certain conditions. (Steele and Meyer 1990:41)

> a lexical function (LF) is a dependency relation between the 'argument' (or the keyword of a function) and a 'value' (or the linguistic expression that realizes the meaning of a function or expresses its syntactic role in relation to its argument) with the function itself defining the kind of dependency relation that is involved." (Steele and Meyer 1990:42)

The functions roughly subdivide into two subclasses: paradigmatic and syntagmatic. ''The paradigmatic class consists of functions that are based essentially on meanings regularly associated with a keyword as an element in a language's system of semantic relations'' (Steele and Meyer 1990:42). Paradigmatic functions divide into four groups: substitution, qualifiers of a keyword, aspects of a keyword situation and qualifiers of actants. ''The syntagmatic class consists of functions that are based primarily on relations arising from the collocational properties of the keyword as an element in a language's speech system'' (Steele and Meyer 1990:42). These subdivide logically into verbal operators (''phrasal patterns linking nouns to verbs'', Steele and Meyer 1990:50) and predicators (''used to carry out synonymic transformations'', Steele and Meyer 1990:50).

A dictionary entry in an ECD is organized in a highly formalized and structured way and is described by Elnitsky (1990) as follows. It subdivides into four main

zones: I. the introductory zone which contains:1. the headword, 2. morphological information, 3. syntactic and stylistic information; II. the semantic zone which contains:4. the propositional form which only applies if the entry is a predicate in the logical sense of the word, 5. the definition, 6. connotations; these consist of sentences where the subject is or includes the entry; III. the syntactic zone which contains:7. the government pattern (GP), 8. restrictions on the GP, 9. examples illustrating GP and restrictions thereon, 10. syntactic modifications of the headword; IV. lexical functions zone which contains:11. lexical functions, 12. examples illustrating a lexeme's meaning and co-occurrence; V. fixed expressions zone. He maintains that at the very least slots, 1, 2, 5 and 12 must be filled in order for an entry to be complete; it would appear therefore that the inclusion of slot 11 (lexical functions) is not considered essential. Yet, with the exception of the lexical functions slot, the information contained in an ECD is very similar to what we have come to expect in general monolingual dictionaries.

An entry in an ECD dictionary may be termed a *vocable* (super-entry) or a *lexème* (entry). The term *vocable* is used to describe an entry which has more than one reading. The term *lexème* is used to describe an entry which has only one reading.

> Un vocable est la famille de tous les lexèmes tels que:
> (i) leurs signifiants sont identiques;
> (ii) les signifiés de deux lexèmes quelconques sont liés soit directement, soit par l'intermédiaire d'une chaîne de liens directs." (Mel'cuk 1984:4)
> (Translation: A super-entry is a family of entries:
> (i) with identical signifiers;
> (ii) where the meanings of any two entries are related, either directly or through a chain of direct links.)
> "Un Lexème est un mot pris dans une seule acception bien déterminée et muni de toutes les informations caractérisant le comportement de ce mot justement lorsqu'il est utilisé dans cette acception." (Mel'cuk 1984:4)
> (Translation: An entry is a word with one sense only, complete with all of the information which characterizes the usage of this word.)

When two entries appear in the same super-entry, they are deemed to be polysemous. When two entries have the same signifier but appear in two different super-entries because there is no direct link between their meanings, they are deemed to be homonyms. The definition, which appears in slot 4 or 5 in the ECD, has the following functions:

> An ECD definition explains a meaning in terms of its semantically simpler components. These are offered in the context of a proposition that is formally equated with the propositional form of the *definiendum*. This propositional mode of definition serves

> to represent the headword as a 'function,' that is to say, as a meaning that has obligatory 'slots' or 'places' for complements, which are represented both in the propositional form and in the definition by the variables X,Y,Z…When these slots are filled, they become the semantic actants of the headword.(Meyer and Steele 1990:66)

The lexical functions of an entry are indicated in slot 11. Below is an example provided by MacKenzie (1990) of the lexical functions for just one sense of the superentry *escape*.

> Example:
> [ESCAPE]
> **I.1a** X *escapes from* Y *through* Z = X, being kept by Y1 against X's will in place or state Y2, such that Y1's intent is to thwart any attempt by X to leave Y2, succeeds in leaving Y2 via Z thereby becoming free.
> **Lexical Functions**
> **Syn** ∩ : break out, run away
> **S_0** : escape I.1a
> **S_1Perf** : escapee
> **S_1** ∩ : runaway, fugitive
> **S_1Able$_1$** ⊂ : escape artist 2
> **S_2** ∩ : guard, jailer, turnkey [=Y1]; place of confinement; jail, prison camp, concentration camp, lockup, dungeon [=Y2]
> **S_3** : escape route
> **A_1** ∩ : runaway, fugitive, fleeting
> **Bon** : daringly
> **Qual$_1$** : kept, imprisoned, guarded
> **Examples**
> He escaped from custody. Dreyfus did not escape from Devil's Island; he was finally released as a result of mounting public outrage. A plot by at least six inmates to use a crossbow to kill a tower guard—or incinerate the tower and then escape from Trenton State Prison over a home-made bridge—has been thwarted, state correction officials said today. He managed to escape from the miner's cabin while the kidnappers were in the kitchen. Three more East Germans have just escaped over the Berlin Wall in a home-made balloon. Four o'clock had come and gone with still no sign of little Billy, and Martha's head was awhirl with visions of the lion that had escaped from Riddington Zoo the previous night.(Mackenzie 1990:97–98)

I.1a above provides the propositional form of the entry which consists of the "headword and the variables representing all of its semantic actants" (Meyer and Steele 1990:65). Syn ∩ means that *Escape* intersects with *break out, run away*.S_0: indicates that *escape* I.1a is a derived noun with the same meaning as the keyword.

S_1Perf, *escapee*: S_1 is the standard name for the first participant in the keyword situation; S_1Perf is the term given to the participant when the process has been carried through to its natural limit. $S_1\cap$:*escapee* intersects with *runaway, fugitive*. S_1Able_1 $\subset$: escape artist 2: the first participant is capable of escape; this reading is narrower than escape artist 2. $S_2\cap$: the second participant intersects with jailer, turnkey etc. S_3: the third participant indicates the 3rd actant of the keyword. $A_1\cap$ means that *runaway, fugitive, fleeting* are typical qualifiers for the first actant of the keyword.Bon(escape) daringly. Bon indicates a standard expression of praise or approval. $Qual_1$:*kept, imprisoned, guarded*; $Qual_1$ entails $Able_1$ with a high degree of probability. In addition to the information provided in this lexical functions slot, we can also expect at least slots 1, 2 and 5 to be filled.

Surprisingly, according to Meyer and Steele (1990:85), examples, which appear in slot 12, can be invented by the lexicographer or drawn from contemporary press or twentieth century literature; if the example is drawn from an authentic source, it should be followed by the name of the author in brackets. The absence of any indication of source would seem to suggest that all of MacKenzie's examples above are in fact invented; this is rather surprising given how easy it has now become to source authentic examples in electronic texts. They also suggest that examples should be ordered according to the order of the elements which they are intended to illustrate. However, as the examples are not numbered, the reader must guess to which part of the entry each example refers. The sample entry provided here only applies to one sense of the super-entry *escape*; it is to be assumed that other senses of the same super-entry will be of a similar length. The space taken up by slot 11 alone is considerable; this would suggest to us that a complete entry in an ECD is likely to be very long, and perhaps also quite difficult to interpret.

The reason why we chose to look at the ECD method is that some authors such as Frawley (1980/1981) have suggested that it is ideally suited to the compilation of specialized dictionaries because the format ensures that all relevant information is provided and that the results are systematic and consistent. Frawley proposes the following 11 lexical functions for entries in specialized dictionaries: taxonomy, synonymy, antonymy, grading (whereby an entry has to be explicit in its serial relations to others, e.g. *linear, quadratic, cubic* etc.), cause, part/whole, source (the entry should specify entries for which it is a source), result, continue (how an entry's continuation is labelled by another entry), inception (other entries which an entry may be said to begin), etymology.

We appreciate the potential advantages of using a systematic and formulaic approach for specifying relations between dictionary entries as it would allow lexicographers to check that all relevant information was present and would give users access to related terms. The information collected by the lexicographers could be used as input for a natural language definition, which would be more user-friendly

than the sample entry presented above. The entry, as presented here, is too complex for human processing but the information contained in the lexical functions slot might be useful if the dictionary were to be converted to a lexicon for natural language processing. Making the information available in a systematic way would certainly make it more tractable. However, while there are clear advantages to using a systematic approach, the absence of any complete dictionary compiled using the ECD approach leads us to conclude that it may be quite difficult to apply on a large scale.

3.4 Explaining meaning

In dictionaries, the meaning of a word can be explained in a number of different ways and different strategies are adopted depending on the type of dictionary involved. Ilson (1986a) considers that dictionaries have four methods of explaining meaning: these are illustration, exemplification, discussion, and definition. Each of these is described briefly below and we then look in detail at different defining strategies.

One method of explaining meaning is by means of illustration. Illustrations are generally presented in the form of pictures or tables and are likely to be found inter alia in children's dictionaries, pictorial and other specialized dictionaries. While they are sometimes used as a supplement to textual explanations, they can also be used as the sole means of explaining a word. In the Duden series of pictorial dictionaries where definitions are not provided, items in illustrations are numbered and the appropriate term is provided in a list accompanying the illustration.

Exemplification can serve one of two purposes; it may be used to exemplify the meaning of the entry (i.e. the referent) or it may be used to exemplify its usage. When the meaning of the word is exemplified, exemplification involves citing examples of the word (e.g. *dog*: spaniel, labrador etc.) and may replace a definition. When usage is exemplified, a definition is also usually provided, and exemplification allows for display not only of the ''meanings of words, but also their syntax, selectional restrictions, collocations and stylistic level'' (Ilson 1986a:216). As already noted, (Section 3.3) different lexicographic traditions view exemplification differently. The OED, for example, prefers to use examples to exemplify both meaning and usage but Cobuild lexicographers prefer to use examples to exemplify usage alone because they believe exemplification should not simply become an extension of the definition.

Discussion is the term which Ilson uses to describe the basic technique used by many dictionaries for explaining function words (i.e. words such as *the, and, of*).

Definitions are perhaps the most common method of explaining meaning and the

classical analytical definition involves defining a word ''per genus et differentiam'' (Ilson 1986b: F3), i.e. in terms of a lexical item's superordinate and a distinguishing characteristic which distinguishes the lexical item from other members of the same group. The relationship between a lexical item and its definition can be expressed in a number of ways and, in general, lexicographers will opt for one of two possible definition styles. According to Ilson (1986a), definition styles can be broadly subdivided into two types, standard (or ''dictionary'') definitions and folk (or ''grassroots'') definitions. Hanks (1987) prefers to talk about different 'explanatory strategies'. These also subdivide into two broad categories, the 'substitutable defining strategy and the 'Cobuild' strategy.

3.4.1 Substitutable defining strategy

''Standard definitions are connected to their definienda by the implicit verb means, so that the basic definitional proposition is: Definiendum [means] Definiendum'' (Ilson 1986a:218). In a standard substitutable definition, the definition ''is supposed to be substitutable for its definiendum in any context in which the definiendum does or can appear'' (Ilson 1986a:218). Hanks traces this desire for substitutability back to the late 17th and early 18th centuries when ''formalism became the spirit of the age'' (Hanks 1987:119). Two expressions were deemed to be equivalent in meaning if one could be substituted for the other. The pursuit of substitutability led lexicographers to ''formulate definitions that could be substituted in any context for the word being defined'' (Hanks 1987:119). Already in Johnson's dictionary, there was evidence of the growing trend towards what Hanks terms dictionary-ese. While Johnson still used many discursive explanations, he also used explanations where the definiens and definiendum were approximately substitutable. By the time the *Oxford English Dictionary* was published, the principle of substitutability had become standard. This resulted in lexicographers going to enormous trouble in order to satisfy the substitutability principle, and also resulted in very unnatural looking definitions. As Hanks suggests:

> As far as I can find out, there was no explicit discussion of the pros and cons of the awkwardnesses in the phrasing of the definitions that resulted and, more seriously, there was no discussion of whether the formulae so concocted faithfully reflected the facts of natural language or whether they introduced distortions. (1987:119)

Substitutability applied not only to the lhs (left hand side) and rhs (right hand side) of the defining statement in the dictionary entry but was also intended to ensure that any entry could be substituted by its definition when used in context. It seems surprising that the notion of substitutability, which is in fact relatively recent, should

have been accepted so unquestioningly by so many lexicographers. There are some detractors, Landau, for example, and of course the Cobuild defining strategy (cf. Section 3.4.2) where the notion of substitutability is rejected. ''Substitutability is often declared to be a principle of defining, but there are so many cases where it is impossible to apply that it is idle to insist that it be universal'' (Landau 1989:132). While Landau accepts that the substitution rule can help readers to understand and even use a word, she believes that ''definition can be given very well without it'' and ''sometimes the effort to make a definition substitutable impairs its clarity by forcing the definer to use a clumsy or ambiguous phrasings'' (1989:134). She also states that ''it is plain that a substantial percentage of the definitions in most dictionaries do not substitute even approximately for their definienda in context. There is no reason why they should'' (Landau 1989:133–134).

As Landau indicates, it is frequently impossible to use the substitutability principle and, while it may be useful as an aid to understanding, there are other better methods for expressing the meaning of words. Landau also refutes the notion that the definiens and definiendum should be substitutable not only within the confines of the dictionary but also in ordinary language usage.

3.4.2 The Cobuild defining strategy

The second defining method involves writing definitions in ordinary prose, hitherto maligned by lexicographers such as Ilson who described these definitions as follows: ''Besides standard definitions, other types of definition, typically non-substitutable, are used by ordinary folk, and even by lexicographers off duty. . . . e.g. (Tired) is how you feel after writing a paper for a learned conference'' (Ilson 1986a:219). Ilson refers to these as folk definitions and rejects them on the grounds that they are not substitutable. Cobuild opted for just such non-substitutable definitions. It decided to abandon the traditional substitutable method of formulating dictionary entries in favour of a method which involved the use of ordinary prose which the target user would be able to understand.

> The dictionary is designed to read like ordinary English. Words appear in their normal full spelling forms and the explanations are written in real sentences. (Sinclair 1987: xvi)

> the definitions (or explanations, as we often call them) are written in full sentences, using vocabulary and grammatical structures that occur naturally with the word being explained . . . . It also enables us to give a lot of information about the way a word or meaning is used by speakers of the language. (Sinclair 1995: xviii)

This innovation of Cobuild represented a major departure from traditional practice. As Hanks (1987:117) explains: ''Each explanation consists of two parts. The first part represents a departure from lexicographic tradition, in that it actually places the word being explained in a typical structure.'' Traditionally, in the substitutable definition, the lhs consisted exclusively of the word being defined. Cobuild chose to use the lhs as a means of showing the entry in a typical structure; thus, in the case of a verb for example, the lhs will show its typical arguments by placing the entry (emboldened) in a sentence. Information about context can be provided by inserting, at the beginning of the entry, an adverbial phrase which specifies the domain in which the entry is used, thereby eliminating the need for a separate subject field reference. As Hanks states: ''In general, then, the first part of each Cobuild explanation shows the use, while the second part explains the meaning'' (Hanks 1987:118).

> The second part of each explanation consists of a more traditional-looking dictionary definition:
> . . . a rectangular block used for building walls, houses, etc.
> These second parts identify the meaning. They are always to be read as stating what is typically the case, not as providing sets of necessary or sufficient conditions.
> (Hanks 1987:118)

To illustrate the difference between the two defining strategies, we have chosen the entry for the verb **steer** in the Cobuild (1987) and Webster New World (1991) dictionaries:

Example from the Cobuild Dictionary

1	When you **steer** a car, boat, plane, etc., you operate it so that it goes in the direction that you want.	**guide**

Example from the Webster New World Dictionary

1 to guide (a ship or boat) by means of a rudder
2 to direct the course or movement of [to *steer* an automobile]

In the Cobuild entry, the lhs of the defining statement shows that **steer** typically has a person in subject position and some type of vehicle, e.g. *car, boat, plane* etc. in object position. The user therefore gains some idea of how to use the verb. The rhs of the defining statement explains the meaning of **steer** in very simple language, and the superordinate or more general word (**guide**) is provided in the column beside the entry. In the Webster entry, the principle of substitutability is applied and a phrase synonymous with the word being defined is provided. The objects of **steer** are indicated in brackets after the superordinate verb but the subject is not indicated and it

is up to the dictionary user to infer that the subject is usually human. In our view, the Cobuild defining method is superior to the substitutability defining method because it manages both to convey meaning and to demonstrate usage in the definition. It seems likely that this approach will have an impact on future developments in general language lexicography. In particular, we would suggest that this strategy might be suited to the formulation of definitions for entries in specialized dictionaries where examples of usage are not provided. A definition formulated in the Cobuild manner would give the user some indication of how the word is used in text and would not necessitate much additional space for the entry.

3.4.3 ISO recommendations for definitions

In *ISO 1087* a definition is defined as ''a statement which describes a concept and permits its differentiation from other concepts within the system of concepts'' (*ISO 1087*:4). Two types of definition are recommended for terms, intensional and extensional definitions. The intensional definition describes a term in terms of its superordinate and the characteristic(s) which distinguish(es) it from its superordinate. ISO stipulates that ''it is necessary to state the closest generic concept that has already been defined or can be assumed to be generally known, and to add the restricting characteristic that delimit the concept to be defined'' (*ISO 1087*:4).

3.5 Recommendations for good defining practice

ISO 1087 makes a number of useful recommendations regarding the formulation of terminological definitions. These relate in particular to the choice of superordinate in an intensional definition, a definition where a term is being defined in terms of its superordinate and a distinguishing characteristic. The reasons why the choice of superordinate is crucial for helping users to understand terminological definitions are outlined. Lexicographers have to make decisions about which terms should be defined in a dictionary. We argue for greater coverage in specialised dictionaries, in particular in relation to what have been described as subtechnical terms (Section 1.6). Opinion is divided about how a dictionary definition should be expressed, particularly with regard to the degree of complexity of the language used in the definition. The arguments for and against a relatively transparent approach are explored.

3.5.1 Selection of a superordinate

As our investigation focuses on definitions of terms which are nouns or noun phrases, the discussion here is confined to this particular category. Such terms are

frequently explained using the classic formula: X = Y + distinguishing characteristic(s). X is the entry, Y is the superordinate and the distinguishing characteristic is intended to distinguish X from other members of the same class. The selection of an appropriate superordinate is crucial to the intelligibility of the defining statement, particularly in the case of terms. As *ISO 1087* stipulates: "For this purpose, it is necessary to state the closest generic concept that has already been defined or can be assumed to be generally known . . . . "(*ISO 1087*:4). The superordinate or closest generic concept selected should preferably be just one step up in the hierarchy from the term being defined; in other words, users will find it easier to fit a term into the hierarchy to which it belongs if the immediate superordinate is specified. Furthermore, it is important that the same superordinate is specified for all terms which belong to the same class and are at the same level of abstraction. To illustrate the problems associated with inconsistency in the use of superordinates, we have chosen some examples from general vocabulary of entries (the pieces of cutlery *knife, fork, spoon*) in two general language dictionaries. It would not be unreasonable to expect the definitions of these pieces of cutlery to share the same superordinate. Yet, we find the following in the *Collins English Dictionary*:

Examples from *Collins English Dictionary* (1991)

> **knife** 1. a cutting instrument consisting of a sharp-edged often pointed blade of metal fitted into a handle or onto a machine.
> **fork 1**. a small usually metal implement consisting of two, three or four long thin prongs on the end of a handle, used for lifting food to the mouth or turning it in cooking, etc.
> **spoon 1**. a metal, wooden, or plastic utensil having a shallow concave part, usually elliptical in shape, attached to a handle, used in eating or serving food, stirring, etc.

The *Collins English Dictionary* makes life difficult for the user who is not familiar with these words by specifying a different superordinate for each of the pieces of cutlery. A user wishing to establish the meaning of each of the superordinates specified would find the following in the same dictionary: the superordinate for *instrument* is mechanical implement or tool, the superordinate for *implement* is piece of equipment, tool or utensil and the superordinate for *utensil* is implement, tool or container. This demonstrates the type of confusion that can be caused by inconsistency in the use of superordinates. The entries for the same examples in the Cobuild dictionary are much easier to understand:

Examples from *Collins Cobuild English Dictionary* (1995)

> A **knife** is an implement used for cutting food. It consists of a flat sharp-edged piece of metal on the end of a handle.

A **fork** is an implement that you eat food with. It consists of three or four long thin prongs on the end of a handle.
A **spoon** is an implement used for eating, stirring, and serving food. One end of it is shaped like a shallow bowl and it has a long handle.

Here, each of the examples has the same superordinate (i.e. *implement*) thereby facilitating comprehension and *implement* is in turn defined as *a tool or other piece of equipment*. While the examples provided here are ordinary examples with which readers are likely to be familiar, it is easy to imagine, on the basis of these, how much more difficult it would be to understand definitions of technical terms where superordinates are not used consistently.

3.5.2 Coverage

One of the issues which lexicographers have to address in the compilation of a specialised dictionary is which terms to include and define in the dictionary. Should all terms used in a particular subject domain be defined or should there be a cut-off point where a decision is made to exclude some terms, such as generic terms or what have been described as subtechnical terms? *ISO 1087* stipulates that all superordinates used in terminological definitions should be defined elsewhere in the same publication unless they ''can be assumed to be generally known'' (1990:4). Superordinates which can be assumed to be known include generic concepts such as *technique, process, device* and it is understandable that these would not be defined in a specialised dictionary. However, all other superordinates should preferably be defined elsewhere in the publication.

Subtechnical terms pose another problem for lexicographers. These are terms which represent notions general to all, or most subject fields (Yang 1986:98). Space constraints may dictate the cut-off point for lexicographers but the omission of such terms from specialized dictionaries can pose problems for users, particularly for users such as translators who may not be familiar with the domain in question. In principle, we would therefore recommend that all terms used in a particular domain, whether specific to that domain or whether also used in other domains should be defined in a specialized dictionary covering the domain in question.

3.5.3 Choice of language for the definition

There are two schools of thought about what type of language should be used in a definition, one which believes that the language should not be more sophisticated than the word being defined and another which believes that it is neither useful nor practical to adopt this approach. Zgusta belongs to the first school and firmly

asserts: ''Nor should the lexicographic definition contain words more difficult to understand than the explained word itself'' (Zgusta 1971:257). Landau, however, does not believe that this is always possible and states: ''Avoid including difficult words in definitions of simpler words'' is a traditional rule that seems to make sense, but like so many lexicographic rules it is often impossible to apply'' (Landau 1989:134). Strehlow suggests that the defining of technical terms (in technical dictionaries) differs from the defining of words in common usage in two respects.

> First, the users of a technical term are concerned with their subject matter in greater detail than is the case with common words, necessitating increased care in maximizing the precision and accuracy of definitions. And second, definitions of technical terms are often required to contain documentation of their source or specific limitations of scope. (Strehlow 1983:16)

It is true that there is a requirement for maximum precision and accuracy in specialized dictionaries but this does not preclude the lexicographer from using simple language in the definition or at least from ensuring that all terms used in the definition are defined elsewhere in the publication. As the main purpose of dictionaries is to facilitate understanding, a definition which does not define simply will frustrate and confuse readers.

3.6 Conclusion

This chapter provided a broad classification of language dictionaries, both general and specialized and described what a typical entry in each of these dictionary types is likely to contain. Three different approaches to lexicography were described: the traditional approach, the ECD approach and the Cobuild approach. The defining strategies used in the traditional and Cobuild approaches were outlined and exemplified. The Cobuild approach is preferred because the definitions provide information not only about the meaning of an entry but also about its usage, and this approach might usefully be applied in the compilation of specialised dictionaries where information about usage and typical grammatical structures is not usually present. A further reason for favouring the Cobuild approach is that it makes use of authentic textual data; decisions are made on the basis of what people actually write and say rather than simply on the basis of lexicographers' intuition.

4 Analysis of Definitions in Text

> Definitions occur frequently in many types of scientific writing because it is often necessary to define certain operations, substances, objects or machines.
>
> (Swales 1971:66)

4.1 Introduction

This chapter examines some of the research which has been carried out into the expression or realization of definitions in text. Our objective was to ascertain whether and how definitions are formulated in text. Given the widespread interest in teaching languages for special purposes, we had anticipated that this was an area which would already have been widely explored previously. In fact, as this chapter demonstrates, research has focused on teaching non-native students, of English primarily, how to formulate definitions when writing scientific prose rather than on examining authentic texts to ascertain whether and how subject specialists formulate definitions when writing. Consequently, the emphasis tends to be on what these researchers consider to be 'typical' definition formulae rather than on what the text actually tells them. There are a few exceptions, namely Selinker, Trimble and Trimble (1976) and Flowerdew (1992). While this appeared at first to be a serious drawback from our point of view, we later found that we were able to use some of the findings as a basis for our analysis. This chapter charts the work of Swales (1971, 1981, 1985), Allen, Widdowson (1974), L. Selinker, R.M. Trimble and L.Trimble (1976) and L.Trimble (1985), Darian (1981), L. Trimble (1985) and Flowerdew (1992).

4.2 Swales

In Swales' work on definition, he is concerned primarily with discovering how to teach students to formulate definitions; he opts for a subject specific rather than a broad based academic approach for this purpose. He argues in favour of a subject specific approach on the basis that the purpose of definitions will vary from one subject to another. For example, while definitions in legal textbooks will frequently

have the same form and function as definitions in science textbooks because, as Swales (1981:109) states "they provide terminological explanations designed to sharpen up a layman's appreciation of the meaning of the terms being defined", such definitions tend to be more complete in legal textbooks than in science textbooks, as legal writers will endeavour to cover all contingencies. Besides the formal type of definition discussed below, Swales argues that there is another category of definitions in legal texts which has a very different function; this category constitutes the law itself, and, as such, cannot be rewritten or rephrased for purposes of clarity or simplicity. Swales cites this category as the reason for his objections to the broad based approach to the teaching of academic language. It is interesting that, although he prefers a subject-specific approach, (in this case, science), many of his examples are not drawn from science; this leads one to wonder about the justification for his claim that different subject fields express definitions differently.

Swales concentrates on the formulation of what Trimble (cf. Section 4.6) later describes as formal definitions, whereby:

> . . . the thing to be defined should be described first in terms of its general class then in terms of its particular properties, qualities, uses, or origins. This could be expressed as
> $T=G + (d_{a,+}\ d_{b\ ,+}\ dc\ \ldots\ d_n)$
> where T equals the thing to be defined
> where = equals *be*
> where G equals a general class word.
> where $d_{a,+}\ d_b$ etc. are the properties which distinguish T from the other members of the general class. (Swales 1971:66)

He suggests that the most common definition formula is:

> An {x,y} *is a/an* general class word + *wh-* word . . . where *x* is a countable noun, where *y* is an uncountable noun. (1971:67)

According to Swales, definitions always commence with the indefinite article; they will not commence with the definite article because definitions are general statements. We would argue, however, that while it is true that definitions will not commence with the definite article, it is not true to conclude that all definitions commence with the indefinite article. In the case of definitions of uncountable nouns, for example, the use of the indefinite article would not be grammatically acceptable, unless the reference was to a particular type of that uncountable (e.g. an aluminium, a cement) in which case it would no longer be an uncountable. With regard to verbs

used in the expression of definitions, Swales suggests that the most common verb is *is* although he acknowledges that the phrase *can be defined* is sometimes used. It is important to note that he does not draw on any authentic data to support this argument; this leaves it open to debate. With regard to the way the remainder of the defining statement is expressed, Swales proposes and exemplifies a number of ways in which this can be done.

i) It can be completed with either active or passive clauses. He provides the following examples: 1) "A dentist is a person who takes care of people's teeth" (1971:68), and 2) "A knife is an instrument which is used for cutting things" (1971:69). These are rather surprising examples if it is being suggested that they are representative of what one finds in scientific prose. They are much closer to the type of definition which we have come to expect in the Cobuild Dictionary, i.e. written in simple language. While it is possible that Swales may have deliberately chosen examples from general language in order to state his case more clearly, the examples cited in (ii) below would suggest that he is not always consistent in doing so.

ii) The definition may also be completed by what Swales terms reduced relative clauses. For example, "Aluminium is a metal produced from bauxite" (1971:70) and "A tangent is a straight line touching a curve at one point" (1971:72). Note the absence of an article before *Aluminium* which confirms that there are indeed situations where the indefinite article would be out of place. In the above examples, the past participle or a verb *+ing* are used instead of a *wh-* word. According to Swales, the verb *+ing* is particularly prevalent with the verbs *contain* and *consist of*. The inclusion of *consist of* and *contain* as main verbs in the defining statement comes as a surprise as Swales states at the start of the article "Other forms of *be* are not common. Other main verbs are also uncommon" (1971:68). Perhaps he is suggesting that the verbs *contain* and *consist of* are only used in the *+ing* form? As he provides no authentic textual evidence to support this, readers are once again left in doubt as to the validity of the statement.

Swales then goes on to say that definitions may be reduced further but only in the context of *used for*:

> A knife is an instrument which is used for cutting.
> A knife is an instrument used for cutting.
> A knife is an instrument for cutting. (1971:71)

whereby the first is an example of the full form, the second is an example of the reduced form and the third is an example of further reduction.

iii) The *wh-* word may be preceded by a preposition "when the subjects of the two parts of the definition statement are not the same" (1971:73). For example,

"Acoustics is a branch of physics in which the properties of sounds are studied" (1971:73).

iv) Swales cites two methods of writing scientific definitions which do not use the relative clause:1) "A triangle is a plane figure with three sides. Tungsten is a metal with the property of retaining hardness at red-heat" (1971:74).

He summarizes these methods of formulating definitions as follows:

An *x/y* is class word
- *which is* verb + *ed*
- *verb* + *ed*
- *for* verb + *ing*
- *wh-* word + *s*
- verb + *ing*
- preposition *wh-* word . . . .
- *with* noun phrase
- *with the property of* verb -*ing* (1971:74)

He makes a questionable distinction between general and specific definitions. In general definitions, "the thing to be defined has usually been a single noun unaccompanied by other nouns or adjectives specifying it" (1971:75). He cites the following example: "A saw is an instrument used for cutting wood" (1971:75). A specific definition is one where a specific type of thing is defined rather than something in general, e.g. "A key-hole saw is a saw with a narrow blade, used for cutting holes in wood" (1971:75). We would suggest that there is no reason to make this type of general/specific distinction. It is always preferable for words and terms to be defined in terms of their immediate superordinate (cf. Section 3.5.1) and if the immediate superordinate is a general class word, then so be it. Thus, one might describe a *knife* as an *instrument* but, in the logical order of things, it is more logical to describe a *carving knife* as a *knife*, thereby allowing it to inherit the characteristics of *knife* rather than simply to describe it as an *instrument*; this would necessitate repetition of the characteristics of *knife* as well as the characteristic which distinguishes the *carving knife* from *knife*. As a consequence of his distinction between general and specific definitions, Swales offers the following revised formula:

$$T + t = \{t \text{ or } G\} + d_a + d_b \ldots d_n \quad (1971:75)$$

Swales notes that definitions are not necessarily confined to one sentence and may be expanded over two or more sentences; examples of expanded definitions are as follows:

Definition formulae +

- Common examples are a, b, c and d
- Typical examples are a, b, c and d
- Main types are a, b, c and d
- such as a, b, c and d
- Therefore, it is used
- As a result, one of its main uses is . . .
- It consists of . . . main parts: . . .
- Its main components are . . . (1971:80)

whereby the general statement is made in one sentence and exemplification or further information is provided in the following sentence. These expanded statements, which Trimble (1985) subsequently terms complex definitions (cf. Section 4.6.4) are of particular interest to us and will be explored in greater detail in chapter eight.

4.2.1 Summary

Swales, as he states, is interested in teaching students of science how to formulate definitions because he believes definitions to be an integral part of scientific prose. Consequently, students need to be able to define. Yet, in 1981, he suggests that definitions are in fact rare in reported research articles but very common in science textbooks. The reason why definitions are rare in reported research articles is because the function of definitions is ''more to furnish explanations of terms than to establish axioms which form part of a logical system of postulates and theorems'' (Swales 1981:107). One wonders then why he chose to teach students something which they were unlikely to use in a productive sense. Had Swales been interested in teaching comprehension, one might understand the emphasis but he explicitly states that:

> It is a writing course . . . . We decided therefore that we could best help them (the students) by concentrating on the productive skills; by teaching, discussing and correcting writing we could improve an aspect of our students' performance that they found particularly difficult to do for themselves. (1985:72)

In spite of these quibbles, his motivations for teaching this particular skill are, in fact, of little concern to us and we are far more interested in establishing whether we can use any of the formulae which he proposes.

Although Swales appears to make use of constructed examples throughout his work, he provides some useful information on the manner in which definitions are

expressed. The notion of generic reference is one which this author had already noted through examination of authentic data (cf. Section 6.5.1) and it is a valid one. The argument that definition statements generally make use of *is* as the main verb is more contentious, and even Swales accepts that *consist of* and *contain* are also used. The use of reduced clauses to introduce the distinguishing characteristic is one which, as we shall see, is actually used. It has also been our experience (cf. Section 7.3) that definitions are not necessarily confined to one sentence. While much of what Swales says appears to be simply what has occurred to him and is therefore not presented in a very systematic manner, it does provide a starting point for more systematic work based on authentic data.

4.3 Widdowson

Widdowson's interest in science language arose out of his involvement in "the teaching of English to students who need to know the language in order to pursue their studies of science and technology in higher education" (1979:21). His aim was "to prepare them (students) for their encounter with scientific communication in English such as they will find in their textbooks" (1979:28). While Widdowson provides definition exercises in the English in Focus series of textbooks which he co-authored with J.P.B. Allen, (cf. *English in Physical Science* 1974:4), and also cites examples of two common forms of definition in scientific discourse,

(a) A [is/are / may be defined as] B which C.

(b) B which C [is/are called / is/are known as] A. (1985:81)

he does not appear to have been concerned with further documenting and explaining how definitions are expressed, opting instead to allow the exercises to speak for themselves. Widdowson does tell us how the exercises were constructed, and it is really quite alarming to realize that someone who is an applied linguist and should therefore be sensitive to the pitfalls of language should so blithely assume that he is capable of writing science. He states that, as students need a course which develops "a knowledge of how sentences are used in the performance of different communicative acts," (1985:74):

> We do this by composing passages on common topics in basic science and presenting them in such a way as to develop in the student an awareness of the ways in which the language system is used to express scientific facts and concepts. The passages are com-

> posed rather than derived directly from existing textbooks for two reasons. Firstly, we are able to avoid syntactic complexity and idiosyncratic features of style which would be likely to confuse students . . . . Our intention is to make linguistic forms as unobtrusive as possible. At the same time, we wish to make their communicative function as obvious as possible, and this is the second reason for composing passages: we are able to 'foreground' features of language which have particular communicative value. (1985:75)

While one can understand that it may be important to ''foreground'' certain aspects of scientific communication, it is still very difficult to understand why Widdowson did not simply choose, for example, to use authentic data or to consult with science writers for assistance. If it is indeed true that syntactic complexity and stylistic idiosyncrasies are a feature of scientific text, then this is even more reason for using authentic texts. Widdowson acknowledges that all of his examples are constructed. When objections are raised, Widdowson simply replies that the passages ''are representative of what we conceive to be certain basic communication processes which underlie, and are variously realized in, individual pieces of scientific writing'' (1985:75). It is such a pity that so much time was lost by linguists such as Widdowson in their failure to appreciate the importance of consulting authentic data.

4.4 Larry Selinker, R.M.Todd Trimble, Louis Trimble

> One of the most important and frequently employed rhetorical functions is that of 'definition'; this function is basic to the scientific thinking and reporting processes. (Selinker, Trimble, Trimble 1976:39)

These authors appear to have based their research on authentic texts because they say that on examining ''paragraphs in naturally occurring EST discourse'' (1976:39) and ''after looking at large amounts of EST discourse'' (1976:40), they found that a discrepancy exists between EST (English for Science and Technology) textbook exercises and EST in reality. ''We do not find, in fact, many specimens of the 'pure' examples that we have typically given EST students to practise on'' (1976:40). They recognize that the 'pure' examples which have been offered to students in the past do not exist in authentic data. These authors recognize the general importance of definition in scientific thinking and reporting. They write of the explicitly stated definition where the reader is given:

> 1) the term being defined, 2) the class of which the term is a member, and 3) a statement of the essential characteristics or differences which distinguish the term from the

> other members of the class. (1976:39)

This is what Trimble later calls a formal definition (cf. Section 4.6.1); it is not as common as Swales would have had us believe (cf. Section 4.2). What is of particular interest in the article by Selinker et al. is their acknowledgement of this fact and their suggestion that defining information is more likely to be provided implicitly rather than explicitly. The defining information is likely to be 'buried' in what they term the 'supporting information' of a paragraph rather than in the core generalization. Moreover, it is not only buried in paragraphs with a defining function (i.e. which begin with a core generalization) but may be buried in paragraphs whose ''primary rhetorical purposes are 'Description', 'Explanation', or 'Classification' or 'Presenting Information on Experimental Procedures'' (1976:40). They cite a number of examples to demonstrate this, showing where the information is buried. Definitions can be constructed by extracting, combining and reordering information buried in the non-core sentences.

This is the only article we have found which gives any real consideration to the existence of implicit defining information. While the article is largely anecdotal, focusing on a selection of examples, and does not therefore allow for easy generalization, it demonstrates that definitions in text may be expressed differently from the defining formulae found in the textbooks written by Swales and Allen and Widdowson and provides a useful basis for further investigation.

4.5 Darian

Darian focuses on the role of definitions in scientific and technical writing and describes *defining* as follows:

> Defining is best understood as a series of interlocking systems dominated by the semantic system, which interacts with the subordinate syntactic, lexical and typographic systems to produce a broad range of definition formulas. (Darian 1981:43)

Of interest is his assertion that there is a broad range of definition formulae. He distinguishes initially between the concepts of *preliminary* and *formal* definitions. When providing a *preliminary* definition of a term, the author provides a brief explanation that clarifies the meaning of the term in the particular context, if s/he is not ''ready to examine it in depth'' (Darian 1981:42). He cites the following example:

> Further reactions occur when solar radiation encounters the gaseous envelope that surrounds the Earth, THE ATMOSPHERE. (1981:42)

Here, the reader is given what Flowerdew subsequently termed a definition by substitution, i.e. a paraphrase or synonym (cf. Section 4.7.2).

A formal definition (DEF), on the other hand, contains a term (T), a genus or class word (CW), and one or more differentiae or limiting features (LF): T = (LF_1) + CW (+LF_2 + LF_3 + LF_n) (Darian, 1981:45). As there is general consensus among all of the researchers about the essential components of a formal definition, we do not intend to reiterate what has already been said regarding the above but prefer to focus on what is innovative in Darian's work. Darian makes a distinction between the use of *general* and *generic* class words in defining statements. *General* class words include words such as *substance, method, device, process* while *generic* class words include words such as *metal, machine, animal, container*. He suggests that the difference lies in the fact that it is difficult to *visualize* general class words but, to this reader, a more likely distinction is that one group consists of abstract nouns and the other, concrete nouns.

Darian, like Swales, distinguishes between definition types, namely *general* and *specific* definitions; in the former, the class word is likely to be fairly general, whereas in the latter, the class word is likely to be the term which is immediately superordinate to the term being defined, and the term is frequently repeated in the definition. Again, we are not sure how useful this distinction is because it is not in fact the term which is repeated but its immediate superordinate which happens to be the same word as one of the words in the term being defined (e.g. a cocker spaniel is a spaniel which . . .).

Darian has devised a very detailed list of semantic features used ''in framing definitions'' but points out that they are suggestive, not definitive. They are:1) classification/category, 2) limiting feature 1 - usually an adjective before the class word, 3) limiting feature 2—usually a phrase after the class word, 4) level 1 example (species), 5) level 2 example (subspecies), 6) level 3 example (i.e. Sheba in a classification of dogs), 7) coordinate classification, 8) synonym, 9) paraphrase or restatement, 10) antonym/contrast, 11) collocation, 12) connotations, 13) semantic modes. While Darian makes some useful observations, his discussion on the whole is not sufficiently focused or supported to serve as a model for the expression of definition statements in text. No verifiable distinction is made between formal and other definitions. He cites examples for many of his features but there is no evidence to suggest that these examples were not invented. The semantic features which he proposes are simply too vague, and it is hard to imagine how one might apply them in any systematic fashion. He also makes other distinctions which appear to be unnecessary (*general/generic, general/specific*).

4.6 Trimble's definition types

While Trimble's methodology for teaching definition was already discussed in Section 4.4 this section explores, in greater detail, a later publication by Trimble where he distinguishes between different definition types. He has been singled out for discussion because the definition types investigated in chapters eight and nine are based on his classification of definitions. Trimble (1985) distinguishes between simple and complex definitions whereby a simple definition is one which is expressed within a single sentence and a complex definition is expressed in more than one sentence. Trimble proposes three types of simple definition: the formal definition, the semi-formal definition and the non-formal definition.

4.6.1 Trimble's formal definition

The formal definition "is, of course, the well-known equation-like '*Species* = *Genus* + *Differentia'*, usually called 'formal' because of its rigidity of form" (Trimble 1985:75–76). A formal definition gives the reader three kinds of information: the name of the *term* being defined, the *class* to which the term belongs and the *difference(s)* between the term and all other members of the class. The difference(s) constitute(s) the essential characteristics of the term. Formal definitions define words in terms of their *physical description, function, use* or *purpose*. Trimble provides the following example of a formal definition, developed by function description: "An anemometer is a meteorological instrument that registers the speed of wind on a dial or gage" (1985:80). This definition does indeed follow the pattern: A *genus* is a *species* which + *distinguishing characteristic (function*). However, Trimble does not indicate how common this and the other types of formal definition are. Nor does he discuss whether there are other ways of expressing the *genus-species* relation (e.g. apposition). As he suggests that this type of definition got its name from the rigidity of its structure, one might have expected some more information about the structures, their format, tense, mood, classes of information presentation.

4.6.2 Trimble's semi-formal definition

> By definition, a semi-formal definition contains only two of the three basic defining elements: the term being defined and the statement of differences. (Trimble 1985:77)

A semi-formal definition gives the reader two kinds of information: the name of the *term* being defined and the *difference(s)* between the term and the other members of the class. The class is not stated, and Trimble suggests that this is because it is

often assumed by the writer either to be obvious or to be of no relevance to the discussion. He provides the following example of a semi-formal definition: "An anemometer registers the speed of the wind on a dial or gage" (1985:80). Here, the *genus* is indeed absent and the term is defined in terms of its function alone. This is the type of statement which was previously (cf. Section 4.4) described as an implicit defining statement because it is unlikely to be the core statement in a paragraph. Trimble makes no comment about other words which might be used to link the *term* and its *distinguishing characteristic*, whether verbs which introduce function are always function verbs, as in this example, or whether other link words or phrases associated with certain types of characteristics (e.g. *used to* to introduce *function*) can be used.

4.6.3 Trimble's non-formal definition

> The function of a non-formal definition is to define in a general sense so that a reader can see the familiar element in whatever the new term may be . . . . Most non-formal definitions are found in the form of synonyms. (Trimble 1985:78)

A non-formal definition gives the reader two kinds of information: the name of the *term* being defined and another word or phrase having the approximate meaning of the term, or giving an outstanding characteristic of the term, e.g. "An arachnid is a spider" (1985:80). As with the previous examples, some discussion of other means of expressing non-formal definitions might be useful. Trimble makes no reference to the use of the conjunction 'or' or other structures such as 'known as', 'called' which, as we shall see in chapter seven, are often used to introduce non-formal definitions. However, he may not consider these to be the preferred method of expressing non-formal definitions, which would explain, but not necessarily justify, the oversight.

4.6.4 Trimble's complex definitions

Complex definitions are expanded versions of the simple definition and, "characteristically, most expanded definitions are developed in paragraph units and have, as a rule, a simple definition—formal or semi-formal—for their core statement" (Trimble 1985:81). They include "definition 1. by stipulation, 2. by operation, and 3. by explication" (Trimble 1985:81). Definition by stipulation is generally used when the author wishes to set a limit, either "in time, in place, in field, in meaning" (1985:81) and involves the use of hedges such as *mostly, as used in this clause, in information theory.*

Trimble's operational definition "tells the reader what to do in order to experi-

ence—physically and/or mentally—whatever is being defined'' (Trimble 1985:82). Below is an example of an operational definition:

> The sound [f] is a voiceless, labio-dental fricative, formed by placing the lower lip lightly against the upper teeth, closing the upper vellum, and forcing the breath out through the spaces between the teeth or between the teeth and the upper lip. (Trimble 1985:82)

The purpose of Trimble's definition by explication ''is to give the reader new information about the key terms in the original definition'' (1985:82). This may consist of explaining terms which are used in a definition in a previous sentence.

4.6.5 Summary

Trimble has devised his classification of definitions using the criteria of completeness of the information provided (formal, semi-formal and non-formal), the type of information provided (physical description, function, purpose), and the manner in which the information is provided (stipulation, explication, operation). He focuses on genus-species relations and makes no reference to part-whole relations. Most of his examples appear either to have been invented by him or to have been taken from students' work (i.e. from work carried out by non-native speakers of English) or to have been taken from textbooks which teach the art of technical writing; the examples taken from these textbooks also appear to have been invented, for the purpose of illustration. The only authentic examples which he provides are examples of complex definitions, i.e. definitions which are expressed in more than one sentence. If he is in a position to cite examples of complex definitions, one wonders why he has not done the same for his simple definitions. Perhaps simple definitions are far less common than he leads us to believe here, as was already suggested in Section 4.4? We suspect that the complex examples offer a more realistic reflection of the way definition information is actually realized in text. It is not that difficult to *speculate* about how a simple, formal definition might be expressed in text, and Trimble does this very well but it is likely to be far more difficult to *locate* many actual instances of this or other simple definition types in text.

4.7 Flowerdew

Flowerdew (1992a:215) adopts Trimble's three main types of definition (formal, semi-formal and substitution) and proposes one minor type (ostension). Unlike

Trimble, who appears to have sourced his examples either by intuition or by consulting ESP textbooks, Flowerdew bases his classification on a series of lectures in biology and chemistry given by native English speakers to non-native speakers. By comparing how lecturers defined in lectures and how the language of definition was presented in ESP textbooks, Flowerdew discovered that there was a 'great discrepancy' between the two media.

> Whilst definitions in lectures, as this study will show, are subject to much variation, the typical EAP course book presentation of definitions tends to be very prescriptive, presenting a formulaic pattern for students to imitate. (1992a:203)

In defence of ESP textbook authors who are concerned with written discourse, it is hardly surprising that spoken discourse should be quite different. We believe, however, that Flowerdew's observation is warranted in the sense that Trimble, and indeed others, appear to be more concerned with teaching formulaic patterns than with taking usage into account.

The most important feature of Flowerdew's work is that it is corpus-based, and authentic data always seem to reveal more interesting facts than mere conjecture. In his study, Flowerdew examined 329 definitions of 314 terms which had been extracted from a series of science lectures. He documented the way in which each of the definition types was expressed.

4.7.1 Flowerdew's formal and semi-formal definitions

Flowerdew found that the formal and semi-formal definitions in his corpus could be further sub-classified, into the following subcategories, according to the semantic content of the specifying characteristic: a) behaviour/process/function, b) composition/structure, c) location/occurrence, d) attribute/property. He also noted that, contrary to what is suggested by Trimble, the term to be defined does not always come first in the sentence. It may appear at the end of the sentence, e.g.: ''now a photo that we take through a microscope we call a micrograph'' (1992a:210).

4.7.2 Flowerdew's definition by substitution

In a substitution, a word, word-part, phrase, or phrases, with a similar meaning, is substituted for the newly introduced term. There are three types of substitution: synonym, paraphrase and derivation (1992a:211). Flowerdew's definition by substitution is the same as Trimble's non-formal definition except that Flowerdew illustrates the many ways in which non-formal definitions are expressed.

4.7.3 Structure of definitions in the Flowerdew corpus

Flowerdew (1992b) provides examples of each of the definition types. He suggests that:

> formal and semi-formal definitions are most commonly used where the information conveyed by the definition is the main focus of the discourse . . . and substitutions are used most commonly where the definition is not the main focus of the discourse. (1992b:170)

The typical syntactic structure for the formal definition is : NP + copula + NP (including relative clause or other pre- or postmodifier). He provides the following example of a formal definition: ''. . . an element is a substance which cannot be broken down into simpler substances'' (1992b:167). Flowerdew does not specify whether the relative clauses or other pre- or postmodifiers are marked in any way which makes them easily identifiable.

The typical syntactic structure for the semi-formal definition is: NP + copula + NP (without relative clause). Flowerdew provides the following example of a semi-formal definition: ''. . . the circulatory system concerns the movement of blood in all animals . . .'' (1992b:168).

The definition by substitution may have the same syntactic structure as the semiformal definition but

> . . . where this occurs, the second noun phrase is usually less complex. Often, however, instead of the two noun phrases being linked by a copula verb, they are placed in apposition, either explicitly marked, usually by or . . . or marked only by intonation. (1992b:171)

He provides the following examples of definition by substitution: ''. . . it increases its girth or fatness . . .'', ''. . . prey is also captured by the cnidoblasts, the stinging cells of hydra'' (1992b:171).

The term in formal and semi-formal definitions does not necessarily appear at the beginning of the definition and, when it does, Flowerdew suggests that this is because the term has already been introduced in the discourse. According to the principle of end focus, the semantic element with most emphasis (i.e. the most salient) is likely to come at the end of a sentence or clause. This means that the distinguishing characteristic is the most salient of the three elements in a formal definition and the term the least salient, if it comes in initial position. While one would normally expect the term to be the most salient of the three elements in a definition,

> where the structure of term, class, characteristic is employed, the term has very often already been introduced into the discourse and is thus given (as opposed to new) information in the definition itself. (1992b:168)

If the term has not been previously introduced, left dislocation (e.g. "ending zeros / *these* are numbers which have . . ." 1992b:168) may be used to establish the term as given or:

> . . . where the term has not been previously established, the semantic ordering of the definition is reversed, with the term coming at the end, in so-called nominal definition, e.g. 'on the ventral surface of the earthworm there are small projections which are known as the *chaetae* . . . (1992b:168)

4.7.4 Linguistic signalling of definitions in Flowerdew corpus

The presence of definitions can be signalled either by *syntactic* or *lexical* devices. The copula is the most common *syntactic* device in the Flowerdew corpus. Unfortunately, Flowerdew does not provide any classification of the types of copula found in his corpus. Approximately half of the definitions in the Flowerdew corpus are signalled lexically by means of expressions such as *we call / is called / are called / called*. Other phrases such as *or, known as* also occur, but much less frequently. He distinguishes between internal and external devices. Internal lexical devices can be further subdivided into boosters and downtoners. Boosters are linguistic items "that signal clearly the illocutionary force of a speech act" (1992b:172). Examples are the expressions cited above, i.e. *we call* etc., Downtoners can downgrade the force of a definition. Frequently used downtoners are adverbials (e.g. just), modal "can", and non-factive predicators (e.g. *one way of defining a* . . . is) . These are similar to the hedges cited by Trimble (1985).

In addition to the internal devices which signal the presence and mark the salience of definitions, external devices, which Flowerdew calls grounders, may be used.

> A grounder is a statement which precedes and prepares the way for a definition . . . . Grounders . . . familiarize hearers with the term to be defined in anticipation of the definition itself. (1992b:173)

An example of a grounder (grounder in italics) is: ". . . *today we're going to start on Chapter 6 / alkenes /* alkenes are . . . (+ definition)" (1992b:173).

4.7.5 Summary

Flowerdew's classification and illustration of definition types provides a very useful starting point for our work. He appears to have restricted his analysis to statements which were clearly marked as definitions. While he does consider the impact of the use of modals and adverbs on a definition's general applicability, he does not appear to have considered the relevance of the presence or absence of the definite article with the term, which is surprising, because, as Swales suggests, it is generally a useful indicator of the scope of a definition.

4.8 Conclusion

In this chapter, we have attempted to provide an overview of the literature which looks at the realization of definition in text. There is general agreement that a formal definition corresponds to the formula *an X is a Y + distinguishing characteristic* whereby *Y* is a class word or superordinate term. The authors all focus on definition in specialized areas (either in the science, or in law). While many examples are provided to illustrate the points made, only Flowerdew, and, to a much lesser extent, Selinker et al., draw systematically on a corpus for their evidence. Widdowson openly acknowledges that he composed his own passages rather than draw on authentic texts which, he believes, are too complex. We learn that definitions may be expressed either by using a full defining statement or by using reduction. Much of the research documented in this chapter, although generally not corpus-based, served as a useful starting-point for developing a methodology for identifying defining statements in authentic texts (cf. chapters eight and nine). Flowerdew's research on linguistic signalling devices in particular provided a very good basis for identifying other linguistic signalling devices in our corpora. Trimble's classification of definitions also proved to be very useful and especially his discussion of complex definitions; these appear to be much more frequent than simple definitions but receive relatively little attention in the literature.

5 Defining as a Performative Act

5.1 Introduction

In the previous chapter we established that much of what had been written about the expression of definitions in discourse was based on how authors imagined definitions would be expressed in text rather than on evidence drawn from authentic data. Flowerdew was the only one of the authors discussed who consistently used a corpus for his work. His corpus consisted of a series of lectures in biology and chemistry given to non-native speakers of English. Given the communicative context of students learning a specialized subject in a language other than their own, definitions were necessarily going to feature as part of the discourse in his corpus. In this chapter, we would like to look at definitions in other types of discourse and to explore further the notion of definition in other domains. We will suggest that definitions, while not always explicitly signalled as such, function as performatives in the sense defined by Austin. To begin with, we will look at the various categories of performatives which have been proposed by Austin and the felicity conditions which he has specified for the successful fulfilment of performatives. We will suggest that it is possible to distinguish between different types of defining act in much the same way as Austin distinguishes between different types of performative. The performative act of defining can be interpreted either as a defining exercitive or as a defining expositive depending on how the definition is expressed and on who is expressing it. The concept of a defining exercitive will be proposed for situations where new concepts are being described, and definitions are being formulated for the first time. The concept of a defining expositive will be introduced for situations where definitions which already exist are being repeated or rephrased for the purposes of clarification or explanation. A set of felicity conditions for defining performatives in text, analogous to those proposed by Austin for performatives in general, will be presented; these conditions must be met in order for a defining performative to be deemed to be successful.

We will examine means of identifying defining performatives. We will find that defining exercitives may be signalled explicitly, and that the signals can be classified as true performance utterances or hedged performance utterances depending on who is making the utterance. It will be suggested that the status of the person mak-

ing the utterance in his/her subject domain has a bearing on the way in which s/he expresses definitions. Defining expositives are unlikely to be signalled explicitly; they are to be found in dictionaries and in certain text types. A distinction will be made between full defining expositives and partial defining expositives; this distinction will correspond broadly to the distinctions made by others such as Trimble and Flowerdew between formal and other types of definition.

5.2 Austin's performatives

> To utter the sentence (in, of course, the appropriate circumstances) is not to describe my doing of what I should be said in so uttering to be doing or to state that I am doing it: it is to do it . . . The term 'performative . . . is derived, of course from 'perform', the vocal verb with the noun 'action': it indicates that the issuing of the utterance is an action—it is not normally thought of as just saying something. (Austin 1962:6).

Austin (1962) uses the term 'performative' to describe verbs which, when used, invoke some form of conventional procedure and which in themselves constitute some form of action. Such procedures include actions as diverse as the performance of a marriage ceremony, the making of a promise, the issuing of a warning. All of these procedures have what Austin later describes as 'illocutionary force' which means that the speaker, in saying something, is actually doing something. Austin describes five categories of performatives. These are: a) verdictives—the giving of a verdict, by a jury or umpire; b) exercitives—the exercising of power, rights, influence—voting, appointing; c) commissives—promising or otherwise undertaking: they commit you; d) behabitives—related to social behaviour: apologizing, congratulating, condoling, cursing; e) expositives—they make plain how our utterances fit into the course of an argument or explanation: I argue, I concede, I illustrate. This chapter will focus on exercitive and expositive performatives.

5.2.1 Austin's felicity conditions and rules governing performatives

For a performative to be valid a number of conditions must be met. When these conditions do not hold, Austin argues that the circumstances are 'infelicitous' and that the performative is not valid. His conditions are outlined and explained below.

Austin's first rule relates to the procedure: "A1. There must exist an accepted conventional procedure having a certain conventional effect, the procedure to include the uttering of certain words by certain persons in certain circumstances" (Austin 1962:26). A procedure must exist. It must be accepted. The procedure is enacted by the use of certain phrases and it is invoked in certain circumstances. The

acceptance of a procedure means that both the speaker and hearer agree that the procedure exists. The uttering of certain words means that the speaker uses fixed phraseology to invoke a performative. This is clearly the case in legal procedures such as marriage, taking an oath, or religious ceremonies such as baptism, granting of the last rites, or ordination.

His second rule relates to appropriateness: ''A2. the particular persons and circumstances in a given case must be appropriate for the invocation of the particular procedure invoked'' (Austin 1962:15). Here, the particular persons must have the authority to invoke the procedure and must do so in the appropriate circumstances. One of the examples which Austin uses is that of appointing someone to a job. If the person making the appointment does not have the authority to do so, the person is not the appropriate person for invoking the procedure and the procedure is therefore null and void. Equally, if someone has already been appointed to the position, and the vacancy no longer exists, the appropriate circumstances do not apply.

His third rule relates to correct execution of the procedure: ''B1. The procedure must be executed by all participants correctly'' (Austin 1962:6). The participants must use the correct phraseology to invoke a procedure. Failure to do so results in flaws. Many performatives involve the use of precise formulae; deviation from these can render the performative incomplete or null and void.

His fourth rule relates to completion of the procedure: ''B2. The procedure must be executed by all participants completely'' (Austin 1962:15). It is not sufficient for a speaker to invoke a procedure. It must be accepted by the hearer and completed by the hearer if the performative so requires. Non-completion of a procedure results in hitches. In a marriage ceremony, for example, there will be a hitch if the man says ''I do'' and the woman says ''I do not''.

The fifth and sixth rules refer to the frame of mind of the participants and to their consent to participate in the performative.

> 1. Where, as often, the procedure is designed for use by persons having certain thoughts or feelings, or for the inauguration of certain consequential conduct on the part of any participant, then a person participating in and so invoking the procedure must in fact have those thoughts or feelings, and the participants must intend so to conduct themselves, and further
> 2. must actually so conduct themselves subsequently.'' (Austin 1962:15)

These two conditions relate to the sincerity, intentions and good faith of the participants.

When the act is not achieved, because of the failure of any one of the four rules A1-B2, Austin refers to these as misfires. When the fifth and sixth rules fail, the act is in fact achieved but the procedure is somehow abused because one or other of the

participants is not sincere in his/her intentions. Austin calls these infelicities abuses. When an act misfires, Austin deems it to be null and void. When an act is abused, Austin says that the act is "'professed' or 'hollow' rather than 'purported' or 'empty' and as not implemented, or not consummated, rather than as void or without effect" (Austin 1962:16).

5.2.2 Austin's criteria for identifying performatives

Austin discusses how one might identify performatives and tries to establish whether there might be grammatical or lexicographical criteria for identifying them. His conclusions are fairly tentative. What follows is a brief summary of the issues which he addresses.

Austin suggests that for a formalized explicit performative to be valid, it must be expressed in the first person singular, indicative active, and present tense. He finds it more difficult to specify the grammatical criteria for less explicit performatives because there are many situations where the performance utterance is not expressed at all. In such situations if the utterance passes the following test, it is deemed to be a performative: "any utterance which is in fact a performative should be reducible or expandible, or analysable into a form, or reproducible in a form, with a verb in the first person singular present indicative active" (Austin 1962:61). An attempt to use vocabulary as a test of performatives was not very successful because performatives can exist without the 'operative words' (e.g. instead of 'dangerous bull' we may write 'bull'); 2) the operative word may be used but not in a performative sense (e.g. "I may say 'you were guilty' or 'you were off-side' or even 'you are guilty' when I have no right to pronounce you guilty or offside).

5.3 Defining as a performative

Austin includes the act of defining in his set of performatives: "When we use the formula 'I define x as y' we have a transition to a performance utterance" (Austin 1962:65). Defining is therefore to be construed as a performative action. When Austin included defining in his set of performatives, it is very likely that he was thinking primarily of defining as a clearly signalled utterance always prefaced by formulae such as 'I hereby define'. We would like to extend the notion of defining as a performative beyond such clearly signalled utterances and to suggest that it can also be used to describe certain metalanguage statements in discourse. Some such metalanguage statements will be readily identifiable as defining performatives because they will be clearly signalled. However, there are many others which we would choose to classify as defining performatives which will not be signalled in

such an obvious way. We hope to demonstrate that this second category of metalanguage statements has the same function as those in the first category. However, before proceeding further with a discussion of the different classes of defining performatives, we propose first to look at the felicity conditions which must hold if a statement is to be considered as a defining performative.

5.3.1 Felicity conditions for defining performatives

The felicity conditions described below, are analogous to those specified by Austin for performatives in general. What we have done is to re-examine each of Austin's rules with a view to ascertaining how they can be interpreted for defining performatives alone. Austin's first and second rules stipulate the use of a conventional procedure by certain competent persons in certain circumstances. In chapter one, we established that there were, broadly speaking, three types of communicative setting in which one was likely to find terms being used as such rather than as ordinary language words. These were a) communication between experts, b) communication between experts and initiates and c) communication between experts and the uninitiated. We called these settings 1, 2 and 4 respectively. It was noted that the author-reader relationship and text purpose were important factors in determining whether a particular text or set of texts would be a suitable term resource. We suggested that texts which had a didactic or informative purpose written by authors with the required level of expertise for an audience which had a professional interest or need to read these texts were most likely to contain terms. We believe that the same text types will also contain metalanguage statements functioning as defining performatives because authors writing within these communicative settings have a tendency to define some of the terms which they use. In the case of defining performatives, therefore, the communicative setting will determine whether the 'circumstances' are appropriate for the execution of a defining performative; settings 1,2 and 4 proposed in Chapter 2 provide the appropriate circumstances. The 'certain competent persons' who are to execute the performative are authors who are deemed to have the required level of expertise to write about their subject and who write texts which fit into settings 1, 2 and 4. The 'conventional procedure' involves using statements which can be reduced, expanded, analysed or reproduced in a form to allow them to be interpreted as definitions. Austin's rules 1 and 2 for performatives in general could therefore be rephrased as follows for defining performatives: defining performatives must be expressed or formulated by competent authors writing texts designed for communicative settings 1,2 and 4.

Austin's rules 3 and 4 refer to the correct execution and completion of the procedure which means that the speaker must use the correct phraseology and the hearer must accept it. As already suggested, there are two types of defining performative,

one which will involve the use of specific phraseology (e.g. *I define*) to signal that a definition is to follow, and another which will not be signalled explicitly but will involve the use of certain syntactic structures. These syntactic structures are explored in depth in chapters seven and eight. If the defining performative is explicitly signalled or is expressed using one of the specified syntactic structures, the correct phraseology is deemed to have been used. If the procedure is being enacted within one of the specified communicative settings, it is assumed that the reader will accept it. Austin's rules 3 and 4 might be rewritten as follows for defining performatives: defining performatives must be expressed or formulated using specific phraseology, and in the settings already defined.

The felicity conditions which we have specified here are very general and are designed merely to specify the general framework within which one is likely to find defining performatives. More detailed discussion of specific felicity conditions is provided in chapters seven and eight.

5.4 Distinguishing between types of defining performative

In this section, we propose to make a distinction between what we have chosen to call 'defining exercitives' and 'defining expositives'. When a definition is formulated and expressed for the first time, we describe the act as a defining exercitive; when an existing definition is being reiterated or rephrased, we describe this act as a defining expositive. The 'defining exercitive' is very similar to Austin's exercitive and the 'defining expositive' is similar to Austin's expositive.

5.4.1 Defining exercitives

The following examples may serve to illustrate what we mean by defining exercitives. When people create new concepts and coin words or phrases for those concepts, they are likely to define them. When they do so, we suggest that they are engaging in an original defining act. This is most likely to occur in texts which match the communicative settings 1 and 2, i.e. texts where experts are writing for their peers (setting 1) or for people who already have some knowledge of the field (setting 2). Another situation is where a concept and word already exist within a particular subject domain but an author wishes to assign a new meaning to that word, thereby altering the definition and *doing* in the same sense as in the first example. This approach is quite common in academic papers (i.e. texts which meet the requirements of communicative setting 1) and it is one to which we will return in our discussion of true performance utterances, cf. Section 5.4.1.2.1.

The defining exercitive is similar to the exercitive defined by Austin. An exercitive is:

> the giving of a decision in favour of or against a certain course of action, or advocacy of it. It is a decision that something is to be so, as distinct from a judgement that it is so. (Austin 1962:155)

> "Exercitives commit us to the consequence of an act, for example of naming." (Austin 1962:159)

The person who is formulating a definition for the first time, as in the examples given, and using a statement such as 'I hereby define x as y' is indeed giving a decision in favour of a certain course of action, namely that some word is henceforth to be understood as now defined. The speaker is committing himself or herself to the consequence of abiding by that definition. We think it is possible to argue that this act is a performative in the sense defined by Austin; the authors are doing, and the statements which follow the performance utterance have the characteristic of being original, of being formulated for the first time. This is an important element in assessing the validity of defining exercitives. We would suggest that a definition can be considered to be a defining exercitive only on the occasion when it is first invoked; all subsequent utterances of that definition must be viewed as clarifications, expositives. We would further suggest that there are two types of defining exercitive: the individual, and the consensual.

The individual defining exercitive involves the identification, naming and description of a new concept by an individual. To qualify as an individual defining exercitive, the definition must be provided by the person who has identified and named the concept. This type of exercitive is likely to occur in the text types described under communicative setting 1 where author and reader are assumed to have a similar level of expertise; such text types include research papers and learned academic texts and journals; it may also occur in communicative setting 2 where the reader is assumed to have a lower level of expertise than the author.

The consensual defining exercitive is a definition provided by an authority such as a standardizing body or professional association which stipulates how a concept is to be understood and used within a particular context. The formulation of a consensual defining exercitive involves the naming and description of a concept, and responsibility for the definition lies with a group of people rather than with one individual. We expect to find consensual definitions in prescriptive documents (i.e. standards) produced by standardizing bodies and professional associations.

Those who formulate individual defining exercitives may or may not use a perfor-

mance utterance (i.e. 'I hereby define') to alert the reader to what they are doing. Consensual defining exercitives are less likely to be preceded by a performance utterance because the status of the publications in which they appear (i.e. standards or prescriptive functions) effectively fulfils that function. The criterion for assessing whether or not a definition can qualify as a defining exercitive will be, as Austin suggests, whether the statement is ''reducible, or expandible, or analysable into a form, or reproducible in a form, with a verb in the first person singular present indicative active'' (Austin 1962:61).

When defining exercitives are preceded by a performance utterance, we propose to call these explicit defining exercitives and when the performance utterance is omitted, we propose to call them implicit defining exercitives. We intend first to discuss briefly implicit defining exercitives and to offer some explanations as to why and how they are used, and then to look at explicit defining exercitives.

5.4.1.1 Implicit defining exercitives

We have said that defining exercitives do not have to be prefaced by a performance utterance which would clearly indicate that a definition is about to follow. The performance utterance may be omitted for a number of reasons. For example, it may be because it is not customary to flag definitions in certain contexts. A scientist who is an acknowledged leader in his/her field may not preface the naming and definition of new concepts with a performance utterance. The fact that the person has in fact formulated a new definition is only established by subsequent reference to that fact by the scientist him/herself or by other people who wish to refer to the scientist's concept and definition. It is difficult to imagine someone such as Stephen Hawking, for example, explicitly marking the introduction of a new theory with a bold statement such as 'I hereby define, declare'. His position of authority within his field exempts him from the need to signal explicitly that he is naming and defining concepts.

Another reason for omitting a performance utterance may be that an author is reluctant to declare openly that s/he is defining her/his own terms and choose instead to define a concept without explicitly stating that that is what s/he is doing, preferring to test audience reaction before publicly committing her/himself. The audience will determine whether or not the definition is acceptable. The author, by not marking the definition, somehow abdicates or postpones responsibility for commitment to the definition. If the definition is accepted by the audience, the author is, however, likely to be given credit for the formulation of the definition and may also explicitly claim credit for it in subsequent publications. The difference between this and the first example lies with the status of the author and the type of text in which these definitions are likely to occur. The first, as previously suggested, may be used by acknowledged experts writing in learned texts (setting 1). The second is

much more likely to be used by people who, while they have the expertise required to write about their subject and write within the same communicative setting, do not have the sort of reputation enjoyed by experts such as Stephen Hawking.

One further explanation for failing to use a performance utterance is that an author may simply not be aware at the time that s/he is in engaging in an act of defining. This appears to be quite common in fields where the terminology is very new and still being clarified. The field of corpus linguistics is perhaps a good example because its terminology is still evolving. It has drawn on other schools of thought for its terms and in many cases, redefined the concepts for its own purposes. Descriptions which appear to be clarifications of existing terms frequently evolve to become definitions in their own right. For example, Sinclair's (1987:110–115) discussion of the principle of idiom in his chapter on collocation leads to the definition of a new concept and the coining of a new term, the 'idiom principle', to which we can find no previous reference in earlier publications.

5.4.1.2 Explicit defining exercitives

Austin has suggested that defining performatives will be signalled by an utterance such as 'I hereby define'. While we have found no evidence of this particular utterance in the corpora and texts which we have examined, we have identified other methods of signalling defining exercitives. These other methods fall into two categories, and we have chosen to call them 1) the true performance utterance which indicates clearly that the author is taking a stand and prepared to stand by his/her definition, and 2) the 'hedged' performance utterance which allows the author to abdicate some of his/her responsibility for the more general applicability of the definition.

5.4.1.2.1 True performance utterances. A true performance utterance is one which shows that the author is unreservedly committing himself/herself to the definition. The author is not afraid to be challenged on it. There may be a link between a person's position within a field and the degree to which they are prepared to take a stand by explicitly marking what they are saying. Given their nature, true performance utterances are more common in academic or research texts than in the corpora on which our discussions in later chapters are based. Consequently, the examples provided here are drawn from a number of academic texts dealing with topics in linguistics.

Here, we propose to look at four examples of true performance utterances and, in the next section, at some examples of hedged performance utterances in an attempt to clarify what the distinctions between them might be.

> it seems clear that to utter the sentence . . . is not to describe my doing of what I should

> be said in so uttering to be doing or to state that I am doing it: it is to do it . . . . What are we to call a sentence or an utterance of this type? I propose to call it a performative sentence, or a performative utterance, or, for short, 'a performative'. (Austin 1962:6)

Austin uses the performance utterance 'I propose to call it' , then introduces two (synonymous) terms for the concept which he has just described, and finally proposes an abbreviated form of these terms. He is clearly signalling the identification, naming and description of a new concept and taking full responsibility for this action.

Let us look now at the following quotation from Sinclair: ''I should like to widen the domain of syntax to include lexical structure as well and call the broader domain structure . . . . I shall define structure as any privileges of occurrence of morphemes'' (Sinclair 1991:104). This quotation is interesting because, in the one sentence, Sinclair is redefining one concept and naming a new one. He appears to be a little diffident about redefining the existing concept, hence the use of the polite form 'I should like to' to introduce the re-definition. While he would like the domain of syntax to include lexical structure as well, he is leaving open the possibility that this proposal may not be accepted. In fact the use of 'I should like to' is more an indication of a proposal for a definition than an indication of an actual definition. When it comes to defining a new concept, however, all signs of diffidence have disappeared. Here, Sinclair simply states 'I shall define' to introduce a new definition. This is a true performance utterance and the author is clearly taking responsibility for it.

In the following quotation, the author is using a true performance utterance, i.e. 'I shall call', and taking responsibility for the definition. ''When a is node and b is collocate, I shall call this downward collocation—collocation of (a) with a less frequent word (b). When b is node and a is collocate, I shall call this upward collocation'' (Sinclair 1991:115–116).

One further example of a true performance utterance is to be found in a definition of *clause relation* provided by Winter. We have chosen to cite it in full because it is interesting for a number of reasons.

> My latest definition of Clause Relations takes the clause as the largest unit of meaning in the sentence, so that relations between sentences are really relations between their constituent clauses. It is as follows:
>
> A Clause Relation is the shared cognitive process whereby we interpret the meaning of a clause or group of clauses in the light of their adjoining clause or group of clauses.
>
> Where the clauses are independent, we can speak of 'sentence relations'. (This revises Winter 1971, 1974, 1977, 1979 and 1982.) It is in no way incompatible with Hoey (1983:19):

> A clause relation is also the cognitive process whereby the choices we make from grammar, lexis and intonation in the creation of a sentence or group of sentences are made in the light of its adjoining sentence or group of sentences. (Winter 1977:91)

We have here an example of someone reformulating a definition which he has already formulated and reformulated. The author claims responsibility for the definition (''my latest definition'') and expressly states that the current version supersedes all previous definitions. He also acknowledges that the concept has been defined by others (i.e. Hoey) and claims that their definitions are not incompatible. Winter cites Hoey's definition, thereby widening the scope of his own definition and incorporating Hoey's definition.

What is noteworthy about the above is that Austin and Sinclair, and perhaps Winter too, are acknowledged as having broken new ground in their respective fields. Has their achievement been recognized because of their willingness to take unreserved responsibility for their claims? Or is it perhaps the fact that they are recognized as being leaders in their field which gives them the confidence to make explicit claims such as those cited above? We would suggest that either of these explanations is possible and that true performance utterances are only likely to be used by people who are acknowledged in their field.

5.4.1.2.2 Hedged performance utterances. We have identified two types of hedged performance utterance:1) indicator of tentativeness, 2) indicator of scope. The indicator of tentativeness indicates that the author is being tentative about his/her claims; the author may be leaving the definition open to challenge and/or leaving him/herself the option of refining the definition further at a later stage. An example of this type of hedge is: ''A phrase can be defined for the moment as a co-occurrence of words which creates a sense that is not the simple combination of each of the words'' (Sinclair 1991:104). Here, 'for the moment' seems to suggest that what we are being offered is not the definitive version. It is complete for now but may be refined at some future date. The use of the modal allows the author to distance himself from the definition, thereby attenuating his commitment to it.

The indicator of scope is used to specify that a definition has local reference only and is not necessarily to be interpreted as a generally applicable rule. The effect of this device is that the author avoids controversy. The use of the hedge functions as a pre-emptive action which guards against possible challenges by others. Examples of this type of hedge are:

> coherence: As used in this book, coherence is a quality assigned to text (q.v.) by a reader or listener, and is a measure of the extent to which the reader or listener finds that the text holds together and makes sense as a unity. (Hoey 1991:265–266)

text: This term is used in two ways in this book . . . . (Hoey 1991:269)

Here, Hoey is specifying that the scope of his definitions does not go beyond the book in which they are used. The reader surmises that the same terms have been defined elsewhere previously but may be willing to accept the author's redefining of the terms once the author has indicated that the refinements are not necessarily applicable beyond the covers of the book.

5.4.2 Summary

In this discussion of defining exercitives, we have shown that it is possible to distinguish between individual defining exercitives and consensual defining exercitives, the distinction being whether the author is an individual or a group. We have seen that defining exercitives are not always explicitly signalled in text. We have looked at both explicit and implicit defining exercitives and have suggested that when defining exercitives are prefaced by a performance utterance, this may tell us something about the status of the author in his/her academic community. Authors who command considerable respect within a specialist community are much more likely to use true performance utterances than authors who are not yet sufficiently well known to expect unquestioning acceptance from their readers. This latter group may choose to use hedges to indicate a certain degree of tentativeness or to restrict the scope of their claims. For example, theses produced by postgraduate students will contain many examples of hedges but are very unlikely to contain true performance utterances.

5.4.3 Defining expositives

There is a second category of defining performative which one encounters in text where the definition is not being formulated for the first time but is being rephrased or repeated for the purposes of clarification or explanation. We chose to call this type of defining act a defining expositive. As Austin explains, expositives are ''used in acts of exposition involving the expounding of views, the conducting of arguments, and the clarifying of usages and of references'' (Austin 1962:161).

The following examples should serve to illustrate what we mean by 'defining expositive'. When a concept and its term already exist, an author may choose to reiterate the definition of that concept for the information of the audience. This typically occurs in textbooks where authors frequently define terms for the benefit of their audience and would correspond to the communicative settings 2 and 4. In other words, the authors are either addressing people who already have some knowledge of a domain but are seeking to further that knowledge or they are addressing beginners who require an introduction to the basic concepts of a particular subject

domain. Unlike defining exercitives, the definition is not being formulated for the first time and consequently, the authors are in fact reporting rather than doing. As many terms are polysemous in the sense that they may have different meanings in different subject domains or they may have a terminological meaning and a general language meaning authors may sometimes feel the need to specify which meaning is designated. This type of situation also arises in texts which fit communicative settings 2 and 4. Here again, authors are simply reiterating definitions which already exist. This type of definition is also quite common in dissertations written by students where they define terms which they use in order to demonstrate that they have fully understood the concepts which they are discussing. However, such examples do not qualify for consideration here because the felicity conditions are not met; the text type 'dissertation' is not included in any of the communicative settings specified in our felicity conditions.

We propose to classify the first two examples as defining expositives. Our reason for doing this is that they involve the clarification of usage; they do not in themselves define; they explain. We have already suggested that a defining performative can be considered as a defining exercitive only on the occasion when it is first invoked; all subsequent utterances of that definition must be viewed as clarifications, expositives. It is more difficult to apply Austin's criterion of originality to defining expositives because they involve the repetition or rephrasing of a definition which already exists. However, the making of the statement, as opposed to the content of the statement, can be construed to be original in that it is the speaker's own and in this sense fulfils Austin's criterion of originality.

We believe that defining expositives are to be found not only in text but also in dictionaries; we propose to call the former text defining expositives and the latter dictionary defining expositives.

5.4.3.1 The realization of defining expositives in dictionaries

Here, we intend to explain why we have chosen to classify dictionary definitions as defining expositives, rather than as defining exercitives. We wish to demonstrate that dictionaries are *rephrasing* (i.e. clarifying usage) rather than *defining* in the sense of *prescribing* as described previously for individual and consensual defining exercitives.

The best way to illustrate this is to look at how dictionaries are compiled. Let us take, for example, the Cobuild dictionary. In very simplistic terms, the compilation of this dictionary involves the collection of two categories of words:1) lists of general words which appear in text, and 2) lists of terms which are also used in general language, and for which definitions already exist.

The definitions of general words are arrived at by examining, analysing and documenting how these words are used. Reports on the usage of each of the words are produced. These reports are edited to become the definitions which one finds in the

dictionary. The lexicographer is not engaging in an exercitive act in the sense that the content of the definitions produced is not in any way unique or original. While the definition itself may be unique because it is being expressed in a new way, the content is a summary of the way in which the word is already being used. While other dictionaries, unlike Cobuild, do not use textual evidence for the formulation of definitions and rely on intuition or other means, we suggest that they too are merely *reporting* rather than *defining* for the first time in the sense in which we have defined *defining* for the defining exercitive.

While we do not know where the definitions of technical terms which one finds in general language dictionaries such as Cobuild are sourced, we would suggest that the definitions are likely to be versions of existing definitions, definitions which have been agreed by an appropriate authoritative body and are rephrased for a more general audience.

When we consult a dictionary, we expect to find answers to questions such as 'what does x mean?' We are making the assumption that the information provided will be true and that it will be structured according to certain conventions. Thus, the felicity conditions specified in Section 5.3.1 are met. We do not generally stop to consider whether the definitions provided are original or consensual, in the sense in which defining exercitives are. We are looking for clarification of usage and of reference and this is why dictionary definitions should be classified as defining expositives rather than as defining exercitives.

5.4.3.2 The realization of defining expositives in texts

Defining expositives involve the reiteration of an existing definition for the purpose of clarification of usage or reference. It seems likely therefore that such performatives will be found in particular in texts where the author-reader relationship is not equal in terms of competence. The ideal candidates for such performatives are texts where the purpose is to impart information, to present existing concepts and to clarify their meaning, i.e. texts which fit communicative settings 2 and 4.

We may also occasionally find defining expositives in texts where the author assumes that the reader has a similar level of expertise (i.e. communicative setting 1, expert-expert communication). In such cases, defining expositives may be used to remind the reader of a concept, to stipulate that when the author uses a certain term, s/he is using it in the same way as, for example, her or his colleagues in another country.

5.4.3.2.1 Identifying defining expositives in text. A defining expositive may be explicitly signalled in texts where the author-reader relationship is more or less equal; in such cases, the author may preface the definition by referring to the original author of the definition, the inventor of the concept. However, it is much more

common for the defining expositive not to be prefaced in any way. The only means of identifying them is by discovering how they are expressed in text. One of the main objectives of this book is to identify and describe the syntactic structures which signal the presence of such performatives. As chapters seven and eight describe and exemplify these structures in detail, we propose merely to introduce the subject in this section, reserving more detailed discussion for the later chapters.

One method of recognizing that a defining expositive is being provided is when phrases such as *is/are defined as, denote(s), consist(s) of, comprise(s*) are used. Such phrases are used when authors have introduced a term which they believe to be unknown to their readers. The term is introduced in one sentence and explained in the following one. Another method of recognizing defining expositives is when phrases such as *is/are known as, is/are called* are used; these tend to be used after the definition has been provided, either in the same sentence or in the following one. The definition is provided first and the word which is being defined is then introduced (note: these are the types of structures already documented by Flowerdew). When these and other signals occur in text, the text segment in which they occur may consist of a defining expositive. However, as chapters seven and eight will demonstrate, these signals alone are not sufficient for identifying defining expositives because the statements retrieved will include many which are not metalanguage statements, i.e. not defining expositives. Consequently, additional conditions have to be specified in order to refine the retrieval process.

5.4.3.2.2 Identifying partial defining expositives in text. Besides the defining expositives described in the previous section where what is provided corresponds to Trimble's formal definition, there are others which provide what correspond to Trimble's semi-formal and non-formal definitions (cf. Section 4.6 for discussion of Trimble's definition types). We have chosen to call these partial defining expositives. Partial defining expositives may name a term without specifying its superordinate; they may name a term without specifying a distinguishing characteristic. They may simply provide information about synonyms or the correct term to be used in a particular context. A set of felicity conditions for retrieving partial defining performatives has been devised; the conditions are explained and exemplified in chapters seven and eight.

5.5 Conclusion

In this chapter, we suggested that Austin's description of performative verbs could be used as a basis for classifying definitions in text. We have suggested that the text types which were suitable sources for terms were also likely to contain definitions

of terms. We have classified definition statements in text as either defining exercitives or defining expositives. Defining exercitives are frequently prefaced by what Austin terms performance utterances and these utterances may be what we have called true or hedged performance utterances. The use of either of these tends to indicate the author's position in relation to the information being provided. Defining exercitives are more likely to be found in academic texts, i.e. in communicative setting 1. Defining expositives, on the other hand, are more likely to be used in communicative settings 2 and 4 and involve the reiteration or rephrasing of definitions which already exist. Defining expositives tend not to be prefaced by performance utterances but the way in which they are expressed alerts us to their presence. Chapters seven and eight deal in depth with the semi-automatic recognition of defining expositives in corpora.

6 Retrieval of terms from the corpora

6.1 Introduction

As noted in chapter two, three collections of texts or special purpose corpora have been selected for investigation in this book. They are the International Telecommunications Union (ITU) corpus (4.7 million words), the GCSE corpus (1 million words) and the Nature corpus (230,000 words). These particular corpora were selected because they have the attributes which we consider necessary for the type of investigation which we have undertaken. In particular, each corpus corresponds to one of the three communicative settings where we expect to find a high density of words functioning as terms rather than as part of general vocabulary. The three settings in question are 1) communication among experts (Nature corpus), 2) communication between experts and initiates (ITU corpus) and 3) communication between experts and the uninitiated (GCSE corpus). An important element of our investigation is the identification and retrieval of terms from the three corpora. It is not the aim of this book to design and implement a high-quality term identification and retrieval system which might result in an exhaustive list of all terms in the corpora but rather to identify some of those terms for which some type of explanation is provided, either in the form of a definition, the specification of a term's superordinate or in the form of synonyms. We need to know what terms look like in order to be able to match them with the patterns which will be specified for retrieving metalanguage statements.

This chapter commences with a brief overview of some of the research which has already been carried out into the automatic identification and retrieval of terms. It continues with a description of how we approached the problem; we started with a manual analysis of each of the corpora and produced term pattern specifications for each corpus based on our analysis. Our objective, in carrying out this task, was to identify as many term patterns as possible, and, using a pattern matcher, to retrieve occurrences of these patterns in the three corpora. This approach inevitably led to the retrieval of many non-terms. We therefore needed to refine the output of the first stage by specifying restrictions which would retrieve only those candidates which were without doubt terms. This refinement process is described in the second half of the chapter. We accept that the process as described is inadequate in some re-

spects and that perhaps not all term formation patterns have been identified but wish to emphasize that we are more concerned with retrieving terms which are partially or fully explained within the corpora than with producing an exhaustive list of all terms in the corpora.

6.2 Previous research into automatic identification and retrieval of terms

The automatic identification of terms has already engaged the minds of a number of researchers, especially those working in information retrieval and natural language processing. The identification of terms in telecommunications texts is well documented, by Béatrice Daille (1994) in her PhD thesis on the topic, by researchers at Dublin City University and UMIST as part of the EU funded Eurotra research project. Yang (1986) devised a technique for identifying scientific and technical terms in a scientific English corpus. At the University of Surrey, work on terminology retrieval has been undertaken in a number of fields including automotive engineering and mammography (Ahmad, Davies, Fulford, Rogers 1994). Jacquemin and Royaute (1994) have worked with the MEDIC corpus which is a bibliographical medical corpus. Bourigault, Gonzalez-Mullier and Gros (1996) have developed LEXTER, a tool for terminology extraction.

Research has generally focused on examining term formation patterns which occur in corpora with a view to tagging the corpora and retrieving term candidates. In a preliminary phase, a manual analysis is carried out to identify the composition of terms in a corpus and a list is drawn up of all possible combinations. These combinations are then used as input to retrieve all term candidates. Nkwenti-Azeh (1992) attempted to identify potential terminological units using a positional and combinational approach. He found that:

> the positional approach removes the need to comprehensively mark up the terms of an . . . input text, especially where we are dealing with a circumscribed corpus: terminological units occurring in the text can be reconstituted if the positions of their elements have been specified in the positional database. (1992:19)

Jacquemin and Royaute (1994) who were in fact more interested in retrieving term variants than actual terms used an existing set of terms and an analysis of the head-modifier relations to establish whether other syntactic patterns containing the same head and modifier could be classified as a variant of a term contained in their list.

Yang used frequency and collocational patterns to identify terms. He states: ‘‘Since terms are highly subject matter specific, it is possible to identify single-

worded terms on the basis of their frequencies of occurrence and distribution. Multi-worded terms are identified on the basis of their collocational behaviour'' (1986:93).

Daille (1994) combined morpho-syntactic and statistical approaches to extract term candidates. She focused exclusively on binary term formation patterns (e.g. adj+noun) in order to write a program for retrieving all such patterns from her corpus. Like Yang, Daille used frequency as an important criterion for assessing the eligibility of a term candidate. When the frequency criterion is used, this means that a term candidate must occur a certain number of times in a corpus before it is considered. The problem with this approach is that it ignores the fact that it is not uncommon for terms to occur infrequently. This may be because the corpus being used for the search is not sufficiently large or because the term in question is more usually referred to in a variant or abbreviated form. We would argue that low frequency should not preclude a term candidate from being considered.

Bourigault et al. (1996) also use a morpho-syntactic approach to develop a noun-phrase extractor which is applied to a wide range of texts. Unlike other researchers who use a morpho-syntactic approach which is based on ''the possible grammatical structures of complex terms'' (1996:772), Bourigault et al. use the ''grammatical configurations which are known not to be parts of terms'' (1996:772). ''The basic principle is then to split the text by locating these potential boundaries, between which noun phrases likely to be occurrences of terms are isolated'' (1996:772). The list of candidate terms which is extracted using the noun-phrase extractor is passed on to a group of experts who decide on the validity of the candidates. The LEXTER system developed by Bourigault and his colleagues is designed to be domain independent.

While the approaches adopted in the research referred to above, and used in a modified form for our initial manual analysis, does indeed provide a useful starting point, it also allows for the inclusion of many words or phrases which are not actually terms. For example, if the pattern *adj+noun* has been specified as a term pattern, all occurrences of modified nouns, regardless of their status, will be included, resulting in the retrieval of a far greater set than is desirable. What will be required is an additional automatic refinement which could greatly reduce the number of non-terms retrieved.

6.3 Identification and retrieval of corpus specific term formation patterns

We are not convinced that there is any real advantage in attempting to devise a list of domain independent term formation patterns which would be valid for our three

corpora. We believe that it is very likely that term formation patterns may vary considerably from one corpus to another and that what might qualify as a term pattern in one corpus might simply be a general language pattern in another. We opted, instead, to produce a set of corpus-specific term formation patterns for each of the corpora. To begin with, we tagged each of the corpora using the CLG tagger devised by the Corpus Linguistics Group at the University of Birmingham; a list of the tags used for our analysis is provided in Appendix I. We then carried out an initial analysis of each of the corpora, using a number of linguistic signals such as *denotes, i.e., e.g.*, as the search key words. We believed that many of the words or phrases which co-occurred with these signals would be terms. Using the material retrieved in this manner, we were able to devise sets of term formation patterns for each of the corpora. As the linguistic signals which we were using did not occur in the Nature corpus, the patterns were more difficult to identify with any degree of certainty; consequently, they are based on observation.

6.3.1 Tag Sequence Pattern matching program

The sets of term formation patterns produced on the basis of the initial manual analysis were used as input for a tag sequence pattern matching program. This program was kindly written for us by Oliver Mason of the Corpus Linguistic Group at the University of Birmingham. The pattern matching program takes as input the term formation patterns which consist essentially of sequences of tags. For example, a typical term formation pattern might consist of the following tag sequence: adj+noun+noun. The program builds up an internal representation of each of the tag sequence patterns. The patterns are then matched against the input streams, i.e. against the tagged corpora. The output from the pattern match procedure consists of the pattern number, the tag sequence and all words or phrases which have been matched. These words or phrases are what we call term candidates.

6.3.2 Decoding the pattern match specifications

Each tag sequence pattern file is based on a different term formation pattern and contains an unlimited number of lines. The specifications for all of the patterns examined in each of the corpora are listed in Appendix A. The tag sequence pattern file allows for two types of refinement. The first of these allows the user to specify alternatives using a pipe. Thus, in the following tag sequence pattern: DT JJ NN|NNS, the use of the pipe allows us to retrieve all singular and plural occurrences of the pattern determiner+adjective+noun(s). The second refinement allows the user to exclude certain items. Thus, in the following tag sequence pattern: !JJ !DT JJ NN NN|NNS, the use of the exclamation mark allows us to specify that the

pattern adjective + noun + noun(s) may not be modified by a determiner or by a second adjective.

If more than one tag sequence pattern is specified in any of the files, the match is carried out in pattern order, and each word in a sentence can only match once. Consequently, if there is an overlap between patterns within the one file, the longer patterns must be specified first, and shorter patterns must follow in descending order. For example, if the two patterns adj + noun + noun(s) and adj + noun(s) are specified in the one file, the longer pattern must come first. As the entire corpus is processed separately for each tag sequence pattern file, there will inevitably be some overlap between some of the sets retrieved.

6.4 Retrieval of term candidates

Here, we provide the term formation patterns for each of the corpora, together with examples. As already noted, the actual specifications for each tag sequence pattern file used to retrieve the patterns are listed in Appendix A, and the number which appears in column three (P.) of the tables indicates which tag sequence pattern file was used. Column four indicates the number of occurrences of each pattern in the corpus. Where { } brackets are used, this indicates that this is the aggregate of occurrences for the lines marked by { } in the tables.

6.4.1 Retrieval of term candidates from the ITU corpus

As the examples show, term candidates can have up to four components, excluding determiners, but instances of 4-component terms will be rare and, frequently, what appear to be 4-component terms are in reality either modified 3-component terms, or terms which have a generic class word such as *system, procedure, function, method* (cf. *path error monitoring function* above) as the head word in the term unit. Single word terms are the most common, followed by two word noun+noun and adj.+noun combinations.

At first glance, some of the terms listed in Table 1 (e.g. *window, effective call*) might not strike one as being particularly 'technical' but we know that these particular examples are indeed terms because they co-occurred with the linguistic signals which we used for identifying terms. However, it is also true that many other term candidates retrieved using the automatic process described above may not actually be terms, hence the need for the refinement procedure which is described in Section 6.5.

On analysing the output from the pattern match files, we discovered that the CLG tagger frequently assigned the incorrect tag to words which were ambiguous on a

Table 1. ITU corpus

Pattern	Examples	P.	Occ.
–det+adj+noun+noun(s)	direct parameter input	1	12,186
+det+adj+noun+noun(s)	a generic test suite	2	13,153
–det+adj+noun(s)	arithmetic subtraction	1	52,157
+det+adj+noun(s)	an effective call	2	44,291
+det+noun+noun+noun(s)	a subscriber line termination	3	8,656
–det+noun+noun+noun(s)	access control administration	4	9,834
+det+noun+noun(s)	a test cycle	3	40,746
–det+noun+noun(s)	envelope delay	4	62,699
+det+noun(s)	a window	3	139,367
–det+noun(s)	interworking	4	385,377
+det+noun+prep+noun(s)	a country of origin	5	{ 8,697
+ det+past part+noun(s)	a position defined parameter	5	}
+det+past part.+noun+noun(s)	the fixed component charge	6	{ 3,708
+det+pres.part.+noun+noun(s)	the ringing tone frequency	6	}
+det+past part. +noun(s)	a confirmed service	6	{ 18,037
+det+pres.part. +noun(s)	the magnifying optics	6	}
+det+noun+noun+pres.p+noun(s)	a path error monitoring function	7	115
+det+verb+noun+noun(s)	the send state variable	8	1,050
+det+verb +noun(s)	a call request	8	4,157

grammatical level. Thus, in the following examples, *charge unit, collect telephone, complete chain, total remuneration*, all of which were retrieved with pattern match file 8, the first word in each unit was always classified as a verb. Yet, it is clear from the examples that this is not their function within the unit.

6.4.2 Retrieval of term candidates from the GCSE corpus

Unlike the terms in the ITU corpus which were frequently complex or multi-word terms consisting of 3 words, the terms in the GCSE corpus are generally single word or 2 word terms. There is a very small number of 3-word terms (cf. pattern 2, *scanning electron micrograph*). The number of term formation patterns is much lower

Table 2. GCSE corpus

Pattern	Examples	P.	Occ.
det+noun+noun(s)	bank cashier	1	14,206
+det+noun+noun(s)	a step-up transformer	3	4,486
-det+noun(s)	chloramphenicol	1	135,200
+det+noun(s)	a carrier	3	44,543
+det+pres.part.+noun+noun(s)	a scanning electron micrograph	2	45
+det+past.part.+noun+noun(s)	a controlled chain reaction	2	420
+det+adj+noun(s)	the dry cell	4	8,680
–det+adj+noun(s)	hydrochloric acid	5	15,889

than in the ITU corpus and the difference is probably due to the fact that the nature and purpose of each of the corpora are quite different. The texts in the GCSE corpus are introductory textbooks written for secondary level pupils. It is therefore not surprising that the use of highly complex terms is avoided. Given the preponderance of simple adj+noun term candidates, it is quite likely that many of these are not terms at all but simply part of general vocabulary. It is anticipated that the refinement procedure described in Section 6.5 will allow us to eliminate many non-term candidates.

6.4.3 Retrieval of term candidates from the Nature corpus

As the Nature corpus tends not to contain linguistic signals of the type found in the ITU and GCSE corpora (i.e. *called, denotes*), the only means which we had for identifying term formation patterns was to examine the corpus and make a note of the patterns observed. Consultation with a subject expert would be required in order to validate our findings. We believe, in fact (see Table 3), that at least one of the patterns identified (adj+adj+noun+noun(s), e.g. *a weak transcriptional activation function*) contains modified terms. In other words, not all of the components may belong to the term. This would suggest that the longest multi-word term in the Nature corpus may only have three components.

6.4.4 General observations on output from first phase

In the ITU, GCSE and Nature corpora, we identified a number of different term formation patterns. In many of the pattern specifications, we specified that the pattern had to be preceded by a determiner. As the determiner tag in the CLG tagger attaches itself indiscriminately to all determiners (e.g. *a, an, the, all, some, any*) and

Table 3. Nature Corpus

Pattern	Examples	P.	Occ.
det+noun+noun+noun(s)	the T-cell antigen receptor	1	141
+det+noun+noun(s)	a replication origin	1	1,285
+det+noun(s)	a protein	1	6,772
+det+adj+adj+noun+noun(s)	a weak transcriptional activation function a low activity state	2	32
+det+adj+noun+noun(s)	a monoclonal antibody	2	431
+det+adj+noun(s)		2	2,534

not just to the one (i.e. *a/an*) which we consider to confirm generic reference to a term (cf. Section 6.5.1), it would be important to be able to discriminate between generic and specific determiners in a more sophisticated retrieval process. A similar problem arises with the term formation patterns where we specified that the first word in the term unit must be an adjective. The adjective tag in the CLG tagger is assigned to all adjectives, including many that are highly unlikely to form part of a term (e.g. *other, same, many*). Again, in a more sophisticated process, it would be necessary to produce a stoplist of such adjectives in order to reduce unnecessary noise in the initial output.

6.5 Refining the term identification process

In the introduction to this chapter, we suggested that the term identification process would occur in two stages. In the first stage, outlined in the preceding sections, the tag sequence patterns were used to retrieve all matches of the patterns examined in the three corpora. In the second stage, outlined in the remainder of this chapter, additional criteria are described and applied, in order to refine the output from the first stage. It is hoped that the application of these additional criteria will eliminate a large number of the non-terms retrieved during the initial retrieval process. While the refinements have not been implemented on a large scale because this would go beyond the scope of this investigation, we have implemented them on a small scale and have chosen a small set of examples from the ITU and GCSE corpora to demonstrate the usefulness of the additional criteria.

6.5.1 Generic reference

The first and, we believe, the most important criterion is that of generic reference. Generic reference is one of the key tenets of the traditional theory of terminology

where a clear line is drawn between generic concepts and individual objects.

> It should always be borne in mind that concepts cannot be taken for the individual object themselves. They are mental constructions serving to classify the individual objects of the inner or outer world by way of a more or less arbitrary abstraction. *ISO/R 704 Naming Principles* (1968:8)

ISO makes a distinction between the generic concept and the individual object, i.e. the realization of that concept by virtue of its location in time and space. Picht/Draskau (1985) also make a distinction between the generic and the individual. They prefer, however, to distinguish between a generic concept and an individual concept rather than an individual object. Like ISO, they argue that the presence or absence of definiteness, i.e. whether or not the concept can be located in time and space will determine whether or not the term is to be considered generic or individual.

An individual concept will be represented by a *name* rather than a *term*. The notion that the absence or presence of an indication of time and space allows us to distinguish between the generic and the individual respectively is an interesting one. However, while the traditional approach to terminology makes this distinction, it does not explain how this distinction is realized in text, and this is what is of particular interest to us.

We have noted in our corpora that terms are referred to in one of two ways; terms are either flagged or unflagged. When we use the expression *flagged term*, we mean that it may be preceded by any one of a number of determiners, with the exception of the indefinite article. In such instances, the reference is likely to be specific. The author is situating the use of the term in time and space, in this instance within a particular text. When we use the expression *unflagged term*, we mean that the term is preceded either by the indefinite article or is not preceded by any article at all. When a term is unflagged, the reference is likely to be generic. When the same term is flagged by a determiner other than the indefinite article, it is not possible to assume that reference is being made to the generic concept. There are, however, exceptions to this rule. For example, there are instances where a flagged term at the beginning of one sentence functions as an anaphoric reference to a generic reference made previously, as in the example below:

> (1) #IT# Cladding Mode Stripper#IQ# The cladding mode stripper is a device that encourages the conversion of cladding modes to radiation modes; as a result, cladding modes are stripped from the fibre. (ITU corpus)

We would suggest that it is not necessary for us to identify flagged terms because

the term will already have been identified by means of the generic reference criterion. We state therefore that the first criterion which a term candidate must meet is: A term candidate must have generic reference. For a term to have generic reference, it must be *unflagged*, i.e. preceded by the indefinite article or by no article at all.

Table 4. Examples of generic and specific reference

Examples from ITU corpus	Examples from GCSE corpus
a state picture	a balk
a cutoff call	a binary system
a bit-error performance measuring equipment	a powder avalanche
a telephone-type circuit	amoeboid motion
double-talk	coastal management
a clear description	a few minutes
a matter of agreement	a small whirlpool
a degree of processing	a balanced diet
*This single sheet	*some other tissues
*the nature of information	*the same way
*the work for examination	*the crushed tablets
* The complete chain	*the following key questions

All of the examples in Table 4 marked with an asterisk which were retrieved during the first phase of the term identification process no longer qualify as term candidates when the generic reference criterion is applied. This is confirmation that the generic reference criterion is very powerful and could usefully be used as a refinement in other term identification systems. However, even when the generic reference criterion has been applied, the remaining set of term candidates will still contain many non-terms, demonstrating that this criterion is not sufficient on its own. In the above set of examples there are some words or phrases which meet the generic reference criterion but may not be terms. These include *a clear description, a matter of agreement* and *a degree of processing* in the ITU corpus, and *a few minutes, a small whirlpool* and *a balanced diet* in the GCSE corpus. It is for this reason that the generic reference criterion alone is not sufficient; hence the condition that all term candidates must satisfy the generic reference criterion and must co-occur at least once with at least one of the linguistic signals specified below.

6.5.2 Linguistic signals

We already mentioned that we used a number of linguistic signals (e.g. *i.e., e.g., denotes*) to draw up our initial set of term formation patterns. It appears that the only

means of ensuring that a term candidate actually is a term is to specify that it satisfies the generic reference criterion and that it co-occurs with one of a specified list of linguistic signals. This list includes, in addition to those already used, the following: *called, known as, ". . .", the term*. It should become clear from the examples provided below, that the occurrence of one of these signals in conjunction with generic reference confirms the presence of a term.

6.5.2.1 The signal 'called'

In the corpora, *called* appears as part of the following phrases: *is/are called*, *is/are often called*, *also called*, *usually called*, *generally called*, *sometimes called*, (*usually called*), (*often called*). If a term candidate retrieved during the first part of the process satisfies the generic reference criterion and occurs immediately after one of the above phrases, it may be considered to be a term.

Examples from the ITU corpus

(2) Alternatively a single piece of equipment *called* a transmultiplexer can be used to
(3) describes an ellipse during one period: this is *called* elliptical polarization.
(4) The MD managed by an organization is *called* a Private Management Domain (PRMD).
(5) A second parameter which is closely related to jitter is *called* wander.
(6) pictorial elements. This picture is *called* a state picture.

Examples from the GCSE corpus

(7) one by a narrow piece of land *called* a balk. Why was this bad farming?)
(8) can be measured by an instrument *called*a barometer. There are two main
(9) pulse of electricity. This code is *called* a binary system. The pulses travel
(10) star and the two stars together are *called* a binary star.The Solar System has
(11) feeding does not need light and is *called* saprophytic nutrition.Some kinds of
(12) length.This kind of cell division is *called* mitosis see figure 1.5). It is

6.5.2.2 The signal 'known as'

Known as appears as part of the following phrases:*known as, also known as, sometimes known as, generally known as, commonly known as*. If a term candidate retrieved during the first part of the process satisfies the generic reference criterion and occurs immediately after one of these phrases, it may be considered to be a term.

Examples from the ITU corpus

(13) This method is also *known as* priority reservation system.
(14) This apparent missing component of offered traffic is *known as* suppressed traffic.
(15) The media access control discipline *known as* carrier sense multiple access (CSMA)is
(16) The different signalling paths so formed are *known as* signalling routes.

Examples from the GCSE corpus

(17) directly to the electorate is *known as* a referendum. Holding a
(18) through the air as a cloud. This is *known as* a powder avalanche. Avalanches
(19) This kind of exchange system is *known as* barter.An economic change
(20) sharing between non-metals is *known as* a covalent bond. Notice that
(21) the amoeba to help it move. (This is *known as* amoeboid motion. An amoeba

6.5.2.3 The signal 'e.g.'

This linguistic signal occurs as part of the main body of the text or within brackets. If a term candidate retrieved during the first part of the process satisfies the generic reference criterion and occurs immediately after the linguistic signal *e.g.*, it may be considered to be a term.

Examples from the ITU corpus

(22) ion are provided by different media (*e.g.* a satellite channel in one direction
(23) whom the charges are to be charged (*e.g.* a branch, a bank or a similar institu
(24) through another communication system (*e.g.* a physical delivery system)that is
(25) *success for a longer period of time (*e.g.*, a few hours), the preferred response
(26) *e done on very short pieces of fibre (*e.g.* a few centimetres).

Examples from the GCSE corpus

(27) following groups: Class I Professional *e.g.* chartered accountants, senior civil
(28) Class II, Managerial and executive *e.g.* pharmacist, departmental manager,
(29) Class VI, Semi-skilled manual *e.g.* painters and decorators, drivers
(30) drugs are produced in laboratories, *e.g.*chloramphenicol # used to combat
(31) years)and are called non-renewable, *e.g.* coal. problems will be caused if any
(32) is naturally stored in porous rocks, *e.g.*chalk, sandstone or limestone,

Examples (25) and (26) from the ITU corpus, marked with an asterisk*, highlight the problem posed by a tagger which classifies all adjectives together.

6.5.2.4 The signal 'the term'

If a term candidate retrieved during the first part of the term identification process satisfies the generic reference criterion and occurs immediately after the linguistic signal *the term*, it may be considered to be a term.

Examples from the ITU corpus

(33) In telephony, use of *the term* ''circuit'' is generally limited to a telecommunication
(34) hierarchical level, *the term* ''frame alignment'' is synonymous with ''multiframe
(35) ''peer protocol'' and ''peer entities''. *The term* ''boundary'' applies to boundaries
(36) telex subscriber over the telex network. *The term* ''notification'' applies to the
(37) *The term* Computerized Communication Terminal (CCT)refers to a device or
(38) *The term* echo control device will comprise both echo suppressors and echo cancellers.

Examples from the GCSE corpus

(39) energy is passed on to animals. The term *biomass* is used to refer to anything
(40) referred to as bureaucrats. *The terms* bureaucracy' and 'bureaucrats' have
(41) are not the Coastal management *The term* COASTAL MANAGEMENT is a useful
(42) camps of the Boer War. *The term* 'concentration camp derives from the
(43) and cell biochemistry. We now use *the term* Neo-Darwinism (neo = new). Several

Frequently, in the ITU corpus, the expression *the term* is followed by a word or phrase in inverted commas. This is a device which is common in specialized texts and which indicates that what appears in inverted commas is either a recently coined term or that, although the lexical unit may look like a general language word (e.g. *boundary, notification* above), it is in fact a term with a precisely defined sense.

6.5.2.5 The signal ". . ."

If a term candidate retrieved during the first part of the term identification process satisfies the generic reference criterion and appears within inverted commas, it may be considered to be a term.

Examples from the ITU corpus

(44) diallable symbols #IQ# As used here, ''distinct'' refers to dissimilarity from oth
(45) ellite connection (echo problems and ''double-talk'') is preferable to the degrad
(46) are talking simultaneously (termed ''double-talking''). To reduce this effect (c
(47) sion section, and generation of AIS ''downstream''; – detection of AIS inside the

(48) is level, one extends the concept of ''downtime'' used in availability specificat

There were no occurrences of this signal in the GCSE corpus.

The use of inverted commas is very useful in one respect but problematic in others. As we have noted, they are frequently used in specialized texts to signal that a term has only recently been coined. They could therefore be used as a mechanism for identifying and retrieving new terms in a corpus. On the other hand, inverted commas can also be used to indicate that the word or phrase which appears between the inverted commas is not the ''correct'' term, but simply a general language equivalent which may be more accessible to the reader.

6.6 Conclusion

In this chapter, a mechanism for identifying and retrieving terms in the three corpora was proposed and outlined. The first step involved (manual) identification of the syntactic patterns which occurred with certain linguistic signals in order to devise a list of term formation patterns for each of the corpora. These patterns were used as input to a pattern matcher which retrieved all instances of these patterns in the corpora. As the output contained many non-terms, further restrictions were specified in order to refine the retrieval process. The first of these restrictions specified that all term candidates had to have generic reference. Only those term candidates which were unflagged, i.e. which were preceded by the indefinite article or by no article at all, were considered to have generic reference. While this is a very powerful criterion, and one which could usefully be applied in other term identification systems, it proved not to be sufficient on its own. We therefore specified a second restriction, namely that all term candidates should also co-occur at least once with one of a specified set of linguistic signals. On the basis of a fairly detailed analysis of the corpora, we concluded that this was a useful mechanism for refining the identification process. It should be noted that the set of linguistic signals described here is not exhaustive. There were other signals which could also be used as part of the refinement process. For example, if a term candidate with generic reference appears at the end of a sentence and is immediately followed at the beginning of the following sentence by phrases such as *This process, This method, This device*, it is very likely that the term candidate is indeed a term. However, as noted at the beginning of this chapter, we were more interested in identifying as many terms as possible which were partially or fully explained in the corpora than in producing an exhaustive list of all terms or indeed of all term formation patterns in the three corpora. There is ample scope for pursuing this line of enquiry which, as outlined here, constitutes only the beginning of what could prove to be a very fruitful basis for further research.

7 Retrieval of formal and semi-formal defining expositives

7.1 Introduction

We suggested in chapter five that authors writing for certain specified communicative settings are likely to explain some of the terms which they are using. The extent to which they do this will depend on the perceived disparity of knowledge between the author and the reader. The explanations provided may correspond to what we termed partial and complete defining expositives or, as we will discover in the next chapter, they may be simply fragments of information scattered throughout a text which have to be combined in order to form a partial or complete defining expositive. We believe that these explanations can be retrieved and used as input for the formulation of specialised definitions.

As already noted in chapter four, several authors make a distinction between formal, semi-formal and non-formal definitions when describing how definitions might be expressed in text. It was also noted that, with few exceptions, authors' views on how such definitions might be expressed was based on their own intuitions rather than on an analysis of textual evidence. What we propose to do here is to look at texts in order to ascertain how one might (semi-)automatically recognize that some form of explanation is being provided. We have discovered that some syntactic patterns in the three corpora under investigation appear to indicate that one of the definition types described previously is being provided.

In chapter five, we made a distinction between defining exercitives and defining expositives. Defining expositives involve the rephrasing of an existing definition for the purposes of explanation or clarification. We suggested that defining expositives could be complete or partial and that their presence in a text could be implicit (i.e. buried or embedded in the text) or flagged explicitly. Complete defining expositives are those which match the formal definition formula proposed by Trimble. In other words, a term is defined in terms of its superordinate and one distinguishing characteristic. Partial defining expositives are those which correspond to what Trimble calls semi-formal definitions whereby a term is explained in terms of the difference between it and other members of its class.

In this chapter, we will start by examining formal defining expositives and will focus initially on retrieving simple formal defining expositives, i.e. expositives where the term and its definition appear in the one sentence. A set of conditions for the retrieval of these statements will be specified and explained. These are the conditions of felicity which must be met in order for statements to qualify as formal defining expositives. We will then discuss how it might be possible to extend the retrieval process beyond the sentence boundary in order to retrieve complex formal defining expositives, i.e. expositives where the term and the defining statement appear in two separate sentences. Examples of simple and complex formal defining expositives in the three corpora are provided. Having discussed how simple and complex formal defining expositives can be identified, we will then look at the retrieval of semi-formal defining expositives. A second set of felicity conditions will be specified for retrieving these. Semi-formal defining expositives will be exemplified with examples from the three corpora.

In the final section of this chapter, we will look at another type of defining expositive which is used in the ITU and GCSE corpora. When examining these corpora, we noted that they appeared to contain what must have originally been dictionary defining expositives. Such expositives look like the type of definition which we would expect to find in a conventional dictionary. These, we believe, are glosses which are consciously and deliberately provided by the authors in order to define terms which they have used and which they believe to be unknown to readers. They consist of both formal and semi-formal defining expositives. We will describe the patterns used in the ITU and GCSE corpora to signal the presence of these definitions and make some suggestions for retrieving them.

7.2 Simple formal defining expositives

The pattern for a formal definition is generally expressed as follows:

Formula 1:*X* = *Y* + distinguishing characteristic, whereby *X* is subordinate to *Y*

The formal definition will provide three types of information (adopted from Trimble 1985:80): the name of the *term* being defined, the *class* to which the term belongs and the *difference(s)* between the term and all other members of the class. Each of these types of information must be present for the statement to be considered as a formal definition candidate. The above formula indicates that the term to be defined appears in sentence-initial position and is followed by the defining statement but, as we will demonstrate, there are many instances in our corpora where the formula appears in reverse order, i.e. as:

Formula 2: *Y* + distinguishing characteristic = *X*, whereby *X* is subordinate to *Y*.

Flowerdew suggests that ''where the structure of term + class + characteristic is employed, the term has very often already been introduced into the discourse and is thus given (as opposed to new) information in the definition itself'' (1992b:168). The evidence of the three corpora which we are using certainly seems to confirm this. When formula 1 is used, *X*, i.e. the term, frequently appears towards the end of the preceding sentence or as a caption or heading preceding the defining expositive. Flowerdew suggests that if a term has not yet been introduced in the discourse ''the semantic ordering is reversed, with the term coming at the end'' (1992b:168). While this is certainly true for many terms in our corpora, there are also instances where *X* appears at the end of the defining statement, even when it has already been introduced in the discourse, as the following example from the GCSE corpus illustrates:

(1) Longshore Drift. One important natural activity which can lead to coastal change is known as Longshore Drift.

However, when *X* has already been introduced, it is generally in the form of a heading or caption rather than as part of a sentence and, in this sense, Flowerdew's contention is correct.

7.2.1 Identifying simple formal defining expositives in text

This section will focus on the identification of simple formal defining expositives, i.e. formal defining expositives which appear within a single sentence. As many statements in the corpora which match formulae 1 and 2 for simple formal definitions are not themselves simple formal defining expositives, we need to specify a number of conditions in order to eliminate statements which do not function as defining expositives. We have specified two separate sets of conditions. The first relates to specifications for the slot fillers for *X, Y* and '='. The second set applies to the statement as a whole and consists of the conditions which formulae 1 and 2 must meet in order to qualify as simple formal defining expositives. Each of these will be discussed and exemplified.

7.2.1.1 Specifying slot fillers for X, Y and '='

7.2.1.1.1 Specifying slot fillers for X. *X* must be a term. To qualify as a term candidate it must match one of the term formation patterns previously specified. In addition, it must, at least once in the corpus under investigation, have satisfied the

generic reference criterion and co-occurred with one of the linguistic signals which we specified.

If *X* appears in sentence-initial position, i.e. on the left hand side of the defining statement, it must be preceded by a) the indefinite article, or b) no article at all. In other words, it may not occur with any modifier other than the indefinite article. This condition allows us to ensure that the particular statement which is retrieved has general rather than specific reference and can qualify as a candidate formal defining expositive. While there are also instances of formal defining expositives in the corpora where *X* is in sentence-initial position and is preceded by the definite article or by the demonstrative adjective, these belong to what Trimble (1985) calls complex formal definitions, i.e. defining expositives where the boundary of the expositive lies beyond the sentence delimiters and are discussed in Section 7.3.

If *X* appears at the end of the defining statement, the only modifiers which may precede it are the definite or the indefinite article. Here it is acceptable for *X* to be modified by the definite article if *Y* (cf. below) has also been preceded by the definite article. What usually happens in this instance is that *Y* has itself functioned as an *X* with generic reference in one sentence and functions as a *Y* (i.e. a superordinate or class word) in the following sentence. In the following sentences from the ITU corpus, *X* appears at the end of the sentence and is modified by the definite article. Here, the definite article is used to indicate that *X* is a specific type of *Y*.

(2) The exchange which is responsible for setting up calls and decides the order in which they are to be connected is called the controlling exchange.
(3) The period during which a component ID is released, but cannot be reallocated, is called the freezing period.
(4) The period in which data is accumulated is called the test period.

7.2.1.1.2 Specifying slot fillers for Y. *Y* must be either a term or one of a set of specified class words. For *Y* to be a term, it must, as in the case of *X*, match one of the term formation patterns previously specified, and must have, at least once in the particular corpus under investigation, satisfied the generic reference criterion and co-occurred with one of the specified linguistic signals. When *Y* is what we call a class word, this means that it is a generic term such as *process, method, function, property*. A number of different class words are used in the corpora. Table 1 lists the most common of these, together with the total number of occurrences in each corpus and the number of occurrences per 100,000 words. *Y* may not be preceded by any modifier other than the definite or the indefinite article and it may also occur without any modifier at all. This applies to all instances of *Y*, i.e. when it is functioning either as a term or as a class word and whether it appears on the lhs or rhs of the connective verb.

Table 1. Class words

Classword	ITU corpus 4.7 m words		GCSE corpus 1m words		Nature corpus 229,676 words	
	Total	100,000	Total	100,000	Total	100,000
technique	341	7.23	7	0.7	9	3.92
method	2,374	50.51	149	14.9	32	13.93
process	1,143	24.32	201	20.1	63	27.43
function	3,105	66.06	54	5.4	133	57.91
property	39	.83	71	7.1	9	3.92
system	7,396	157.36	712	71.2	135	2.87
class	1,598	34.00	41	41.6	81	35.27
device	746	15.87	69	.9	4	1.74

7.2.1.1.3 Specifying the slot fillers for '='. A number of verbs or verb phrases may fill the '=' slot and these have variously been termed *hinges* or *connective verbs*. The term *hinge* is used by Barnbrook and Sinclair (1994) to describe part of the structure in the Cobuild dictionary definition. In a Cobuild definition, a *hinge* is a word such as *if* and *when* when it is used to introduce the lhs of the definition and a word or phrase such as *is/are, were, means, consists of* when it is used to introduce the rhs of the definition (1994:21). Sager (1980) uses the term *connective verb* to describe the verbs which link the lhs and rhs of a description statement. Such descriptive statements tend to be very simple, ''consisting merely of two nominals linked by verbs such as to be, to have or to give'' (Sager 1980:186). He considers that these verbs are devoid of meaning and that sentences containing these verbs consist of a grammatical subject and its descriptive predicate with the lhs listing the names of objects and the rhs describing them, resulting in a tabular-type presentation. In both cases (i.e. *hinge* and *connective verb*), it is assumed that *X* will be on the lhs and *Y* on the rhs of the descriptive statement or definition. As the term *hinge* has two senses, we have chosen to use the general term *connective* to describe all verbs and phrases which connect a term with information about the term; such information may range from the provision of a formal defining expositive to simply the specification of the term's superordinate. We use the term *connective verb* to encompass all verbs and verb phrases which are used to link a term with a partial or complete defining expositive and the term *connective phrase* to describe other connectives which link a term with information about a term but where the connective is not a verb or verb phrase.

We had originally planned to use the set of connective verbs provided by Sager as a basis for our analysis but encountered some problems with this approach. We took the list of connective verbs provided by Sager and examined the three corpora

Table 2. Connective verbs

Connective verbs	ITU corpus 4.7 m words		GCSE corpus 1 m words		Nature corpus 229,676 words	
	Total	100,000	Total	100,000	Total	100,000
comprise(s)	534	11.36	0	0	13	5.66
consist(s)	1,173	24.95	67	6.7	72	31.35
define(s)	1,514	32.21	14	1.4	18	7.90
denote(s)	188	4.00	0	0	5	2.18
designate(s)	110	2.34	0	0	0	0
is/are	61,478	1,308.04	9,462	946.2	3,906	1,700.65
is/are called	257	5.47	598	59.8	3	1.31
is/are defined as	320	6.81	1	.1	8	3.48
is/are known as	28	.60	37	3.7	2	.87

under investigation to establish 1) whether these occurred and 2) whether they were used in the manner described by him. We found that many of the connective verbs in Sager's list hardly ever occurred in our corpora; for example, *entail* occurs only once in the GCSE corpus, *imply* occurs only four times in the same corpus and *is ascribed to* never occurs in the GCSE corpus. When connective verbs are used in the corpora, they do not always function in the way that Sager has suggested. We decided therefore to devise our own set of connective verbs based on corpus evidence. This set is much smaller than the one originally envisaged by Sager.

Sager's assertion that connective verbs are devoid of meaning is slightly misleading and we would argue that the connective verbs which we have identified may have meaning in the sense that they may determine the type and content of the defining statement which is being provided. They may point to the provision of intensional information, extensional information, information about synonyms or a combination of all of these. A connective verb such as *to be* will provide intensional information, a connective verb such as *consist* may provide extensional information and a connective verb such as *known as* may provide synonyms. The connective verb may tell us something about the type of distinguishing characteristic which is about to follow. For example, a connective verb such as *be used to* is likely to be followed by a statement of purpose of function.

Table 2 contains a list of the verbs/verb phrases which function as connective verbs for formal defining expositives in the corpora. The table lists the connective verbs, the total number of occurrences of these connective verbs in each of the corpora and the number of occurrences per 100,000 words.

As *designate(s*) occurs only in the ITU corpus and c*omprise(s)*, and *denote(s*) occur in the ITU and Nature corpora but not in the GCSE corpus, we decided to eliminate these three connective verbs from the investigation. The connective verbs

which we have chosen to investigate for simple formal defining expositives are *is/are, is/are called, consist(s), is/are defined as, is/are known as*.

We have distinguished between two classes of connective verb, class 1 and class 2 connective verbs. With class 1 connective verbs, *X* appears on the lhs and *Y* + the distinguishing characteristic appear on the rhs of the formal defining expositive. With class 2 connective verbs this order is reversed. Class 1 connective verbs include *comprise(s), consist(s) of, define(s), denote(s), describe(s), designate(s), is/are*. Class 2 connective verbs include *is/are, is/are called, is/are known as*.

7.2.1.2 The expression of simple formal defining expositives

In addition to specifying the slot fillers for *X, Y* and '=', we have specified a further set of conditions relating to the expression of the defining expositive. These are discussed below.

7.2.1.2.1 Location of the statement. The '*X = Y*' part of the defining expositive must constitute the main clause of the sentence (or the first main clause if there are two main clauses in the sentence) and may not be prefaced by clauses or phrases other than those which fill the *X, Y* and '=' slots. This allows us to exclude phrases or clauses at the beginning of the sentence which might affect the general applicability of a defining expositive. Such phrases or clauses might be restrictive, confining the scope of the defining statement to the text in which it appears. Expressions such as *for our purposes, here* and many others are used to introduce such restrictions, as the following example from the ITU corpus illustrates.

(5) As used here, ''distinct'' refers to dissimilarity from other symbols compared with them visually, or aurally.

In this example, the use of the phrase *as used here* restricts the scope of the reference in the defining statement; what is being provided is a specific rather than general definition of *distinct*. However, it is worth noting that clauses or phrases which appear before the defining statements are not always restrictive and can sometimes contribute to the content of the defining statement, particularly in cases where the defining statement is preceded by a prepositional phrase which specifies the domain in which the term is used, as the following examples from the GCSE corpus illustrate.

(6) In biology, a fruit is a ripened ovary, together with its contents, the seeds, and anything that grows attached to it.
(7) In science, a law is a natural relationship that always holds true for a given set of conditions.

For reasons of computational difficulty, we have chosen not to include such phrases in the analysis of formal defining expositives but we plan to pursue this research further at a later stage. While we have excluded sentences with sentence-initial clauses or phrases which do not form part of the defining statement, we do not exclude the possibility of a second clause after the defining statement when this clause is introduced by the co-ordinating conjunction *and*.

7.2.1.2.2 Form of the connective verb. If the connective verb is the main verb in the defining statement it must be in the present tense, indicative mood. It may be in active or passive voice. The use of the present tense gives the statement general applicability; this would not be possible were the connective verb in a past or future tense where the statement is more likely to be a description of a particular event than a definition.

It is not acceptable for connective verbs to be used in combination with any modal verb with the exception of *can*. Modals are excluded from consideration because they restrict the scope of defining statements and introduce an element of doubt. For example, if *X may be defined as Y* + distinguishing characteristic, this suggests hedging on the part of the author and restricts the general validity of the statement. It could be argued that the modal *can*, when used in a phrase such as *can be defined as*, plays a similar role. However, corpus evidence seems to suggest that when *can* is used in a defining statement, the author is indicating that the definition being provided is one of a number of accepted ways in which the term in question can be defined, as the following example from the ITU corpus illustrates:

(8) A model can be defined as an abstraction of a reality as seen from a certain viewpoint.

Connective verbs which are qualified by negating words such as *not, never* do not qualify for consideration. We are more interested in defining terms in terms of what they are rather than what they are not.

7.2.1.2.3 Exclusion of focusing adverbs. Focusing adverbs (e.g. *generally, usually, commonly*) are used quite frequently in the GCSE, ITU and NATURE corpora. Table 3 provides a list of focusing adverbs and the number of occurrences in each of the corpora. The function of focusing adverbs in defining statements in the corpora is either to render a statement generally applicable or to restrict the scope of the reference. We have included those which enhance the general applicability of a statement and excluded those which do the reverse. Those which we accept are *commonly, generally, usually*. When authors use these particular adverbs in a defin-

Table 3. Focusing Adverbs

Adverb	ITU corpus 4.7m words		GCSE corpus 1m words		Nature corpus 229,676 words	
	Total	100,000	Total	100,000	Total	100,000
chiefly	5	.02	9	.9	1	.44
commonly	91	1.94	19	1.9	4	1.74
especially	176	3.74	142	14.2	21	9.14
exceptionally	88	1.87	1	.1	4	1.74
exclusively	61	1.30	3	.3	6	2.61
frequently	93	1.98	59	5.9	10	4.35
generally	496	10.55	140	14.0	35	15.24
mainly	98	2.09	125	12.5	20	8.71
mostly	11	.23	85	8.5	18	7.84
occasionally	18	.38	13	1.3	1	.44
often	213	4.53	574	57.4	37	16.11
only	4,504	95.83	1,457	145.7	385	167.63
on the whole	2	.04	13	1.3	1	.44
predominantly	12	.26	5	.5	6	2.61
primarily	118	2.51	4	.4	19	8.73
principally	9	.19	0	0	2	.87
purely	61	1.30	10	1.0	3	1.31
rarely	22	.47	31	3.1	1	.44
solely	93	1.98	5	.5	5	2.18
sometimes	111	2.36	252	25.2	11	4.80
specifically	175	3.72	2	.2	27	11.76
usually	354	7.53	408	40.8	17	7.40

ing statement, they are indicating that the definition which they are providing is the usual definition for this particular term. It is not possible to say the same for any of the other focusing adverbs listed in Table 3. Consequently, we reject these from our considerations. What is interesting about this latter group is that, in normal circumstances, one would consider that the use of some of these adverbs with any of the connective verbs would suggest that the statement was generally applicable and therefore had some form of definitional status. The focusing adverb *often* is a case in point. As we can see from the following example from the ITU corpus, of which there are many in all three corpora, the use of *often* restricts the general applicability of the statement.

(9) The cladding mode stripper often consists of a material having a refractive index equal to or greater than that of the fibre cladding.

When a term is described as *often* having a particular characteristic, it is not possible to conclude that it *always* has this characteristic.

7.2.1.2.4 Specifying the introduction of the distinguishing characteristic. *Y* may be followed by one of the following: a preposition, a relative pronoun, a past participle. It may not be followed by the co-ordinating conjunctions *and, or, but*. If *Y* is followed immediately by *and*, this may signal the introduction of another noun phrase or the beginning of another clause rather than the introduction of the distinguishing characteristic, as the following example from the ITU corpus illustrates.

(10) Telex is a message transfer service and therefore interworking between telex and videotex should be limited to the exchange of alphanumeric text between terminal equipments.

If it is followed immediately by *or*, this is likely to signal the introduction of another noun phrase. The use of *but* tends to signal an exception rather than the rule.

7.2.1.3 Summary
The conditions which must be met in order to facilitate the retrieval of simple formal defining expositives within the corpora have now been specified. We have specified the slot fillers for each of the components in the defining expositives and have specified additional restrictions which apply to the expositive as a whole. The next section contains examples of expositives which have been retrieved from the corpora using these conditions.

7.2.2 Examples of simple formal defining expositives

All of the examples provided here fulfil the above specified conditions. To facilitate analysis, we have classified the examples according to the class of connective verb used and, for each connective verb, according to the class of word used to introduce the distinguishing characteristic. As previously indicated, two broad classes of connective verb have been used for retrieving formal defining expositives. In the first class, *X* appears on the lhs of the connective verb and the defining statement on the rhs while, in the second class, the order is reversed. The first class consists of: (*is/are*, *consist/consists, is/are defined as*). The second class consists of: (*is/are known as, is/are called*).

7.2.2.1 Examples with class 1 connective verbs
The first pattern is as follows:*X is/are Y* + distinguishing characteristic. In this pat-

tern, the distinguishing characteristic may be introduced by a) a past participle, b) a relative pronoun or c) a preposition.

1.a Distinguishing characteristic introduced by past participle

Examples from ITU corpus

(11) A misdelivered frame is a frame transferred from a source user to a destination user other than the intended destination user.
(12) A videotex service centre is a computer used by the videotex service provider to authorize access to a videotex service.

Examples from GCSE corpus

(13) Bile is a green liquid made in your liver and it is stored in your gall bladder.
(14) Sulphur dioxide is a gas given off by some fuels as they burn.
(15) An abscess is a cavity filled with pus.

Example from Nature corpus

(16) INTERLEUKIN-1 (IL-1) is a cytokine1 produced primarily by mononuclear phagocytes.

While the only example of this pattern found in the Nature corpus contains a focusing adverb, it is still eligible for consideration because the focusing adverb is not modifying the connective verb.

1.b Distinguishing characteristic introduced by relative pronoun

Examples from ITU corpus

(17) An expedited-data-unit is a service-data-unit which is transferred and/or processed with priority over normal service-data-units.
(18) Telewriting is a communication technique that enables the exchange of handwritten information through telecommunication means.
(19) Multi-endpoint-connections are connections which have three or more connection-endpoints.
(20) Lexical rules are rules which define how lexical units are built from characters.

Examples from GCSE corpus

(21) A pike is a carnivorous fish which lives in lakes and rivers.
(22) Vorticella is a unicellular animal which lives in ponds, puddles and sewage filters.
(23) A robot is a machine that tries to copy one or more human functions.

Examples from Nature corpus

(24) Kinesin is a motor protein that uses energy derived from ATP hydrolysis to move organelles along microtubules.
(25) A pCM is a nonlinear crystal that forms a hologram of the incident beam.

1.c Distinguishing characteristic introduced by preposition

Examples from ITU corpus

(26) A control circuit is a telephone-type circuit between the point of origin of the programme and the point where it terminates (recording equipment, studio, switching centre, transmitter, etc.)used by a broadcasting organization for the supervision and coordination of a sound or television transmission.
(27) A bearer channel is a unidirectional, digital, transmission path from the transmit unit of one DCME to the receive unit of a second associated DCME and which is used to carry concentrated traffic between two associated DCMEs.

Examples from GCSE corpus

(28) Magnetic tape is a plastic tape with a thin coating of metal oxide (or sometimes pure metal)particles attached.
(29) A cheque is a written order to a bank telling it to make a payment.

Example from Nature corpus

(30) Stearothermophilus PFK is a tetramer of identical subunits with 319 amino acids.

There were very few occurrences of this pattern in the GCSE and Nature corpora.

Pattern 2: *X consist(s) of Y* + distinguishing characteristic

This particular connective verb is interesting because one would normally expect it to introduce a list of component parts. While it is indeed used to introduce lists in the corpora, it is also used in the same way as the connective verb, *to be*, i.e. to

introduce a superordinate term + distinguishing characteristic. It is easy to distinguish between the two uses because when it is used to introduce a list, *Y* is either modified by a number or followed by the conjunction *and*. When *Y* is not modified by a number and/or not followed by the conjunction *and*, we know that *consist(s) of* is being used to introduce a superordinate. In the Nature corpus, this particular connective verb is not used to introduce formal defining expositives of the type described here and is used only to introduce part-whole relations. Consequently, no examples from the Nature corpus are provided here.

2.a Distinguishing characteristic introduced by past participle

Examples from ITU corpus

(31) A parameter name defined parameter consists of a parameter name followed by a parameter value from which it is separated by an = (equal sign).
(32) A supergroup consists of a supergroup link connected at each end to terminal equipments.
(33) Direct parameter input consists of an optional parameter block entry sequence preceded by the separator : (colon).

Example from GCSE corpus

(34) Skeletal muscle consists of bundles of muscle fibres held together by connective tissue.

2.b Distinguishing characteristic introduced by relative pronoun
Example from ITU corpus

(35) Plain language consists of words that present an intelligible meaning in one or more of the languages admitted for international telegrams, which include at least French, English and Spanish.

Example from GCSE corpus

(36) A nephron consists of a cup-shaped, holllow Bowman's capsule (Fig. 22.5) which leads into a long, narrow tubule.

2.c Distinguishing characteristic introduced by preposition

This pattern does not occur in any of the three corpora.
Pattern 3: *X is/are defined as Y* + distinguishing characteristic

As the connective verb *be defined as* occurs only once in the form *is/are defined as* in the GCSE corpus, and, when used in combination with the modal *can*, does not fulfil all of the other conditions, no examples are provided. As it does not occur with a past participle or preposition in the Nature corpus, no examples are provided for these patterns.

3.a Distinguishing characteristic introduced by past participle

Example from ITU corpus

(37) A final frame is defined as the last frame transmitted prior to an expected response from the distant station.

3.b Distinguishing characteristic introduced by relative pronoun

Examples from ITU corpus

(38) A confirmed service is defined as a service which results in an explicit confirmation from the service-provider.

(39) An unconfirmed service is defined as a service which does not result in an explicit confirmation.

Examples from Nature corpus

(40 A unit of inhibitory activity is defined as the amount that produces the same level of inhibition of the binding of 1ng 1251-labelled L-1 alpha protein to El4–6.1 cells as 1ng recombinant IL-1 alpha protein.

3.c Distinguishing characteristic introduced by preposition

Examples from ITU corpus

(41) Through-connection delay is defined as the interval from the instant at which the information required for setting up a through-connection is available for processing in an exchange,….
(42) A circuit is defined as the complete transmission path between the switch points of the two private exchanges concerned.
(43) A model can be defined as an abstraction of a reality as seen from a certain viewpoint.

7.2.2.2 Examples with class 2 connective verbs

Pattern 4: *Y* + distinguishing characteristic *is/are called X*

As this pattern occurs only three times in the Nature corpus, and as none of these corresponds to the patterns described here, no examples are provided.

4.a Distinguishing characteristic introduced by past participle

Examples from ITU corpus

(44) A message described by a probe is called a described message.
(45) An MD managed by an organization other than an Administration is called a private management domain (PRMD).
(46) Money orders and postal cheques transmitted by telegraph are called ''POSTFIN telegrams''.
(47) Transport protocol data units (TPDUs) carrying transport service (TS)user information or control information are called blocks .

Example from GCSE corpus

(48) The black substance made from the remains is called humus.

4.b Distinguishing characteristic introduced by relative pronoun

Examples from ITU corpus

(49) A functional object that provides one link in the MTS' store-and-forward chain is called a message transfer agent (MTA).
(50) Transceivers that meet warm-start requirements are called warm-start transceivers.

Examples from GCSE corpus

(51) A material that is made up of only one type of atom is called an element.
(52) Animals which feed on plants are called herbivores.

4.c Distinguishing characteristic introduced by preposition

Examples from ITU corpus

(53) A model with several independent variables is called a multiple regression model.

(54) A connection (if any) between an incoming and an outgoing circuit at interfaces to other exchanges/networks is called a transit connection.

Examples from GCSE corpus

(55) The rush of water up the beach following each wave-break is called the swash.
(56) The tough white outer coat of the eye to which the eye muscles are attached, is called the sclera.

Pattern 5: *Y* + distinguishing characteristic *is/are known as X*
While this particular pattern is quite common in the ITU and GCSE corpora, it never fulfils all of the conditions specified for simple formal defining expositives.

7.2.2.3 Observations
While all three corpora contain simple formal defining expositives, these expositives are more common in the ITU and GCSE corpora than in the Nature corpus, even when the differences in corpus size are taken into consideration. Given that the texts in the Nature corpus are written by experts for their peers, this is not surprising. In this particular communicative setting (expert-expert communication), there is likely to be a high density of terms but rather few explanations. This is because readers are expected to understand the terminology being used. When simple formal defining expositives are provided in the Nature corpus, they relate to what appear to be very technical terms and the language used in the definitions is not readily accessible to the non-expert.

Simple formal defining expositives are more common in the ITU corpus which, again, is not surprising given that the corpus corresponds to a communicative setting where some terms are likely to be explained. The ITU corpus was written for people who already have some knowledge of the field of telecommunications but need to learn more about it. The language used in the simple formal defining expositives provided in the ITU corpus is again fairly technical and there is an assumption that readers will already have some knowledge of the field.

Simple formal defining expositives are most common in the GCSE corpus and it appears that very many terms are defined. The language used in the simple formal defining expositives in the GCSE corpus is accessible to the ordinary reader and requires no prior knowledge of the subject.

The restrictions which have been specified for the retrieval of simple formal defining expositives make it possible to retrieve most, if not all, such expositives. The exclusion of some modals and some focusing adverbs in particular ensure that only those statements which have general validity are retrieved.

We chose to classify the patterns according to the connective verb used and ac-

cording to the grammatical category of the word used to introduce the distinguishing characteristic. We had hoped that the grammatical category might indicate what type of information was likely to follow. Unfortunately, the examples listed here and other examples found in the corpora do not allow us to do this because there are not sufficient instances of any one of these occurring in the form of simple formal defining expositives. In a larger corpus, this might be possible and it is certainly worthy of further investigation. For example, the use of the preposition *for* might indicate a characteristic of purpose; the use of the past participle *made* might indicate a characteristic of composition. In general, the simple formal defining expositives described in this section are much less likely to occur than other types of defining expositives. However, they are also easier to retrieve than any of the other types. This is because they occur within a single sentence and because we have imposed a fairly restrictive set of conditions which they have to satisfy in order to qualify as candidates for simple formal defining expositives. The conditions specified allowed us to retrieve and use the statements as they appeared in the corpora. The retrieval of complex formal defining expositives, defined in the next section, is much more complicated but essential if we are to consider using corpora for building terminological definitions.

7.3 Complex formal defining expositives

In addition to the patterns specified for simple formal defining expositives, there are many other patterns in the corpora which have all of the elements of a simple formal defining expositive but which would require further processing in order to be presented in the formula *X* = *Y* + distinguishing characteristic. These are complex formal defining expositives, i.e. expositives which span more than one sentence. We have chosen to present some of these here with a view to establishing how they might be transformed to simple formal defining expositives. Not all of the restrictions specified for simple formal defining expositives will always apply for the retrieval of complex definitions; however, we still expect all expositives to contain the separate elements specified for a simple formal defining expositive (i.e. *X*, *Y*, =, distinguishing characteristic) in order to qualify as a complex formal defining expositive, even if they do not appear in the same order or in the same mode as in the simple formal defining expositives. Where necessary, we have specified a separate set of conditions for each of the patterns considered. We have chosen to examine two connective verbs, namely *is/are* and *is/are called*. Examples from the ITU and GCSE corpora are considered separately. We have chosen not to discuss the Nature corpus because complex formal defining performatives are even more infrequent than simple formal defining performatives in the Nature corpus.

7.3.1 Complex formal defining expositives in the GCSE corpus

7.3.1.1 Pattern: 'Defining statement. This is called . . . .' in the GCSE corpus

In the GCSE corpus there is a tendency to avoid the use of long sentences and we assume that this is to facilitate comprehension. Definitions are frequently split into two sentences, with the defining statement appearing in the first sentence and the name of the term at the end of the second. The defining statement of the first sentence is replaced by the demonstrative pronoun at the beginning of the second sentence. When the second sentence consists solely of the expression *This is called* followed by a term (i.e. *X*) which is not modified by any modifier other than the indefinite article, the preceding sentence will almost always be a defining expositive. When the term in the second sentence is not modified by any determiner, it is likely to be a *process* or *method*. When the term in the second sentence is modified by the indefinite article, the term may be a *product, process* or *method*. In the first set of examples which follows, the term in the second sentence is unmodified.

(57) To get pure lines the plants are pollinated with their own pollen. This is called self-pollination.
(58) During sexual reproduction, a male sperm joins with a female egg. This is called fertilisation.
(59) Some plants and animals live closely together to help each other to survive better. This is called symbiosis.
(60) Some bacteria can change nitrogen gas into nitrates. This is called fixing nitrogen
(61) When the fruit grows, the cotyledon stays below the ground. This is called hypogeal germination.
(62) Pseudopodia can push out in any direction from anywhere on an amoeba. This is called amoeboid movement.
(63) Plants lose water into the air through the stomata in their leaves. This is called transpiration.

To transform any of these examples into simple formal defining expositives should require little computational effort. We allow the term in the second sentence to become *X* in our formula for simple formal defining expositives. We then insert the phrase *is a process whereby* before the main clause of the first sentence. This would result in the following rewritten simple formal defining expositives.
Rewritten expositives

(57a) Self-pollination is a process whereby plants are pollinated with their own pollen to get pure lines.
(58a) Fertilisation is a process whereby a male sperm joins with a female egg during sexual reproduction.

(59a) Symbiosis is a process whereby some plants and animals live closely together to help each other to survive better.
(60a) Fixing nitrogen is a process whereby some bacteria can change nitrogen gas into nitrates.
(61a) Hypogeal germination is a process whereby the cotyledon stays below the ground when the fruit grows.
(62a) Amoeboid movement is a process whereby pseudopodia can push out in any direction from anywhere on an amoeba.
(63a) Transpiration is a process whereby plants lose water into the air through the stomata in their leaves.

In the second set of examples below, the term in the second sentence is modified by the indefinite article.

(64) If a bar of iron is placed inside the coil it will become magnetic when the current is flowing. This is called an electromagnet.
(65) One of the H atoms in ethane is substituted by a bromine atom forming bromoethane. This is called a substitution reaction.
(66) If the ligaments are torn during an awkward fall, the bones may be pulled out of position. This is called a dislocation.
(67) Sometimes your brain is tricked and you don't see objects as they really are. (See figure 36.12.)This is called an optical illusion.
(68) Diseased kidneys can be replaced by healthy ones from other people. This is called a kidney transplant.
(69) Blood passes through your heart twice on its way round your body. This is called a double circulation.

There is no easy formula for converting these sentences to simple formal definitions. Some of the terms (e.g. *kidney transplant, double circulation*) could be described as processes and could be rewritten in the same way as the examples in the previous set of sentences. Others (e.g. *optical illusion, dislocation*) describe an effect and the defining statement in the first statement describes the cause. It is difficult to recommend how one might transform these. Nonetheless, the essential elements for a simple formal defining expositive are present and our main objective here was to demonstrate this.

7.3.1.2 Pattern: ' Heading. This is …' in the GCSE corpus

Frequently, in the GCSE corpus, a term is introduced on its own in what appears to be a heading, and the definition is provided in the following sentence. We say 'appears to be a heading' because the corpus does not contain any tags to indicate that this is what they are. We suspect, however, that this is indeed what they are and discuss our reasons for this assertion in greater detail in Section 7.5. The word or

phrase in the heading must be a term which matches one of the term formation patterns previously specified and the term may not be modified by any modifier other than the definite or indefinite article. The sentence which follows the heading must commence with *This is. This* effectively replaces the *X* in our formula for simple formal defining expositives. The second sentence must meet all of the conditions which we specified for the retrieval of simple formal defining expositives with the exception of the specifications for the slot fillers for *X* . Here, the slot filled by *X* is *This*.

(70) The iris This is a flat ring of muscle which controls the amount of light that enters the eye.
(71) The stomach This is a flexible bag of muscle that can enlarge to accept the food that arrives from the oesophagus.
(72) The liver This is a large red-brown organ situated in the abdomen beneath the diaphragm.
(73) The pancreas This is a gland which pours pancreatic juice into the duodenum.
(74) Rayon This is a synthetic fibre that is made from wood pulp.

While the noun phrases which follow the expression *this is* are sometimes modified by modifiers which might not appear to be part of the term (e.g. *flat, flexible, large*) we accept them because they provide additional information relating to the characteristics of the term. We believe that it would be relatively straightforward to transform the above examples into simple formal definitions by deletion of the word *This*.

7.3.1.3 Pattern: 'x. This is a ….'in the GCSE corpus

In the GCSE corpus, there are also many instances of terms being introduced at the very end of one sentence and being defined in the following sentence when the following sentence commences with *This is*. Here, the word which is introduced at the end of one sentence must be a term which matches one of the patterns previously specified and it may not be modified by any modifier other than the indefinite article. The sentence which follows must meet all of our specifications for a simple formal defining expositive with the exception of the specifications for *X* which in this instance, is replaced by *This*.

(75) . . . a ring mains system. This is a loop of cable that runs from the consumer unit round the house and back to the unit.
(76) . . . a mortgage. This is a long-term loan either at a fixed or a changing rate of interest.
(77) . . . a business enterprise. This is a company, set up in law to make its own deals, buying and selling with other people.

(78) . . . an industrial tribunal. This is a law court that deals only with legal matters to do with work.
(79) . . . an exoskeleton. This is a hard outer protective covering made of chitin.

The above examples convert very easily to simple formal definitions. This can be done by deleting *This* and conflating the two sentences (e.g. (77a) A business enterprise is a company, set up in law to make its own deals, buying and selling with other people).

7.3.2 Complex formal defining performatives in the ITU corpus

There is also evidence of complex formal defining expositives in the ITU corpus but retrieval of these is less straightforward than in the GCSE corpus. The definitions in the GCSE corpus are quite visible in the sense that the reader can tell immediately when a definition is being provided. This is because the defining expositives, whether simple or complex, stand on their own. In other words, they do not contain any information which is extraneous to the definition. We suspect that this is intentional because it makes the text more accessible to the reader. Complex formal defining expositives in the ITU corpus are often less visible. They may be embedded in sentences containing other types of information which may be a comment on the definition or may be completely unrelated to it. However, there are some patterns in the ITU corpus which are very similar to the patterns which we identified for complex formal defining expositives in the GCSE corpus. We have chosen to focus on these.

7.3.2.1 Pattern: 'Heading. This is…' in the ITU corpus

As in the GCSE corpus, whenever a term is introduced on its own in the form of a heading, the definition of that term is provided in the following sentence. Specifications for the sets of characters which may co-occur with headings are provided below. These allow us to identify when explicit definitions are being provided. When these sets of characters are absent, different conditions need to be specified, and these are as follows.

The word or phrase in the heading must be a term which matches one of the term formation patterns previously specified and the term may not be modified by any modifier other than the indefinite article. The sentence which follows the heading must commence with *This is* followed by a word or phrase which must be a term or one of the specified set of class words and may not be modified by any modifier other than the indefinite or definite article. We have not placed any restrictions on what should follow this word or phrase but it seems that what follows tends to be the same type of information, only more of it, as the information which was retrieved for simple formal defining expositives.

(80) preference indicator. This is an indicator contained within the forward call indicators parameter field of ISUP, sent in the forward direction indicating whether or not the user . . . .
(81) digital speech interpolation (DSI). This is a technique whereby advantage can be taken of the inactive periods during a conversation, creating extra channel capacity.
(82) mean holding time. This is the total holding time divided by the total number of seizures and can be calculated on a circuit group basis or for switching equipment.

The above definitions are longer than those retrieved from the GCSE corpus. For instance, sentence two in example (82) above has two consecutive main clauses all of which are part of the definition of the term *mean holding time*. Example (80) is also a rather complicated sentence but it too seems to contain only information which relates to the definition. It seems therefore that if a term appears as a heading, the sentence which follows will be definitional, regardless of its length. We need only specify that a sentence must be immediately preceded by a term with generic reference and must commence with *This is* and that *This is* must be followed by a term or one of a generic set of class words.

7.3.2.2 Pattern: 'Defining statement. This is called/known as . . .' in the ITU corpus
When a sentence commences with one of the following *This is called/This is known as* and the phrase is followed by a term which is not modified by any modifier other than the definite article or indefinite article, the whole of the previous sentence is likely to constitute a defining statement.

(83) Signals can be sent from the DCME to the exchange to busy-out part of the route when the quality criteria are violated. This is known as Dynamic Load Control (DLC) and can be an effective control method.
(84) In the case of both-way circuits, it may only be necessary to inhibit one direction of operation. This is called directionalization.
(85) Charging is by the minute and any fraction of a minute shall be charged as for one minute. This is known as one plus one.
(86) . . . for an impedance match at the point of interconnection and choose the value of this impedance to be equal to the design resistance of measuring instruments. This is known as the impedance matching technique (previously referred to as the constant electromotive force technique).
(87) . . . Network protocol address information: information encoded in a Network protocol data unit to carry the semantics of a Network service access point address. (This is known as an "address signal" or as the "coding of an address signal" in the public network environment).

What distinguishes some of these examples from the examples in the GCSE corpus is that the sentence which commences with *This is* may contain more information than just the term alone. Examples one and two above have an additional clause which are evaluative statements and add nothing to the definition. In the second last example, the term is followed by an indication, in brackets, of another term for the same technique. ITU policy in relation to round brackets is that ''indications in round brackets are qualifiers or alternative terms in general use in addition to the principal term.'' In fact, what we have here appears to be a deprecated term because the text specifies ''previously referred to'' rather than an alternative. ITU generally reserves square brackets for indicating deprecated terms. The last example is interesting because sentence 1 becomes a simple formal defining expositive if we substitute *is* for the semi-colon ': '. The following sentence is in round brackets and contains alternative terms for *network protocol address information* when it is used in a different context, i.e. in the public network environment.

7.3.3 Observations

Complex formal defining expositives appear to be quite common in the ITU and GCSE corpora. We have examined only a very small number of devices for identifying the presence of complex formal defining expositives but have found these to be very productive. While we feel that we have succeeded in demonstrating that such structures exist, we believe that we have only touched the tip of the iceberg and that this is an area which warrants a great deal of further investigation

7.4 Semi-formal defining expositives

In chapter four, section , we discussed Trimble's definition of a semi-formal definition. He states that ''by definition, a semi-formal definition contains only two of the three basic defining elements: the term being defined and the statement of differences. 'Semi-formal' refers to the form of the definition and indicates that it is not complete: the class is left out'' (1985:77). Trimble suggests that the class may be omitted because a term may be too high up in a conceptual hierarchy to warrant the assignment of a class word or a superordinate term. We would suggest that there may be other reasons for omitting the class word or superordinate term when semi-formal defining expositives occur in text. It may be that the class word or superordinate has already been specified in the text, perhaps in the previous sentence. Thus, what appears to be a semi-formal defining expositive may in fact be part of a complex formal defining expositive. Alternatively, the semi-formal defining expositive may complement a formal defining expositive which has already been

provided in the previous sentence(s). Here, the semi-formal defining expositive may specify additional differences, relating, for example, to the purpose, material or property of the term previously defined.

In this section we will examine how semi-formal defining expositives can be retrieved from the three corpora. The formula which we wish to identify is:

X = distinguishing characteristic(s)

Semi-formal defining expositives are deemed to be partial defining expositives because *Y* is absent from the definition. Below, we have specified the slot fillers for *X* and '=', and we have also provided specifications relating to the expression of the expositive. Where the specifications are the same as for formal defining expositives, the reader is referred to the appropriate section, in order to avoid repetition.

7.4.1 Specifying the slot fillers

7.4.1.1 Specifying the slot fillers for X

The specifications for *X* are the same as for simple formal defining expositives (cf. 7.2.1.1.1) except that *X* may be modified by the definite article as well as the indefinite article regardless of its position in the sentence, i.e. whether it appears at the beginning or at the end of the defining statement. When *X* is modified by the definite article this is generally an indication that it has already been introduced in the form of a caption or heading (cf. Section 7.5) or that it has been referred to by means of a paraphrase.

7.4.1.2 Specifying the slot fillers for '='

A much greater number of words may fill the '=' slot in semi-formal defining expositives than in formal defining expositives. Connective verbs in semi-formal defining expositives in the three corpora include the following:*contain(s), has, is/are used for, is/are used to, include(s), involve(s), is/are characterized by, is/are described as, produce(s), provide(s)*. We have chosen to restrict our investigation to the following connective verbs:*is used to, is used for, has/have*. These connective verbs are common to all three corpora.

7.4.2 The expression of semi-formal defining expositives

Some of the following specifications were already specified for the expression of formal defining expositives. When a specification has already been discussed, the reader is referred to the appropriate section for discussion of same.

The connective verb must appear in the main clause of the sentence and must not be prefaced by clauses or phrases other than those which fill the *X* slot. When the connective verb is the main verb it must be in the present tense, indicative mood, in active or passive voice. The connective verb may not be qualified by any modal verb other than the modal *can*. The defining statement may not be qualified by any focusing adverb other than *commonly, generally, usually*.

7.4.3 Examples of semi-formal defining expositives

All of the examples provided below meet the criteria specified for semi-formal defining expositives. To facilitate analysis, we have classified the examples according to the type of connective verb used and, for each connective verb, according to the class of word used to introduce the distinguishing characteristic.

Pattern 1:*X is/are used to* + distinguishing characteristic

This pattern is invariably followed by a verb in infinitive form. The connective verb *is/are used to* indicates the purpose or function of the term being defined. In most of the examples examined, *X* was not modified by any modifier other than the indefinite article. Where *X* is modified by the definite article, *X* appears as a caption immediately preceding the defining statement (cf. 7.5.2 for further discussion of these).

Examples from ITU corpus

(88) Digital transfer links are used to interconnect interface adaptors to form signalling data links.
(89) Analogue transfer links are used to interconnect data modems located within, or adjacent to, international switching centres, thus forming signalling data links.
(90) Graphic elements. Graphic elements are used to disiplay text, including symbols or pictures.
(91) Photographic elements. Photographic elements are used to render an image by the transmission and display of an array of individual picture elements (pixels)within an active drawing area.
(92) Customer sub-account number. The customer sub-account number is used to provide the card holder with telecommunications expense control where multiple PIN numbers are associated with a single primary account number.
(93) Call waiting tone. The call waiting tone is used to advise a subscriber who is engaged on a call that another subscriber is attempting to call.

Examples from GCSE corpus

(94) Oxygen is used to convert iron into steel.

(95) Flax is used to make linen which is very strong and hard wearing.
(96) A vaccine is used to encourage the body to make its own antibodies.
(97) Hardboard is used to make doors and cupboards.
(98) Microbes are used to produce cheeses and yoghurts from milk.
(99) Neutralization is used to remove carbon dioxide from the air in air-conditioned buildings.
(100) Nitric acid is used to produce fertilisers such as potassium nitrate and explosives like TNT (trinitroluene)and dynamite.

Example from Nature corpus

(101) Satellite measurements are used to quantify the atmospheric greenhouse effect, defined here as the infrared radiation energy trapped by atmospheric gases and clouds.

There was only one example of this pattern in the Nature corpus.

Pattern 2: *X is/are used for* + distinguishing characteristic
When the connective verb *is/are used for* is followed by a present participle, the defining statement describes the purpose of the term, *X*. When the connective verb is followed by a noun or lists of nouns, the noun or lists of nouns are either the products made of the material specified by *X* or the purpose of *X* when these nouns are deverbal nouns.

Examples from ITU corpus

(102) Configuration information is used for a network management data base at exchange level.
(103) The timing signal is used for synchronizing the sampling frequency of the analogue/digital converters producing the digital sound-programme signal.
(104) The public switched telephone network is used for carrying the telewriting information.
(105) The DCN is used for communications between central opertaions systems and distributed telecommunications centres.
(106) The reset procedure is used for recovering from a restart of a home location register.

Examples from GCSE corpus

(107) Autoclaves are used for killing bacteria on instruments needed for operations in hospital.

(108) Impure salt is used for deicing roads.
(109) PVC plastics are used for raincoats, coverings for tables and shelves (Fablon), floor tiles (vinyl tiles), upholstery, records and electrical installations.
(110) Rigid PVC is used for records and for gas and water pipes.
(111) Flexible PVC is used for toys and for insulating cables.

Examples from Nature corpus
This pattern does not occur in the Nature corpus.

Pattern 3:*X has/have* + distinguishing characteristic

Examples from ITU corpus

(112) The Featherset headset has an insert type receiver and a noise cancelling electret microphone which is held near the side of the mouth by a boom.
(113) The ISDN-UP has an interface to the SCCP (which is also a level 4 User Part) to allow the ISDN-UP to use the SCCP for end-to-end signalling.
(114) An (N)-connection-endpoint has an identifier, called an (N)-connection-endpoint- dentifier, which is unique within the scope of the (N + 1)-entity which is bound to the (N)-connection-endpoint.
(115) A service primitive has a direction which is either: a)from a service-user to the service-provider; b)from the service-provider to a service-user.
(116) Power spectrum SU32 has a code spectrum modified by the conditional coding rule compared to random ternary signalling.
(117) The special information tone has a tone period theoretically equal in length to the silent period.

Examples from GCSE corpus

(118) Stereo records have grooves that have different sides.
(119) A seed has a tough coating or testa around it.
(120) Graphite has a very high melting point.
(121) A female has a pair of X genes on her sex chromosomes while a male has an X and a Y gene.
(122) Red cells have haemoglobin in them.

Examples from Nature corpus

(123) The P400 protein has a relative molecular mass (Mr)of 250,000 (250K)1,3 and is phosphorylated4 in a cyclic AMP-dependent manner.
(124) Calpastatin has four repeated domains, each of which effectively inhibits calpain in vitro 38,39.

(125) The cDNA has a single open reading (ORF) coding for 177 amino acids that is preceded by a relatively short 5' untranslated region (UTR) of 14 nucleotides (exclduing the EcoRI 11-nker sequence) and is followed by a long 3′ UTR of 1,133 nucleotides excluding the poly(A)tail.

(126) The output light beam has the role of the axon, broadcasting the signal from each neuron.

7.4.4 Observations

Simple semi-formal defining expositives are much more common in all three corpora than simple formal defining expositives. Our examination of the corpora confirms that, contrary to what Trimble (1985) suggests, *Y* is unlikely to be omitted because *X* is too near the top of its conceptual hierarchy. Many of the terms defined in the above examples do not fit into this category. *Y* is much more likely to be omitted in the semi-formal defining expositive because it has already been specified previously in the text, and what appears to be a semi-formal defining expositive is in fact an extension of a simple formal defining expositive previously expressed. There are also many instances where it is omitted and there are no apparent reasons for doing so.

7.5 Dictionary type definitions

In addition to providing implicit definitions which are buried in the corpora, both the ITU and GCSE corpora provide definitions which are intended to be interpreted as such and are marked in a specific way in each of the corpora. We have chosen to call these dictionary defining expositives because their structure frequently resembles the structure of a dictionary definition. It is not surprising that authors should provide definitions in such an obvious way in these corpora as the function of the texts is informative and the authors are expected to have a greater level of expertise than their readers. They therefore provide definitions of terms which are perceived to be unknown to the reader. We have found no evidence of this type of definition in the Nature corpus which is what we might have expected given the assumption that author and reader have a similar level of expertise.

7.5.1 Dictionary defining expositives in the ITU corpus

The ITU corpus signals that a dictionary-type definition is about to follow by enclosing the term to be defined in one of the following sets of characters:1) #GR# term #AS#, or 2) #IT# term #IQ#. It appears that #GR# #AS# refers to the French

language equivalent for the emboldening feature (i.e. *gras*), and #IT##IQ# refers to the French equivalent for *italics*. This makes it relatively easy to retrieve dictionary defining expositives from the corpus. However, as the bold and italics features are also used for other purposes, we need to specify a certain number of conditions in order to retrieve only those which signal that a definition is about to follow. The conditions are broadly similar to those specified previously for the retrieval of formal and semi-formal definitions and they are:

1. The word or phrase which appears within the above sets of characters must be a term (henceforth referred to as *X*) which may not be modified by any modifier at all.
2. What follows #AS# or #IQ# must be a statement which commences with a term or one of our generic set of class words (henceforth referred to as *Y*) preceded by the indefinite article whereby the 'a' of the indefinite article must be in upper case.
3. It is not necessary for the defining statement to contain a connective verb. Where a connective verb is used, it does not have to be one of the connective verbs specified for the retrieval of formal definitions.
4. The defining statement may not be qualified by modals, focusing adverbs or negating words.
5. The distinguishing characteristic which follows the connective verb, or *Y* where no connective verb is used, must be introduced by one of the following: infinitive, present participle, past participle, preposition, relative pronoun.

Pattern: #GR# term #AS#

(127) #GR# Data acknowledgement (AK) #AS# A data acknowledgement message is used to control the window flow control mechanism which has been selected for the data transfer phase (UDT).
(128) #GR# restoration link equipment #AS# A transmission link equipment which is used for tranmission when the normal link equipment is not available.
(129) #GR# SCCP Route #AS# A SCCP route is composed of an ordered list of nodes where the SCCP is used (origin, relay(s), destination) for the transfer of SCCP messages from an orignating user to the destination user.
(130) #GR# main cable #AS# A cable used in the local line distribution network between the main distribution frame and a cross connection point.
(131) #GR# restoration control program #AS# A decision making programme which controls restoration processes.
(132) #GR# Videotex form #AS# A form is a frame where one or several fields are defined for the collectionof user data.
(133) #GR# byte #AS# A bit string that is operated upon as a unit and the size of which is independent of redundancy or framing techniques.
(134) #GR# Suspend message (SUS) #AS# A message sent in either direction indicating that the calling or called party has been temporarily disconnected.

(135) #GR# periodicity pattern #AS# A pattern which indicates which days are recording (or results output) days and which are not.
(136) #IT# Subsequent address message (SAM) #IQ# A subsequent address message (SAM) is used to transmit additional address signals not available when the initial address message is formed.
(137) #IT# Field #IQ# A field is a part of a window (sometimes the entire window area), which is used for entering or displaying information.
(138) #IT# channel gate #IQ# A device for connecting a channel to a highway, or a highway to a channel, at specified times.
(139) #IT# Digital terminal circuit section #IQ# A digital terminal circuit section comprises the two directions of transmission, for one equivalent voice-frequency signal, through a digital terminal.
(140) #IT# Telecommunications management network #IQ# A telecommunications management network (TMN) provides the means to transport and process information related to network operations, administration and maintenance.

In these examples, the information required to satisfy the formal defining expositive conditions is provided in one of two ways. It may be provided as a straightforward simple formal defining expositive where the *x, y* and '=' slots are filled (e.g. (137) A field is a part of a window (sometimes the entire window area), which is used for entering or displaying information). Alternatively, the *x* and '=' slots at the beginning of the defining statement may not be filled but the structure allows the reader to understand that they are implied, as the following example from the corpus illustrates.

(141) #GR# main cable #AS# A cable used in the local line distribution network between the main distribution frame and a cross connection point.

We also find instances of semi-formal defining expositives, i.e. where *Y* is not stated, as the following example illustrates.

(142) #IT# Subsequent address message (SAM) #IQ# A subsequent address message (SAM) is used to transmit additional address signals not available when the initial address message is formed.

7.5.2 Dictionary defining expositives in the GCSE corpus

We were working with an electronic version of the GCSE corpus and did not have access to the original paper version of the corpus. We suspect that when the corpus was converted to electronic form, the original line breaks in the corpus were not retained, and we have recently received confirmation this is indeed what happened

(private communication from Professor John Sinclair). It appears that the original texts may have contained headings which contained a term which were followed on the next line by a dictionary type definition of that term. We have drawn this conclusion because it is the only logical explanation for some patterns in the corpus which, if they appeared in a single continuous line and in the same sentence, would be ungrammatical. These patterns do not have any main verb. They are very similar to conventional dictionary definitions where the dictionary entry is followed by a phrase which describes the term. To illustrate what we mean, some examples are provided below with the exact same layout as in the corpus.

(143) Fungus a simple plant which has no chlorophyll and which is made of hyphae (plural: fungi).
(144) Element a chemical substance which cannot be broken down by chemical reactions into a simpler substance; carbon, nitrogen and oxygen are elements.
(145) Flagellum a long thread growing from a cell; used to cause movement (plural: flagelli).
(146) Stoma a small pore in a leaf (plural: stomata).
(147) Hypha a very small thread-like part of a fungus (plural: hyphae).

As there is no punctuation between what we perceive to be the term and the defining phrase, the resultant syntactic pattern contravenes normal noun phrase rules. Thus, we can stipulate that wherever a sentence begins with a single word term and is not followed by any punctuation mark but is followed immediately by a definite article, this sentence is likely to constitute a dictionary-type definition. Computationally, this pattern is relatively easy to retrieve.

There are further examples of dictionary defining expositives where the term to be defined and the phrase which follows together contravene noun phrase rules. These include patterns where the term is a single noun and is followed by an adjective or a cardinal number, as the following examples illustrate.

(148) Villi small finger-like structures in your ileum which absorb digested food (singular: villus).
(149) Atrium one of the two top chambers in the heart (plural: atria).
(150) Vertebrae small bones making up the backbone (singular: vertebra).

These can be retrieved by stipulating that a sentence must begin with a single word term and must be followed immediately (i.e. not separated by a punctuation mark) by an adjective or cardinal number.

There are other instances of this type of defining expositive where the term and the head of the defining phrase together do not contravene noun phrase rules and could therefore be perceived to be a single noun phrase rather than as two nouns to

be read separately, as the following example illustrates:

(151) *Bacteria organisms made of one cell: they are neither animals nor plants (singular: bacterium).

''*Bacteria organisms'' corresponds to a valid term formation pattern but the unit is clearly not a term; bacteria and organisms should be interpreted as two separate terms. There is no simple means of identifying this particular pattern except when it is preceded in the corpus by the phrase *Important Words*. This phrase frequently appears before the expression of a dictionary defining expositive; unfortunately, it is not always present.

7.6 Conclusion

In this chapter, a set of conditions for the retrieval of simple formal definitions from all three corpora was specified. It was necessary to specify the conditions in order to exclude the retrieval of statements which were not formal defining expositives. There was ample evidence of simple formal defining expositives in the ITU and GCSE corpora but little evidence of this type of expositive in the Nature corpus. This confirmed our expectation that defining expositives were more likely to be provided in informative texts where the level of expertise of author and reader is different. A set of conditions for the retrieval of complex formal defining expositives was also provided. These are defining expositives which satisfy the criteria for formal definitions but which are expressed in more than one sentence. The search for these cross-sentence defining expositives proved to be very fruitful, and it is an area which warrants further investigation.

Semi-formal defining expositives, expositives where the superordinate or class word is omitted, were present in all three corpora. We devised a set of conditions for retrieving this type of defining expositive. As the conditions are less restrictive than those specified for the retrieval of formal defining expositives, further processing would be required to transform these into proper defining expositives. We feel, nonetheless, that it was useful to attempt to retrieve this type of defining expositive because it is used quite frequently in all three corpora. We also found that many semi-formal defining expositives were in fact extensions of simple or complex formal defining expositives, and could serve as a complement to these in the formulation of an adequate terminological definition. In chapter nine we will investigate how information about a particular term which is expressed both formally and informally can be combined to produce a terminological definition.

In the search for formal and semi-formal defining expositives which were implicit

in the corpora, we noted that both the ITU and GCSE corpora contained explicit defining expositives, i.e. dictionary defining expositives which were explicitly flagged. We described how it would be possible to identify these in the ITU corpus, using the tags which had been used to signal the presence of these defining expositives. The retrieval of such defining expositives from the GCSE corpus was more difficult because the original layout of the texts was not preserved when the texts were converted to electronic form. We suggested that, in the original texts, the term was likely to have been presented in the form of a heading with the definition following immediately afterwards. We made some proposals for identifying where these occur in the corpus.

8 Synonymy, substitution and paraphrasing

8.1 Introduction

Chapter seven focused on the design of specifications for the retrieval of formal and semi-formal defining expositives from the three corpora. These specifications were deliberately restrictive and were designed to retrieve only those syntactic patterns which corresponded to formal and semi-formal defining expositives. However, the connectives which were specified in the previous chapter and other connectives (connective phrases) which will be discussed in this chapter are also used to provide information which does not necessarily correspond to a formal or semi-formal defining expositive. As this chapter will demonstrate, connective phrases are potentially a very rich source of additional information. They may serve to complement information which has already been retrieved about a particular term, or terms, using the specifications for formal and semi-formal defining expositives or they may be the sole source of information about a term, or terms, in a corpus, in which case they may serve as a starting point for the formulation of a definition. Whichever the case, it is very important that connective phrases should not be overlooked in the design of a corpus-based method for the retrieval of terminological information.

Connective phrases are used to signal the presence of any one of a number of relations; the relation may be one of synonymy, it may be one of equivalence, or it may be a genus-species relation. As terms such as synonymy and equivalence are not used unambiguously in the literature, we start by defining what we mean by these terms. We explore the concept of synonymy by discussing dictionary definitions of synonymy and definitions that have been proposed by *inter alia* Trimble, Landheer, Carter and ISO. It will be noted that almost all of these definitions, with the exception of Carter's, focus on equivalence in meaning alone and do not address the issue of equivalence of usage. We will suggest that true synonymy is in fact extremely rare if words are examined in terms of the company which they keep but that it may nonetheless be a useful lexicographic device for explaining a word or phrase in terms of another which has the same meaning. The concepts of non-formal definition and relations of equivalence as defined by Trimble, Barnbrook and Sinclair are also discussed. We find that Trimble's non-formal definitions may

account for synonymy and genus-species relations. We find that Barnbrook and Sinclair's definition of equivalence which includes explanation by means of paraphrase and substitution may allow us to account not only for the formal and semi-formal defining expositives discussed in the previous chapter but also for some of the patterns discussed in this chapter.

The connective phrases which we have identified as possibly signalling the presence of additional information include phrases, punctuation marks and abbreviations such as *i.e., e.g*. Whereas in the previous chapter, we chose to specify quite restrictive conditions in order to minimize the retrieval of invalid statements, we have chosen in this chapter simply to present the connective phrases with what appear to be their typical collocation patterns in order to establish 1) whether different connective phrases are used for different purposes and 2) how the output might be refined to eliminate 'uninteresting' information. Apart from specifying that the output must contain one of the connective phrases, we have not imposed any restrictions on the type of pattern which can be retrieved. The manner in which the information is presented in this chapter therefore contrasts quite sharply with the very formal presentation in the previous chapter. We have chosen to present the information in this way in an attempt to underline the enormous potential of a corpus-based method for retrieving information about terms.

8.2 Defining our terms

8.2.1 Synonymy

This section will examine some definitions of synonymy and outline some of the problems associated with the definitions which we have found in the literature. We start by looking at the Cobuild dictionary definition of synonym, followed by ISO, Landheer, Trimble and Carter.

> A **synonym** is a word or expression which means the same as another word or expression. (*Collins Cobuild English Language Dictionary* 1987)

From this, we infer that what is important for establishing synonymy is equivalence of meaning. This is borne out by the following examples of synonyms from the Cobuild dictionary where the = (equals) sign is used to signal equivalence of meaning. The entries for *afraid* and *fearful* each cite *frightened* and *scared* as synonyms. The entry for *frightened* cites *scared* as a synonym, while the entry for *scared* cites *afraid* as a synonym. (Interestingly, *fearful* is not cited as a synonym for any of the other three words). It would appear therefore that *afraid, frightened* and *scared* are

to be interpreted as being synonymous words in terms of their meaning because each of them is cited as a synonym for the other. However, we would argue that the similarity between these words ends there.

First, there are grammatical differences in the sense that the grammatical rules governing the use of these are not always identical. For example, *afraid* is predicative and never occurs immediately after the word which it is qualifying. *Frightened, scared* and *fearful* can all appear either as attributive or predicative adjectives. As *afraid* is never used as an attributive adjective, it is therefore not possible to state that, in terms of its grammatical rules, *afraid* is synonymous with *frightened, scared* and *fearful.*

Second, there may be another difference; these words may not always be used in the same contexts. For example, the Cobuild dictionary specifies that the use of *fearful* when it means *afraid* is a formal use which means that they are unlikely to be interchangeable. The contextual difference is discussed further below in our analysis of the ISO definition of synonymous and quasi-synonymous terms.

> **5.4.3 synonymy**: Relation between designations (5.3.1) representing only one concept (3.1) in one language.
>
> EXAMPLE
> sodium chloride; NaCl
>
> NOTE–Terms (5.3.1.2) which are interchangeable in all contexts (6.1.5.7) of a subject field (2.2) are called synonyms (5.4.3); if they are interchangeable only in some contexts (6.1.5.7), they are called quasi-synonyms. (*ISO 1087 Vocabulary of Terminology* 1990:5)

In its definition, ISO refers to situations where terms are equivalent in meaning and in usage. Synonymous terms are interchangeable in all contexts of a subject field. A context is defined as ''text or part of a text in which a term (5.3.1.2) occurs'' (*ISO 1087*:10). We wish to adopt this particular definition for our purposes. However, ISO's distinction between synonyms and quasi-synonyms is not very clear, and, as no examples are provided, we have chosen not to use this particular label.

Landheer (1989) defines a synonymous relationship as a bilateral relationship, one where the left hand side (lhs) and right hand side (rhs) are equivalent in meaning and where one side can be substituted for the other without loss of meaning.

> un rapport synonymique étant un rapport d'inclusion bilatérale, symétrique. Donc (4) Un vélo est une bicyclette au même titre que (5) Une bicyclette est un vélo. (Landheer 1989:140)

> Translation: as a synonymous relation is symmetrical, a relation of bilateral inclusion, a (4) *vélo* is a *bicyclette* in the same way as (5) a *bicylette* is a *vélo*.

He makes a distinction between *synonymie absolue* and *synonymie approximative. Vélo* and *bicyclette* are absolute synonyms while *livre* and *bouquin* (an informal or colloquial word for *livre*) are only 'approximate' synonyms. *Vélo* and *bicyclette* are interchangeable while *livre* and *bouquin* are not. In a dictionary, *bouquin* will be defined as *livre* but the reverse will not be the case. There are two reasons for this. One is that *bouquin* is an informal word so *livre* is unlikely to be defined as *bouquin*. The second reason is that *livre* can function not only as a term with its own referent but also as a class-word which is the superordinate for *livre, bouquin* and similar publications. Either of these reasons would be sufficient to prevent the terms from being interchangeable. Although Landheer does not state this explicitly, it is possible to infer that there is absolute synonymy when two words have the same referent and are in a two way replacement relation whereas there is 'approximate' synonymy (synonymie approximative) when two words have the same referent but are not interchangeable. We would add that even when there is what Landheer terms 'absolute synonymy', his definition does not account for differences in usage and that there is no guarantee that because two terms have the same referent, they will be interchangeable in terms of their usage.

Trimble (1985) does not define synonym *per se* but includes synonym in his definition of non-formal definitions:

> The function of a non-formal definition is to define in a general sense so that a reader can see the familiar element in whatever the new term may be . . . . Most non-formal definitions are found in the form of synonyms; that is, they attempt to substitute a word or phrase familiar to the reader for one presumably unfamiliar. (Trimble 1985:78)

Trimble's definition is provided in the context of his definitions for formal and semi-formal definitions (cf. Section 4.6). A non-formal definition gives the reader two kinds of information: the name of the *term* being defined and another word or phrase having the approximate meaning of the term. Trimble adds that the term and the word are not necessarily interchangeable and he uses the example *An arachnid is a spider* to illustrate this point. Regardless of the fact that Trimble has probably made up the example, it demonstrates his point that the two are not interchangeable because the term that is being defined (i.e. *arachnid*) is not on ''the same level of generality'' as *spider*. In fact, what is being provided is closer to a genus-species relation. Another example which he provides, *Native means indigenous*, seems to us to be closer to a synonymous relationship. In the *Collins Cobuild English Language Dictionary*, the entry for *indigenous* lists *native* as synonym but *indigenous*

is not listed as a synonym for *native* which means that it may not be unlike the *livre/bouquin* relationship, in the sense that *indigenous* may have a dual function, i.e. as a class-word and as an ordinary word with its own referent.

Trimble, like Landheer, appears to distinguish between absolute synonymy where a word or phrase is substituted by another which is likely to be more familiar to the reader but having the same meaning (e.g. *native, indigenous* above), and approximate or quasi-synonymy where both words have approximately the same meaning (e.g. *arachnid, spider* above). As with the previous authors, Trimble does not address the issue of equivalence of usage.

Carter defines synonymy as follows:

> *Synonymy* – is essentially a bilateral or symmetrical sense relation in which more than one linguistic form can be said to have the same conceptual or propositional meaning. This does not mean that the words should be totally interchangeable in all contexts; but where synonyms are substituted changes in the propositional meaning of the sentence as a whole do not occur . . . . However, stylistic differences limit substitutability.
>
> (Carter 1987:19)

Carter, too, asserts that synonymy implies equivalence of meaning, a bilateral relationship between the lhs and rhs, but points out that factors such as style or context may preclude substitutability. Two words or terms may indeed have the same referent but may not necessarily be used in the same contexts. The earlier Cobuild examples of *afraid* and *fearful* illustrate this very well, as do Landheer's examples of *livre* and *bouquin.*

To summarize the various points of view presented here, there appears to be a general consensus that a synonym is a word or phrase which has the same referent as another word or phrase, and when used as a defining device, is a means of explaining one word in terms of another. Landheer makes a distinction between absolute and approximate synonymy. Carter addresses the issue of synonymy in relation to usage and specifies that synonymous words are not necessarily substitutable in context. ISO defines synonymy in terminology and stipulates that synonymous terms are indeed interchangeable. We would support the notion that synonymy denotes equivalence of meaning and would agree with Carter that this does not necessarily imply substitutability in terms of usage when we are dealing with general language words. However, when we are dealing with terms, we subscribe to the ISO definition, namely that synonymous terms are interchangeable. For the purposes of this investigation, we propose to adopt the following definition of *synonymy*: a synonym is a term which means the same as another term used in the same communicative setting.

8.2.2 Equivalence

Barnbrook and Sinclair (1995), in their analysis of Cobuild dictionary entries, define an equivalence relation as follows:

> Essentially the two parts of the sentence are held to mean the same thing. (From equivalence arise the two powerful notions of *paraphrase* and *substitutability*. Paraphrase is defined as the replacement of a word by its definition, or vice versa. Substitutability is defined as a segment of text which stands in an equivalence relation with another. (1995:8)

As Barnbrook and Sinclair suggest, paraphrase and substitutability are two methods of stating equivalence. A paraphrase may be the replacement of a word by its definition, in which case we expect it to contain a superordinate and one distinguishing characteristic. This was also the pattern which we were seeking when attempting to retrieve formal defining expositives from the corpora in the previous chapter where restrictive conditions were specified for retrieving these expositives. However, paraphrasing is also used with some of the connectives discussed in this chapter even when it does not necessarily meet the felicity conditions specified previously. Substitutability, where ''a segment of text stands in an equivalence relation with another'' (Barnbrook and Sinclair 1995:8) is another means of expressing equivalence. Barnbrook and Sinclair do not specify the segment of text but, for our purposes, we assume that it may be a word or an entire phrase which stands in an equivalence relation to another word or entire phrase. Within the framework of this investigation, substitutability is not to be confused with synonymy. In cases of synonymy, the two words are interchangeable within a particular communicative setting. In cases of substitutability, the two words or phrases are not to be considered as interchangeable. For example, a term may be explained in terms of its general language equivalent as the following example from the GCSE corpus illustrates: 'small hairs *called* cilia'. Here, the referent is the same in both cases but only one of the two, i.e. the term, is appropriate in a specialized communicative setting.

Ogden and Richards (1923:110) suggest that when we wish to define *words*, we use substitution and when we wish to define *things*, we use 'real' definitions. Substitution involves using one lexical item in an equivalence relation to another. Thus, one might say that *to be afraid* is *to be scared*. Equivalence is established but the meaning of the phrase is not provided. Ogden and Richards' substitution appears to be not unlike what has been defined as synonymy for the purposes of this investigation.

In this chapter, we propose to use the term *synonym* to describe a word or expression which has the same referent as another word or expression in the same

communicative setting. We propose to use Barnbrook and Sinclair's definition of substitutability to describe situations where a segment of text is equivalent to but not necessarily synonymous with another and their definition of paraphrasing to describe situations where a term is replaced by its definition.

8.3 In search of synonyms, paraphrase and substitution

In the previous chapter, we called the connectives which linked the lhs and rhs of formal and semi-formal definitions *connective verbs*. In chapter six, we used the term *linguistic signals* to describe words or phrases which signalled the presence of terms. For reasons of clarity, we have chosen to use the term *connective phrase* to denote those words (or punctuation marks) which signal the presence of synonyms, paraphrases, or substitution. *Linguistic signals* and *connective phrases* intersect but they are not identical and this is why we have chosen to call them by different names. Moreover, even when there is considerable overlap between the two sets, they may play quite a different role in each set.

An analysis of each of the three corpora revealed that when certain connective phrases were present, it was sometimes possible to conclude that the words or phrases which co-occurred with these were in some way equivalent, whereby equivalence includes relations of synonymy, paraphrasing and substitution. On the basis of initial results, it appeared that it would simply be sufficient to retrieve all pairs of words or phrases co-occurring with each of these connective phrases and to replace the connective phrases with the connective verbs *is/are* in order to produce a statement of synonymy or substitution. Unfortunately, this proved not to be the case. In many situations where the connective phrases are apparently being used to denote a relation of equivalence, they are in fact functioning as connective phrases of genus-species relations, with the superordinate appearing on the left hand side and the subordinate appearing on the right hand side of the connective phrase. In other instances, what appear to be cases of synonymy are in fact cases of substitution where one of the referents is a general language word.

The connective phrases which we propose to discuss include: *i.e., e.g., known as, called, (*)*. Table 1 contains a list of these connective phrases, together with the total number of occurrences of these connective phrases in each of the corpora and the number of occurrences per 100,000 words. We propose to examine and exemplify each of the connective phrases in turn in order to identify with which relations they are associated. As the examples provided from the corpora are not all of equal length, the connective phrases appear in bold type in order to make them more visible. Where necessary, and possible, we will attempt to specify some selection restrictions in order to refine the output.

Table 1. Connective phrases

Indicator	ITU corpus		GCSE corpus		Nature corpus	
	Total	100,000	Total	100,000	Total	100,000
i.e.	2,198	46.77	44	4.4	0	0
e.g.	3,135	66.7	105	10.5	0	0
called	2,473	52.62	877	87.7	20	8.71
known as	108	2.3	170	17	11	4.79
the term	410	8.53	7	0.7	1	0.44
(*)	66,999	1,425.51	5462	546.2	3423	1,490.36

8.3.1 Analysis of the connective phrase i.e.

This connective phrase occurs only in the ITU and GCSE corpora. It does not appear to be used to express genus-species relations and is used only to express relations of equivalence. These can be expressed in one of four ways:1) phrase or clause on lhs of connective phrase followed by equivalent phrase or clause on rhs of connective phrase; 2) phrase or clause on lhs of connective phrase followed by appropriate term on rhs of connective phrase; 3) term on lhs of connective phrase followed by equivalent word or term on rhs of connective phrase; 4) term on lhs of connective phrase followed by equivalent phrase or clause on rhs of connective phrase. The total number of occurrences of the connective phrase in the two corpora is specified in brackets.

1) Phrase or clause on lhs of connective phrase followed by equivalent phrase or clause on rhs of connective phrase

Examples from GCSE corpus (44)

(1) By insisting on a logging cycle of 20 years, i.e. leaving an area alone for 20 years after tree felling.
(2) By keeping cattle or sheep at low densities, **i.e**. few animals per hectare, as in Australia, the grass was able to support some farming.

Examples from ITU corpus (2,198)

(3) User information is transferred in both directions simultaneously, **i.e**. both terminals are simultaneously a source as well as a sink.
(4) the ability to distinguish between destinations that are easy to reach (ETR)and destinations that are hard-to-reach (HTR), **i.e**. destinations with a low answer bid ratio . . . .

The above contain examples of substitution, where one phrase stands in an equivalence relation to another. In both the ITU and GCSE corpora, the substitutable phrase or clause appears on the rhs of the connective phrase. In the GCSE corpus, the substitutable phrase on the rhs of the connective phrase corresponds to a noun phrase on the lhs (i.e. *a logging cycle of 20 years, cattle or sheep at low densities*). In the ITU corpus, the substitutable phrase or clause on the rhs corresponds to a clause on the lhs. In the examples from the GCSE corpus, the substitutable segment on the rhs of the connective phrase corresponds to a noun phrase (i.e. *a logging cycle of 20 years, cattle or sheep at low densities*) on the lhs. In the examples from the ITU corpus, the substitutable statements (phrase or clause) on the rhs correspond to a clause on the lhs. Specifications for retrieving this pattern would therefore have to account for both clauses and noun phrases on the lhs and rhs of the connective phrase.

2) Phrase or clause on lhs of connective phrase followed by appropriate term on rhs of connective phrase

Examples from GCSE corpus

(5) But how can solids change to liquids (i.e. melt) or liquids to gases (**i.e**. boil or evaporate)?
(6) If the eye moves to look at a nearer object, the lens must become fatter, **i.e**. more convex

Example from ITU corpus

(7) The acceptance of the call by the terminating user, **i.e**. answer, causes the indications to be removed. During any one call, message flow is in one direction only, **i.e**. simplex working.

The term, which always appears on the rhs of the connective phrase, is explained by means of paraphrasing or substitution on the lhs. In the examples from the GCSE corpus, the term on the rhs corresponds to a clause on the lhs while in the ITU corpus, it corresponds to either a noun phrase or a clause. We specify for this and all subsequent patterns with the connective phrase *i.e*. that the word or phrase which appears on the rhs must be a term and that the cut-off point for identification of the right hand side is a punctuation mark such as a full stop, comma or brackets.

3) Term on lhs of connective phrase followed by equivalent word or term on rhs of connective phrase

Examples from GCSE corpus

(8) Energy is needed for muscles to contract (i.e. shorten).
(9) The noble gases all exist as separate single atoms (**i.e**. monatomic molecules).

Examples from ITU corpus

(10) The ability to simulate motion (**i.e**. animation) is a potential enhancement that can be achieved by several means.
(11) As for any other service or product, the establishment of a tariff, **i.e**. a sales price for telecommunications services . . .

In both the ITU and GCSE corpora, the terms being explained can appear either before or after the connective phrase. These appear to be cases of substitution where the term is explained in terms of an equivalent word or phrase drawn from general language. The term appears after the connective phrase when the connective phrase is enclosed in brackets and before the connective phrase when there are no brackets. The grammatical categories of the words or phrases which appear on the lhs and rhs are the same. Thus, where the connective phrase is preceded by a verb, it is also followed by a verb. The same applies to noun phrases and nouns where their number will also correspond. We may therefore be able to specify, for the identification of this pattern, that the grammatical category and number on both sides of the connective phrase must correspond. This specification may be particularly appropriate when the connective phrase and term are enclosed in brackets.

4) Term on lhs of connective phrase followed by equivalent phrase or clause on rhs of connective phrase

Examples from GCSE corpus

(12) (g) for gas and (aq) for an aqueous solution (i.e. a substance dissolved in water).
(13) workers were paid piece-rates, **i.e**. the wage was based on how many they made.
(14) These stones are quarried, **i.e**. removed from the ground by careful blasting which must not shatter the rock.
(15) In this case there is a double recessive gene, **i.e**. there is no dominant gene to mask the abnormal one.

Examples from ITU corpus

(16) a quiet code, **i.e**. a PCM signal corresponding to decoder output value number 0 (-law) or output value number 1 (A-law) (with the sign bit in a fixed state . . . .

(17) An all-zero signal, **i.e**. a signal unit consisting of 20 zeros with the correct check bits, may cause a discontinuity in the transmitted signal unit sequence.

The above examples of paraphrases and substitution contradict what was specified for the previous set. The word or phrase contained within the brackets is not necessarily the term being explained and, in this set of examples, the term which is being explained always appears on the lhs of the connective phrase, whereas in the previous set, it always appeared on the right. To retrieve these examples, we could specify that the word or phrase which appears on the lhs must be a term which matches one of the term formation patterns previously specified.

8.3.2 Analysis of the connective phrase e.g.

This connective phrase occurs only in the ITU and GCSE corpora. It is used *inter alia* to indicate the following: 1) superordinate term on lhs followed by one or more subordinate terms on the rhs (expression of a genus-species relation); 2) effect on lhs followed by cause on rhs. It does not appear to be used to signal the existence of a synonym or of a substitutable phrase. The term being explained will always appear on the lhs of the connective phrase and the explanation or subordinate term on the rhs.

1) superordinate term on lhs followed by one or more subordinate terms on the rhs

Examples from GCSE corpus (105)

(18) Cells or organs which can detect stimuli (e.g. smell, temperature, touch, taste, sound, light).
(19) If the ash contains oxides of reactive materials (**e.g**. sodium oxide and calcium oxide), it will form an alkaline solution with water.
(20) Now fill a shallow tray with water and sprinkle fine powder (**e.g**. talcum or lycopodium) on the surface.
(21) terracing of the land, plans to reduce wool consumption by using other sources of energy (**e.g**. wind power) and attempts to train local people in improved farming methods.
(22) Row crops (**e.g**. strawberries) can be protected from frost by covering them with plastic cloches or covers.
(23) if you know: that you can separate mixtures by using their properties, **e.g**. solubility, density, particle size, whether they are magnetic or not, any special
(24) cell types, **e.g**. root-hair cell, egg cell (ovum), sperm cell, muscle cell, skin cell, leaf cell
(25) Mammals (**e.g**. cats, horses, people) fertilise and grow the egg inside the body.

Examples from ITU corpus (3,135)

(26) measured (or specified) items of equipment (**e.g**. feeding-bridges, cable pairs, audio inputs to channel translating equipment, etc.)
(27) a number of methods of estimating the coefficients (**e.g**. least squares, varying parameter methods, nonlinear regression, etc.)
(28) maintenance information (**e.g**. line errors)
(29) some are steady-state impairments (**e.g**. loss, noise, quantization distortion, phase jitter, harmonic and intermodulation distortions, envelope delay distortion, echo, and attenuation distortion)

All of the above are examples of genus-species relations. A superordinate term or generic class word is explained by means of one or more subordinate terms. In order to retrieve this pattern, we could specify that the word or phrase on the lhs of the connective phrase must be a term or generic class word and that the words or phrases which appear on the rhs must be terms which correspond to the term formation patterns previously specified and must correspond to the same grammatical class as the word which precedes the connective phrase. It should, however, be noted that the superordinate term does not always appear immediately before the connective phrase. It may also appear in the following way (example from GCSE corpus):

(30) Some antibiotic drugs are produced in laboratories, **e.g**. chloramphenicol: used to combat typhoid; tetracycline—used against a wide range of . . .

where *antibiotic drugs* rather than *laboratories* is the superordinate. While this pattern appears to be quite rare (only 1 example in the GCSE corpus), it is difficult to imagine how one might ensure the correct reading. As *laboratories* corresponds to a valid term formation pattern, it is eligible for consideration. We suspect that in the absence of some form of semantic tagging, it would not be possible to prevent mismatches such as this one.

2) effect on lhs followed by cause on rhs

Example from GCSE corpus

(31) An increase in temperature (e.g. from direct sunlight) increases the rate of evaporation of water.

This was the only occurrence of this pattern in the GCSE corpus.

Examples from ITU corpus

(32) exceptional circumstances causing a major degradation or disruption of service (**e.g**. natural disasters, strikes, facility outrages, etc.)
(33) For short duration failures, **e.g**. solar interference on satellite network management plans may be the complemented by the detection of malfunctioning of particular items, **e.g**. loss of power supply, loss of incoming signal, loss of frame alignment.

While this pattern which specifies a cause-effect relation, whereby the effect appears on the lhs and the cause on the rhs, was quite common in the ITU corpus, it is difficult to specify how it might be identified. It would not be correct to state that the connective phrase is always preceded by a term because to do so would eliminate many valid instances from consideration. However, it would be useful to specify that the grammatical class of the words/phrases or terms which appear before and after the connective phrase should be the same.

8.3.3 Analysis of the connective phrase called

In the previous chapter, we examined the connective verb *is/are called* for the retrieval of formal and semi-formal defining expositives. Here, we are interested in examining other occurrences of the connective *called*, i.e. instances where it does not co-occur with *to be*. An attempt to retrieve all such occurrences from the ITU corpus proved to be problematic because *called* is also frequently used as an adjective to refer to the person or party who has been called, e.g. *the called subscriber, the called party*, or as a past participle of the verb meaning *to ring* (using a telephone). Further analysis showed that it was possible to prevent these readings from being retrieved. In order to exclude the adjective *called*, and the past participle of the verb *to call* (meaning *to ring*), we have specified that *called* may not be immediately preceded by any of the following: the definite or indefinite article, a demonstrative adjective, a preposition, an adjective or adverb. Attempts to retrieve instances of *called* functioning as a connective phrase proved to be less problematic with the GCSE corpus. There are some instances where *called* is being used as a verb in the perfect tense but, as the subject of the sentences where this occurs is generally a personal pronoun or a proper name, it should be relatively easy to eliminate occurrences of *called* functioning in this way.

Called appears in the following phrases:*often called, also called, called, usually called, generally called, sometimes called, (usually called), (often called*). While we excluded the focusing adverbs *often* and *sometimes* from our deliberations in the previous chapter because they restricted the scope of formal and semi-formal defining expositives, we have chosen to include them here because they serve as useful indicators of other types of information.

Called can be used to signal the presence of one of the following:1) general language word on lhs followed by appropriate term on rhs; 2) term on lhs followed by synonym on rhs of connective phrase; 3) superordinate term or phrase on lhs of connective phrase followed by subordinate term on rhs. The latter category contain further examples of genus-species relations.

1) general language word or phrase on lhs followed by appropriate term on rhs

In the GCSE corpus in particular, terms which appears on the rhs of the connective phrase are frequently preceded on the lhs by an equivalent word or phrase, drawn from general language, with which the reader is likely to be more familiar. The term and general language word or phrase are to some extent equivalent but they are not intended to be understood as being interchangeable. These are all examples of substitution. In this particular context, substitution allows the author to achieve two objectives: to provide the reader with an indication of the meaning of the term by using a known lexical item in the same sentence; to tell the reader what the correct technical term is in the particular context.

Examples from GCSE corpus (877)

(34) organisms are built of 'bricks' called cells. Because cells are so small we
(35) Amoeba. It swims using short hairs **called** cilia. The cilia also <P 19> make a
(36) spores are made inside spore cases **called** sporangia. When the sporangia burst
(37) the support. Peas have small shoots **called** tendrils which grip the support. (>
(38) be fixed to the stem with a stalk **called** a petiole. Leaves have three main >
(39) and gets inside through a tiny hole **called** the micropyle. Then the male gamete
(40) by a green substance in leaves **called** chlorophyll. Some of the sugar made >

The nature of the author-reader relationship in the GCSE corpus requires that the author use words which are known to the reader in order to introduce new terms. We found no occurrences of this pattern in either the ITU corpus or the Nature corpus which is understandable as the author is more likely to use known equivalent or superordinate terms, rather than general language words, to introduce new terms.

2) term on lhs followed by synonymous term on rhs of connective phrase

Examples from ITU corpus (2,473)

(41) in Fig. 5.8b, where a dialectric plate (or metallic fin) called a half-wave plate is inserted in a circular waveguide.
(42) of the continuous-time signal. A low-pass filter (often **called** anti-aliasing filter) is needed to prevent aliaisng caus

(43) our-phase (4-PSK). Modulation with two phase conditions, **called** bi-phase modulation (2-PSK), or with eight phase conditi
(44) by the satellite in separate non-overlapping time slots **called** bursts in which information (e.g. PCM telephony) is buff

In the preface to the ITU corpus, the authors specify that round brackets are used to indicate the existence of synonyms. It is therefore possible to infer that if the word or phrase which precedes the connective phrase and the word or phrase within the connective phrase correspond to term formation patterns, the two terms are synonymous. This particular pattern does not occur in either of the other two corpora.

3) superordinate term or phrase on lhs of connective phrase followed by subordinate term on rhs

Examples from GCSE corpus

(45) we must use a special microscope called an electron microscope. The electron
(46) the extra water as a weak solution **called** urine. Sea water is stronger than
(47) growing points of a seed. Chemicals **called** enzymes digest the stored food and
(48) this stimulus with a growth movement **called** geotropism. Stems usually grow
(49) Tropisms are caused by hormones **called** auxins. Plants make auxins near the
(50) can be killed by steam. A machine **called** an autoclave is used for this. An
(51) blood for food. They cause a disease **called** scabies. The head louse is a light >

Examples from ITU corpus

(52) therefore be avoided. In order to prevent this, a device **called** a tone disabler is added which inhibits the function of
(53) iplex signals. Alternatively a single piece of equipment **called** a transmultiplexer can be used to perform the functions
(54) Users of the MH service, and Dls, can be identified by a name, **called** a directory name.
(55) It was superseded by a program **called** CATNAP (COMPUTER-AIDED TELEPHONE NETWORK ASSESSMENT PROGRAM),

Examples from Nature corpus

(56) from a variant of erythroid ankyrin **called** ankyrin 2.2 (or protein 2.2). Ankyri
(57) binding site. A part of the intron **called** the internal guide sequence (IGS) is
(58) encodes a very large muscle protein, **called** twitchin, which consists of a protei

These are examples of genus-species relations where the genus (a term or generic class word) appears on the lhs and the species or term being explained appears on the rhs. In order to retrieve this pattern, we can specify that the rhs must consist of a term which matches one of the term formation patterns previously specified, and that the term may not be modified by any modifier other than the definite or indefinite article. The connective phrase must be preceded by an NP which may be modified.

8.3.4 Analysis of the connective phrase known as

This connective phrase is used to introduce one of the following types of information: 1) noun phrase or term on lhs followed by appropriate term on rhs; 2) defining phrase on lhs followed by a term on rhs; 3) superordinate on lhs, term on rhs.

1) noun phrase or term on lhs followed by appropriate term on rhs

Examples from GCSE corpus (170)

(59) Much of the coastline to the south of Flamborough Head consists of a soft sedimentary deposit **known as** boulder clay.
(60) In addition, large vertical shifts of the sea bed can produce enormous waves of water **known as** tsunami which can cause severe flooding in coastal areas.
(61) The acid makes the solution turn back into threads of pure cellulose **known as** rayon.
(62) Heroin, sometimes **known as** 'H', is injected directly into the bloodstream, usually into a vein in the arm. This is **known as** 'mainlining'.
(63) If the monomer is ethene, it will make the polymer poly (ethene)—usually **known as** polythene.

Examples from ITU corpus (108)

(64) a signal of limited duration **known as** a ''measuring signal''
(65) Jitter tolerance (also **known as** jitter accommodation)
(66) are identified by a unique code **known as** a point code (Recommendation Q.704

Examples from Nature corpus (11)

(67) However, because of concerns that deoxyxanthosine might undergo depurination, another complementary base, 3-B-D-ribofuranosyl-(1-methyl-pyrazola[4,3-d]pyrimidine-5, 7(4H,6H)-dione) also **known as** 7-methyl oxoformycin B, and trivially designated here as n, was prepared by the route shown in Fig. 2b.

Some of the examples appear to be synonyms (e.g. *jitter tolerance* and *jitter accommodation* (65)*, heroin* and *H* (62)), while others are examples of substitution (e.g. *enormous waves of water* and *tsunami* (60), *threads of pure cellulose* and *rayon* (61)). In the case of the ITU corpus, we know that a term enclosed in brackets is synonymous with the term which precedes it. When a word or phrase in the ITU corpus is enclosed in inverted commas, this seems to signal that it has terminological status.

2) defining phrase on lhs followed by term on rhs

Examples from GCSE corpus

(68) floodplain areas of Amazonia which are seasonally flooded are known as varzea.
(69) What is left are the indigestible parts of what you eat: mostly the tough stringy parts of vegetables, tomato skins, sweetcorn husks **known as** fibre.

Examples from ITU corpus

(70) A function which provides the user with the means to control system functions via MML inputs and outputs; also **known as** an IT function.
(71) Automatic calling procedures which make use of only the 100-series interchange circuits, are **known as** ''serial'' automatic calling and are defined in Recommendation V.25.
(72) In the case of a particular telegram, the linked office through which the telegram enters the system is **known as** the linked entry office; the linked office through which the telegram leaves the system is **known as** the linked exit system.

While the above examples do not match the conditions specified for formal defining expositives, they actually provide broadly similar information in the form of paraphrases. The terms, which appear on the rhs, are generally explained in terms of their superordinate and at least one distinguishing characteristic on the lhs. There were no examples of this pattern in the Nature corpus.

3) superordinate on lhs, term on rhs

Examples from GCSE corpus

(72) They made new irrigated farmlands in settlements known as kibbutz.
(73) He developed a theory, **known as** the Iron Law of Oligarchy, which states that any large scale organisation requires leadership to be successful and survive and inevitably those who lead cannot be controlled by those who are beneath them.

(74) each individual country used to charge taxes, **known as** tariffs, on goods imported into the country.

Example from ITU corpus

(75) a coefficient controlling its breadth, by a procedure **known as** ''completing the square''.

Examples from Nature corpus

(76) The activity rises again during meiosis II and is stabilized at high levels at metaphase II in fully matured ococytes (or unfertilized eggs) by a calcium-sensitive factor **known as** cytostatic factor (C5F)2,3. MPS, originally found in the unfertilized amphibian egg4, is now known to be ubiquitous in eukaryotes and promotes G2-M transitions in both meiosis and mitosis1,5,6.

(77) Such a plot avoids problems with a mild instability in the finite-element code **known as** ''pressure checkerboarding'' (ref.19) that develops under the application of large viscous strains.

The above contain mainly examples of genus-species relations but there also appears to be at least one example of substitution (e.g. (76) *a calcium-sensitive factor* and *cytostatic factor*). To retrieve this example, one could specify that the words or phrases which precede and follow the connective phrase must belong to the same grammatical class and must correspond to one of the term formation patterns specified in chapter seven.

8.3.5 Analysis of connective phrase the term

Examples from GCSE corpus (19)

(78) **The term** palisade stems from the fact that they have the appearance of the wall or palisade of an old, wooden fort.

(79) **The term** 'concentration camp' derives from the Boer War, 1899-1902, when the British herded Boer families into compounds to prevent aid being given to the Boer guerillas.

(80) **The term** 'means' covers a wide range of behaviour, some of which conforms to what is regarded as normal, some of which is seen as deviant.

(81) ***The term** coastal management is a useful one to describe how we treat and look after the coast.

Examples from ITU corpus (401)

(82) **The term** ''mean value'' is understood as the expected value in the probabilistic sense.
(83) **The term** ''Reversed Charge'' is used to mean collect, credit card and third number paying calls.
(84) **The term** ''delivery of messages'' applies to the forwarding of messages, which were input into an SFU by an originating telex subscriber, to a telex subscriber over the telex network.
(85) **The term** ''notification'' applies to the forwarding of an advice of delivery/non-delivery of a message to the originating telex subscriber over an international telex circuit.
(86) **The term** default implies that the value defined should be used in the absence of any assignment or negotiation of alternative values.
(87) **The term** real time call establishment refers to a set of procedures based on which the communication can be started in a relatively short time (i.e. in the order of a few seconds) after the request is made.

Examples from Nature corpus (1)

(88) Surface uplift (**The term** is used to mean that the average elevation of the ground increases) on a regional scale is difficult to demonstrate.

Whenever this connective phrase appears at the beginning of a sentence, it is almost invariably followed by a term and a defining paraphrase, the single occurrence of this pattern in the Nature corpus and the last example cited from the GCSE corpus being notable exceptions. This pattern is particularly common in the ITU corpus.

8.3.6 Analysis of connective phrase (*)

In the ITU and Nature corpora, this connective phrase can signal the presence of one of the following patterns:1) term before connective phrase followed by synonymous term within connective phrase; 2) term before connective phrase followed by term abbreviation within connective phrase. In the GCSE corpus, it is used to indicate 3) word or term on lhs, word or term in brackets, 4) word or term on lhs, defining phrase in brackets, or vice versa.

1) term before connective phrase followed by alternative term within connective phrase

Examples from ITU corpus (66,999)

(89) be available from analysis of call vouchers (dockets). For derivation of the efficiency
(90) the satellite and the receiving earth station (down-link) of each country receiving the
(91) international telephone service:- distress (emergency) calls;- government calls;
(92) international data communication centres (gateways). #TITRE# 4.3 Interconnection
(93) overflow often have direct first-choice (high-usage) routes, and indirect alternative

Examples from Nature corpus (3,423)

(94) that are currently available on optical disk (CD-ROM), the most convenient form for
(95) agents which include compactin and lovastatin (mevinolin), competitively inhibit the
(96) than the horizontal stretching. The vi-scous (non-lithostatic) stresses developed in the
(97) detect both longwave (infrared) and shortwave (solar) radiation. The ERBE scanners,
(98) swells The idea of great continental uplifts (swells) with rifted crests is an old one

As already noted, the authors of the ITU text specify that round brackets are used to indicated alternative terms or synonyms. Consequently, we can state that when the connective phrase (*) is preceded by a term which matches the term formation patterns previously specified and contains a term which also corresponds to the term formation patterns, these terms are deemed to be synonymous. It appears, from our analysis of the Nature corpus that we can draw the same conclusion about the terms which appear both before and within the brackets sign.

2) term before connective phrase followed by term abbreviation within connective phrase

Examples from ITU corpus

(99) #GR# busy-flash seizure ratio (BFSR) #AS# BFSR gives the relationship
(100) and data circuit terminating equipment (DCE) for terminals operating in the
(101) digital circuit multiplication equipment (DCME) and A/-law converters are not
(102) INIC [or a Data Network Identification Code (DNIC)] to identify the individual

(103) #IT# Fixed daily measurement period (FDMP) #IQ# With this method
(104) initiated when the general directory number (GDN) is called. One line in the group is
(105) b) the aeronautical ground earth station (GES); and c) the mobile aircraft earth
(106) —to pass Higher Layer Compatibility (HLC) Information in order to support
(107) identification of the home location register (HLR) of the mobile station. #TITRE# 2.5
(108) service, the Home Public Land Mobile Network (HPLMN) will know the location of
(109) #IT# High Speed Data (HSD) #IQ##IT# channels #IQ# High-speed 56
(110) or Fixed Daily Measurement Hour (FDMH) and is described in detail in

Examples from Nature corpus

(111) alpain or a calcium-activated neutral protease (CANP) has been shown to be well
(112) from the Global Digital Seismograph Network (GDSN), chromosome condensation
(113) (MPF) which causes germinal vesicle breakdown (GVBD) and chromosome
(114) inhibitor isobutl methyl xanthine (IBMX) (0.5 mM) (Fig. 5). These agents
(115) a photo-detector and a light-emitting diode (LED). An incident signal above threshold
(116) of a major histocompatibility complex (MHC) molecule expressed n the thymus
(117) dominates mid-ocean-ridge basalts (MORBs) worldwide. Depletion of the
(118) analysis by secondary-ion mass spectromety (SIMS) apparently showed a

We can retrieve this pattern by specifying that the letters of the entry within the connective phrase must match the initials of the corresponding number of words which appear before the connective phrase. Thus, if the entry has three letters, the initials of the three words preceding the connective phrase must correspond to the three letters within the brackets.

3) word or term on lhs, word or term in brackets

Examples from GCSE corpus (5,462)

(119) Label the spore cases (sporangia).
(120) These wastes must be removed (or excreted).
(121) The digested food is soaked up (or absorbed).
(122) The plant has a mineral deficiency (or shortage)

Unlike the examples in the ITU and Nature corpora where brackets are used to indi-

cate synonymy, they appear to be used for substitution in the GCSE corpus. The term which is being explained can appear within the brackets (e.g. *sporangia*) or before the brackets (e.g. *mineral deficiency*) and the word or phrase which is being used to explain the term will appear within the brackets or before them. To retrieve this pattern, we could specify that the grammatical class of the word/phrases which appears before and after the connective phrase must match and at least one of them must correspond to a valid term formation pattern.

4) word or term on lhs, defining phrase in brackets, or vice versa

(123) Many important antibiotics (substances which kill harmful bacteria)
(124) the piece of bread was covered with a grey pin mould (Mucor)
(125) Bacteria which cause decay (sacrophytic bacteria)

These are again examples of substitution and, as with the previous examples, the term being explained may appear within or before the brackets. The following example from the GCSE corpus, like some others in this particular corpus, does not fit into any particular pattern.

(126) Buttercup flowers have stamens and carpels (hermaphrodite)

Here, there appears to be an assumption that the reader already knows the term *hermaphrodite* and will therefore understand the function of *stamens* and *carpels*.

8.4 Conclusion

At the beginning of this chapter, we explored the concepts of synonymy, paraphrase and substitution. When we use the term synonym, we are referring to equivalence in meaning and usage. Synonyms are in fact quite rare in the corpora and appear mainly to be provided in brackets. The two other methods of expressing an equivalence relation, i.e. paraphrase and substitution, are much more common. Substitution often involves the use of a general language word or phrase to explain a term. The general language words are not to be perceived as being synonymous. Substitution is particularly useful in situations where the reader may not be familiar with the specialized vocabulary of a particular subject field. This may explain why it appears to be especially common in the GCSE corpus (cf. examples for *called, i.e.*). Having identified a number of connective phrases which seemed to signal that some relation of equivalence was being expressed, we looked at examples of each of these connective phrases. We noted that many of the connective phrases are used to signal

more than one type of equivalence relation. We were therefore unable to specify that if 'x' connective phrase is present, 'x' equivalence relation is being expressed. However, in most cases of substitution and synonymy, and of genus-species relations it was possible to stipulate that the grammatical class of the word/phrase(s) which preceded and followed the connective phrase had to match in order to qualify as valid statements.

The information retrieved using these connective phrases can be used to complement information which has already been retrieved using the specifications described in chapter eight, or as a starting point for the formulation of a definition where no other information is available about a term. Here, we have concentrated on a fairly small set of connective phrases and the work as presented here is more indicative than conclusive. However, we believe that it has enormous potential and that there may be many other linguistic signals (e.g. generic class words) which can be exploited in a similar manner in the design of a corpus-based system for retrieving information about terms.

9 Using the term as the node

9.1 Introduction

In the two previous chapters we used connectives as the node for retrieving metalanguage statements about terms. In this chapter we look at what type of information can be retrieved by using a term as the search node. The method involves producing a concordance file for terms and examining the environment of the search node to see whether it is possible to glean additional information about the meaning and usage of the term under investigation. The approach described in this chapter is, in many respects, analogous to the more usual method of proceeding in corpus-based lexicography where concordances of lexical items are analysed in order to ascertain their meaning and usage. Lexicographers working with general language words usually have access to corpora running into many millions of words and can therefore expect to find a reasonable number of occurrences of all of the words which they are investigating. This is not yet the case for lexicographers dealing with terms. Special purpose corpora are still relatively small and the number of terms which will occur sufficiently frequently to warrant an investigation which uses the term as the node may not be that large. However, as interest grows in corpus-based terminography, one can expect increasingly larger corpora to become available and problems relating to frequency to diminish.

To illustrate what sort of information can be obtained by using the term as the search node, one term has been selected from two of the corpora: *ankyrin* from the Nature corpus, and *respiration* from the GCSE corpus. A concordance file has been produced for each of these terms and the full concordance files are provided in Appendix B. The concordances were examined in order to establish whether they contained additional information about the meanings of these terms which had not previously been retrieved when connectives were being used as the search node. In addition to looking for information about the meaning of the terms themselves, we were also interested in information relating to usage and information about related terms.

It is envisaged that the combined output from a search procedure of this type and from the search procedure described in the two previous chapters would be submitted to subject experts for validation. The information retrieved needs to be collated and prepared for the validation process. The information will be entered on a record sheet

which contains a number of different fields, each designed to record different types of information. A sample record sheet is provided and the rationale for each of the proposed fields is explained. A record sheet is completed for the term *ankyrin 2.2*.

9.2 Evaluating occurrences of Ankyrin*

With the procedure adopted in the two previous chapters, ankyrin 2.2 was selected as a term which co-occurred with one of the connectives used for the retrieval of metalanguage patterns. Initially, *ankyrin 2.2* was used as the search node and it very quickly became apparent that *ankyrin 2.2* was just one of several types of ankyrin and that it might be interesting to continue the investigation further up the conceptual hierarchy. We chose therefore to use *ankyrin** as the search node and used the wild card to ensure that plural occurrences would also be retrieved. Ankyrin* occurs either as a term on its own or as the head word or as some form of modifier of complex nominals. While we were making an assumption that *ankyrin* was likely to be a term, we were aware that not all of the complex nominals which would be retrieved using the node *ankyrin** would prove to be terms. The intervention of subject experts would be required at a later stage to determine which of the complex nominals retrieved actually were terms. Ankyrin* appears as head word of the following complex nominals: brain ankyrin, brain-type ankyrins, erythrocyte ankyrin, human erythrocyte ankyrin, erythrocyte ankyrins, erythroid ankyrins, full-length ankyrin, 'built-in' ankyrins, native ankyrin, ankyrin 2.2, unphosphorylated ankyrin. It appears as some form of modifier in the following complex nominals: ankyrin homologues, ankyrin repeats, ankyrin-like repeats, deduced ankyrin sequence, erythrocyte ankyrin clones, erythrocyte ankyrin gene, erythroid ankyrin repeats, erythroid ankyrin sites, ankyrin variant 2.2, ankyrin defects, ankyrin genomic structure, ankyrin amino-acid sequence, ankyrin sequence, ankyrin cDNA sequence, ankyrin RNAs. Of particular interest are the complex nominals where ankyrin* functions as the head word because these are all types of ankyrin. By taking these complex nominals alone, it is already possible to start building a conceptual hierarchy with *ankyrin* as the superordinate term in the hierarchy. To illustrate how this is possible, a segment of the concordance file is provided below. The following complex nominals occur in this concordance segment: erythroid ankyrin, erythrocyte ankyrin, human erythrocyte ankyrin, ankyrin 22, full-length ankyrin.

(1) oid ankyrin called ankyrin 2.2 (or protein 2.2). **Ankyrin** 2.2 is particularly interesting because it is an 'activated' molecule with the ability to bind to all brain and
(2) ing from a variant of erythroid ankyrin called **ankyrin** 2.2 (or protein 2.2). Ankyrin 2.2 is particularly interesting because it is an 'activated' molecule with

(3) n (protein 2.1) but not a smaller variant called **ankyrin** 2.2 (Fig. 1e). By contrast, an antibody to a C-terminal peptide stains both ankyrins (Fig. 1e). Because the
(4) nents and cytoskeletal elements. Erythrocyte **ankyrin** (protein 2.1) is the best-characterized isoform. lt attaches the spectrin skeleton (membrane skeleton) to
(5) of which have been identified. Erythrocyte **ankyrin** contains two principal chymotrypsin-resistant domains that were originally judged to have relative molecu
(6) Analysis of cDNA for human erythrocyte **ankyrin** indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins Samuel E.
(7) from proteolytic digests of erythrocyte **ankyrin** or its chymotryptic domains. Each of these sequences was found in the translatedankyrin cDNA sequence (Fig.
(8) define the structural domains of erythrocyte **ankyrin**, the first-described member of the family, and identify a cluster of characteristic repeats also present in
(9) using expressed segments of erythrocyte **ankyrin**. The availability of erythrocyte ankyrin clones will also facilitate cloning of other ankyrins, analysis of ankyrin
(10) first determined the structure of erythrocyte **ankyrin**. Complementary DNA cloning We cloned ankyrin from a human reticulocyte cDNA library3l to ensure isola
(11) the sequence of human erythrocyte **ankyrin**. They find alternative sequences at the C terminus that probably result from differential splicing: (1) deletion of bp
(12) and closely resemble the repeats in erythroid **ankyrin** (compare Figs 3b and 4b). The principal differences are: (1) residue 3 is much more polar in the ankyrin-like
(13) a consensus sequence of the erythroid **ankyrin** repeats and found very similar repeats in several invertebrate, yeast and viral proteins (Fig. 4). Multiple searches w
(14) of complementary DNA for human erythroid **ankyrin** indicates that the mature protein contains 1,880 amino acids comprising an N-terminal domain binding
(15) other is missing from a variant of erythroid **ankyrin** called ankyrin 2.2 (or protein 2.2). Ankyrin 2.2 is particularly interesting because it is an 'activated' mole
(16) there are hints that other ankyrins exist9-11. **Ankyrin**s are found in many other, and perhaps all, cells8-17. They have been provisionally subdivided into 'erythro
(17) to a peptide in the insert stains full-length **ankyrin** (protein 2.1) but not a smaller variant called ankyrin 2.2 (Fig. 1e). By contrast, an antibody to a C-terminal pep
(18) and differentiation. Discussion In summary, **ankyrin**s constitute a new family of proteins that seem to function as 'molecular brokers'; that is, they coordinate

From the above, we know that *ankyrins* constitute a family of proteins (18). We can deduce that erythroid ankyrins and erythrocyte ankyrins are types of ankyrin. Ankyrin 2.2 is a variant of erythroid ankyrin (15). One can surmise that human erythrocyte ankyrins (6) and human erythroid ankyrins (14) are respectively types of erythrocyte ankyrins and erythroid ankyrins found in humans. The genus-species relations identified in this manner can be used in definitions for the terms *ankyrin, ankyrin 2.2, erythroid ankyrin and erythrocyte ankyrin*. If we now look at the concordances for each of these four terms, we find that further useful information can be retrieved. A selection of concordances for each of these is examined below.

Ankyrin

(19) GAG codon), Asp for pAnk15 (GAC codon). **Ankyrin** contains three structural domains The deduced ankyrin sequence is divided into three regions correspond
(20) present at positions 834, 1,378 and 1,500. **Ankyrin** is also acylated by palmitic acid, and the rapid turnover of fatty acid indicates a regulatory role. Unfortunate
(21) Complementary DNA cloning We cloned **ankyrin** from a human reticulocyte cDNA library3l to ensure isolation of the erythrocyte isoform. We obtained a
(22) other tissues. Among the proteins containing **ankyrin** or ankyrin-like repeats, only ankyrin is available in amounts suitable for structural analysis and binding
(23) there are hints that other ankyrins exist9-11. **Ankyrin**s are found in many other, and perhaps all, cells8-17. They have been provisionally subdivided into 'erythro
(24) domain that regulates the binding of **ankyrin** to spectrin and the anion-exchange protein3O. The function of ankyrin is also regulated by phosphorylation34-36. In
(25) sites, and to additional sites of its ownl7. **Ankyrin**s therefore facilitate and probably control interactions between integral membrane proteins and cytoskeletal
(26) and differentiation. Discussion In summary, **ankyrin**s constitute a new family of proteins that seem to function as 'molecular brokers'; that is, they coordinate interaction between various integral membrane proteins and cytoskeletal elements.

What do we learn about ankyrins from these eight concordance lines? Ankyrins constitute a new family of proteins (26) . Ankyrin has three structural domains (19). Ankyrins may be found in all cells (23). They can be acylated by palmitic acid (20). They can be cloned from a human reticulocyte cDNA library (21). They function as 'molecular brokers'; they co-ordinate interaction between various integral membrane proteins and cytoskeletal elements; they may also facilitate and control these interactions (26). The function of ankyrin can be regulated by phosphorylation (24). Clearly, a terminological definition would not incorporate all of the above information and a subject expert examining this data might even consider that some of the information was not relevant for a terminological definition. However, the data provides raw material which subject experts can use as they see fit, whether as part of a definition or for providing phraseological information.

Ankyrin 2.2

(27) ankyrin called ankyrin 2.2 (or protein 2.2). **Ankyrin** 2.2 is particularly interesting because it is an 'activated' molecule with the ability to bind to all brain and erythr
(28) from a variant of erythroid ankyrin called **ankyrin** 2.2 (or protein 2.2). Ankyrin 2.2 is particularly interesting because it is an 'activated' molecule with the
(29) (protein 2.1) but not a smaller variant called **ankyrin** 2.2 (Fig. 1e). By contrast, an antibody to a C-terminal peptide stains both ankyrins (Fig. 1e). Because the
(30) for palmitoylation has been identified. **Ankyrin** 2.2 lacks some of the regulatory domain The central third of the C-terminal 55K domain contains a 486-bp se

(31) of protein 2.2. The absence of this insert in **ankyrin** 2.2 'activates' the protein and enhances binding to spectrin and to sites on membranes, especially kidney

What does the data reveal about *ankyrin 2.2*? *Ankyrin 2.2* is a variant of erythroid ankyrin (28). It appears also to be known as protein 2.2 (28). It is an 'activated' molecule (28). In concordance line 27 above, *ankyrin 2.2* is followed by an evaluative statement (*is particularly interesting*) and what appears to be a definition of the term *ankyrin 2.2*. When the concordance line was expanded to retrieve the full sentence, the following was obtained:

(32) Ankyrin 2.2 is particularly interesting because it is an 'activated' molecule with the ability to bind to all brain and erythroid ankyrin sites, and to additional sites

The amount of information which can be retrieved from a total of only 5 concordance lines for *ankyrin 2.2* in the Nature corpus is surprisingly rich.

Erythroid ankyrin

(33) and closely resemble the repeats in erythroid **ankyrin** (compare Figs 3b and 4b). The principal differences are: (1) residue 3 is much more polar in the ankyrin-like
(34) other is missing from a variant of erythroid **ankyrin** called ankyrin 2.2 (or protein 2.2). Ankyrin 2.2 is particularly interesting because it is an 'activated'

There were only two occurrences of *erythroid ankyrin* in the Nature corpus and the concordance lines do not reveal very much about the term itself. The second concordance line in particular reveals more about ankyrin 2.2 than it does about *erythroid ankyrin.*

Erythrocyte ankyrin

(35) and cytoskeletal elements. Erythrocyte **ankyrin** (protein 2.1) is the best-characterized isoform. lt attaches the spectrin skeleton (membrane skeleton) to band 3,
(36) of which have been identified. Erythrocyte **ankyrin** contains two principal chymotrypsin-resistant domains that were originally judged to have relative molecular masses of 90,000 (Mr90K) and 72K on the basis of their mobility on SDS
(37) from proteolytic digests of erythrocyte **ankyrin** or its chymotryptic domains. Each of these sequences was found in the translatedankyrin cDNA sequence (Fig.
(38) define the structural domains of erythrocyte **ankyrin**, the first-described member of the family, and identify a cluster of characteristic repeats also present in
(39) using expressed segments of erythrocyte **ankyrin**. The availability of erythrocyte ankyrin clones will also facilitate cloning of other ankyrins, analysis of ankyrin
(40) first determined the structure of erythrocyte **ankyrin**. Complementary DNA cloning We cloned ankyrin from a human reticulocyte cDNA library3l to ensure isola

The first two concordance lines above (35 and 36) have been expanded because there was an indication that there might be further useful information beyond the existing concordance lines. In fact, none of the concordance lines reveals very much about *erythrocyte ankyrin*. We learn only that *erythrocyte ankyrin* contains two principal chymotrypsin-resistant domains (36), and that the term appears to have a synonym: protein 2.1 (35).

9.3 Evaluating occurrences of respiration in the GCSE corpus

Respiration was selected as a search node because it is clearly a term; it corresponds to one of the term formation patterns previously specified and co-occurs with one of the specified connectives in the GCSE corpus. There are 91 occurrences of *respiration* in the GCSE corpus. The full concordance file for *respiration* is provided in Appendix B. As *respiration* is a deverbal noun from the verb *to respire*, a concordance file was also produced for *respire* and this too is provided in Appendix B.

Respiration occurs either as a single word term or as head word of the complex terms *aerobic respiration, anaerobic respiration* and *cellular respiration*. It also occurs as part of the complex nominals, *respiration rate* and *rate of respiration*. In this section, the concordance lines for *respiration, aerobic respiration* and *anaerobic respiration* will be examined.

Respiration

A selection of the concordance lines for *respiration* is provided below.

(41) minerals needed for plant growth. TOPIC 13 **RESPIRATION** Every living organism needs energy to keep itself alive. Movement, growth, reproduction, feed-
(42) for humans and animals. Photosynthesis and **respiration** are very complex processes involving several steps. photosynthesis is the reverse of respiration. Respi-
(43) release carbon dioxide (in a process called **respiration**), and it is also formed when fuels burn. Carbon dioxide is used up when plants photosyntesize (see sec-
(44) in our bodies by a chemical process called **respiration**. During respiration, foods react with oxygen forming carbon dioxide and water. This is why we breathe out
(45) which happen in all living cells. This is called **respiration**. It is one of the most important chemical reactions that happens in cells. There are two kinds of respira-
(46) set free the energy in its food. This is called **respiration**. Oxygen diffuses into Amoeba from the higher concentration of oxygen in the water (figure 3.4a). An
(47) a chemical process called respiration. During **respiration**, foods react with oxygen forming carbon dioxide and water. This is why we breathe out carbon dioxide
(48) el + oxygen-- carbon dioxide + water During **respiration**, foods containing carbon and hydrogen react with oxygen to form carbon dioxide and water: food + ox
(49) O6 + 602 ——=== 6CO2 + 6H2O + energy **Respiration** happens in both plants

and animals, but photosynthesis can only happen in plants. Respiration and photo-
(50) fossil fuel savings run out. 7 Foods as Fuels **Respiration** Foods are broken down in our bodies by a chemical process called respiration. During respiration, foods
(51) urface of ponds in the heat of the summer. In **respiration**, oxygen is used up and carbon dioxide is produced. Fish and the plankton in the water consume oxygen.
(52) cycle (Fig. 5.4). One of the products of **respiration** is water: Some animals, for example gerbils. are adapted to living in a desert. They may never drink water,
(53) in particular the chemical waste products of **respiration**, such as carbon dioxide and urea. . Have a nutrition system. This means being able to collect, absorb or
(54) steps. photosynthesis is the reverse of **respiration**. Respiration is exothermic. lt uses up food and oxygen and produces carbon dioxide, water and energy. ln con-
(55) that happens in cells. There are two kinds of **respiration**: aerobic and anaerobic. In most cells both can happen at the same time. Aerobic respiration uses oxygen
(56) photosynthesis can only happen in plants. **Respiration** and photosynthesis are important in the carbon cycle (figure 2). This shows how carbon, in carbon diox-
(57) photosynthesis is the reverse of respiration. **Respiration** is exothermic. lt uses up food and oxygen and produces carbon dioxide, water and energy. ln contrast pho-
(58) with them as a single cycle. Besides water, **respiration** has another product, carbon dioxide. In order to get the energy from food, all living things must respire
(59) vapour and why urine is mainly water. **Respiration** is also exothermic and energy is given out. Because of this, foods are sometimes described as 'biological fuels'.
(60) from aerobic respiration. Important words **Respiration** the setting free of the energy in food by chemical reactions in cells. Aerobic respiration sett

What do we learn about *respiration* from the above concordance lines? *Respiration* is a chemical process in which foods are broken down in our bodies (50). *Respiration* involves the setting free of the energy in food by chemical reactions in cells (60). It happens in both plants and animals (49). *Respiration* is exothermic (59). During *respiration*, foods react with oxygen forming carbon dioxide and water (44). In *respiration*, oxygen is used up and carbon dioxide is produced (51). There are two types of *respiration*: aerobic and anaerobic. The information can be collated to produce a definition which could be phrased as follows: r*espiration* is a chemical process which takes place in plant and animal cells whereby foods react with oxygen to form carbon dioxide and water. The terminological entry for *respiration* should also indicate that there are two types of respiration (*aerobic respiration* and *anaerobic respiration*), and these two terms should be defined elsewhere in the same publication. The concordance lines for *aerobic respiration* and *anaerobic respiration* are also worthy of investigation, and a selection of each of these follows.

Aerobic respiration

(61) energy. More energy is set free in aerobic **respiration** than in anaerobic respiration. Too much lactic acid will soon stop muscle cells working: the sprinter

(62) food by chemical reactions in cells. Aerobic **respiration** setting free the energy in food by using oxygen. Anaerobic respiration setting free the energy in food wit
(63) To find out if oxygen is used up in aerobic **respiration** Put a few small animals such as woodlice into a syringe. A piece of sponge should separate the woodlice
(64) anaerobic respiration is different from aerobic **respiration** because in anaerobic respiration:1. the energy in sugar is set free without using oxygen to do it, 2. sugar
(65) food and enzymes. 2 Oxygen for aerobic **respiration** of seed tissue. This releases the energy needed for growth and development. 3 A suitable temperature for
(66) both can happen at the same time. Aerobic **respiration** uses oxygen to set energy free. Anaerobic respiration does not use oxygen to set energy free. AEROBIC
(67) not use oxygen to set energy free. AEROBIC **RESPIRATION** Energy is usually set free from fat or sugar in respiration. 'to respire' means 'to do respiration'
(68) get oxygen quickly enough for aerobic **respiration**, they change to anaerobic respiration. For the first few metres of a sprint your muscles use aerobic respira-
(69) is involved in the process it is called aerobic **respiration**. Most plant and animal cells respire aerobically and the reaction can be represented by this equation: The
(70) efflcient at releasing energy than aerobic **respiration**. Anaerobic respiration in yeast is used commercially in baking and brewing. Yeast breaks down glucose to
(71) less energy is set free than from aerobic **respiration**. Important words Respiration the setting free of the energy in food by chemical reactions in cells. A
(72) metres of a sprint your muscles use aerobic **respiration**. Anaerobic respiration is used for the rest of the race. (See figure 13.5.) Your muscles can work hard with

We already know, from the information which has been retrieved about *respiration*, that *aerobic respiration* is one type of *respiration* which means that the superordinate is already known. What we now need to establish is the distinguishing characteristic which distinguishes *aerobic respiration* from *respiration* and indeed from other types of *respiration* such as *anaerobic respiration*. The concordance lines reveal that in *aerobic respiration*, oxygen is used to set energy free (62, 66, 67).

Anaerobic respiration

(73) red. ANAEROBIC RESPIRATION In anaerobic **respiration** the energy in food is set free without using oxygen to do it. When you walk, your muscles are work
(74) uses oxygen to set energy free. Anaerobic **respiration** does not use oxygen to set energy free. AEROBIC RESPIRATION Energy is usually set free from fat or
(75) energy than aerobic respiration. Anaerobic **respiration** in yeast is used commercially in baking and brewing. Yeast breaks down glucose to obtain its energy and
(76) both flasks must be compared. ANAEROBIC **RESPIRATION** In anaerobic respiration the energy in food is set free without using oxygen to do it. When you
(77) energy in food by using oxygen. Anaerobic **respiration** setting free the energy in food without using oxygen to do it. Fermentation a kind of anaerobic re

(78) to do it. Fermentation a kind of anaerobic **respiration** which makes alcohol. Things to do Bread is quick and easy to make, so bake yourself a loaf (in a kitc
(79) muscles use aerobic respiration. Anaerobic **respiration** is used for the rest of the race. (See figure 13.5.) Your muscles can work hard without oxygen for about 15
(80) from fermentation. This is a kind of anaerobic **respiration** in which alcohol is made instead of lactic acid. Figure 13.7 shows that in fermentation, organisms:
(81) dough lighter and better to eat. Anaerobic **respiration** is different from aerobic respiration because in anaerobic respiration:1. the energy in sugar is set free with-
(82) free in aerobic respiration than in anaerobic **respiration**. Too much lactic acid will soon stop muscle cells working: the sprinter cannot carry on sprinting! At the
(83) there is no oxygen. This is called anaerobic **respiration**. It is less efficient at releasing energy than aerobic respiration. Anaerobic respiration in yeast is used
(84) respiration, they change to anaerobic **respiration**. For the first few metres of a sprint your muscles use aerobic respiration. Anaerobic respiration is used for the
(85) respiration because in anaerobic **respiration**:1. the energy in sugar is set free without using oxygen to do it, 2. sugar is not completely broken down to carbon

The difference between *aerobic respiration* and *anaerobic respiration* is that 1. the energy in sugar is set free without using oxygen to do it (73, 76), 2. sugar is not completely broken down to carbon dioxide and water (85). It is less efficient at releasing energy than *aerobic respiration* (83). *Fermentation* is related to *anaerobic respiration* in that it is a kind of *anaerobic respiration* which makes alcohol.

As *respiration* is a deverbal noun from the verb *respire*, we also produced a concordance of *respire** to check whether we had missed out on any information which was relevant for *respiration, aerobic respiration* and *anaerobic respiration. Respire** occurs thirteen times in the GCSE corpus and the full concordance file is provided in appendix B. We find confirmation of our findings about *aerobic respiration*, namely that 'when organisms respire aerobically they use up oxygen and sugar (or fat), give off carbon dioxide and water. We find some additional information about *anaerobic respiration*, namely that all cells of the body respire and most organisms respire aerobically. When cells *respire* anaerobically they use up sugar, make lactic acid and set free energy.

9.4 Collating the Information

The objective of the work described in this book was to investigate whether and how corpora could be used in terminography. The investigation has shown fairly conclusively that certain types of corpora can indeed be used for this purpose. Terms have been identified. Metalanguage statements about terms have been retrieved either by using a set of connectives or terms as the search node. Meta-

Table 1. Sample record sheet

Language
Term Grammatical category
Gender
Plural
Subject field
Definition
Conceptual information is_a:
is_part_of:
Lexical information synonym: abbreviated form: short_for: also_known_as:
Corpus attested collocations Verbs: Nouns: Adjectives:
Related terms:

language statements include full defining expositives, partial defining expositives, indications of synonyms, genus-species relations, part-whole relations, explanation by analogy using general language words. When the term is used as the search node, information about the use of the term and about related terms can also be retrieved. Once all of the information about a particular term or group of terms has been retrieved, it must then be assessed by subject experts who will decide on the validity and suitability of the information for inclusion in a terminological entry.

The next step in the process, therefore, is to envisage how the information which has been collected can be collated and prepared for validation purposes. The simplest means of doing this is to devise a record sheet which can accommodate all of the information retrieved. The sample record sheet provided here (Table 1) is the type of record sheet which would be suitable for recording information retrieved from the corpus. It has a number of different fields, each of which will be used for

different types of information. Naturally, not all of these fields will always be completed for every entry. In some instances, some of the fields may simply not be appropriate for a particular term. In other instances, the information required for a particular field may simply not have been specified in the corpus, in which case the field will have to be completed by a subject expert.

Here, a brief description is provided of the type of information which might be entered into each of the fields. The *language* field is the field which indicates the language of the terminological entry; it would not normally be necessary in a monolingual context but it would be particularly useful in a bi- or multilingual context. The *term* field specifies the term as it appears in the corpus. Here, it is important to enter the term in the form in which it most commonly appears. If, for example, a noun term appears only in plural in the corpus, it should be entered in plural form in this field. It is important to specify the *grammatical category* of a term as this may not be immediately obvious if a term is examined out of context. The *gender* field is only necessary for languages which make gender distinctions. The *plural* field is used to indicate whether a term also occurs in the plural. It is particularly useful for terms which have irregular plurals such as *hypha* and *fungus* in the GCSE corpus (*hyphae* and *fungi* respectively). When a term does not appear in plural in the corpus (e.g. *noise* in the ITU corpus), this should also be indicated in the plural field. When the subject field has been indicated in a metalanguage statement in the corpus, it should be entered under *subject field*. The *definition* field will require the most post-processing by subject experts. Initially, it will contain all of the concordance lines which contain some information about the meaning of a particular term. These may be full or partial defining expositives The concordance lines will then be examined by subject experts and condensed to form a terminological definition. The *is_a* field is used to indicate genus-species relations when these have been provided in the corpus. This field can be used subsequently for retrieving all terms with the same superordinate term. If the superordinate is a generic term such as *process, method* or *device*, this may not be particularly useful but if the superordinate is a term such as *ankyrin*, it will make it possible to group together all terminological entries which have *ankyrin* as their superordinate. The *is_part_of* field is used to indicate part-whole relations when these have been indicated in the corpus. Thus, in the following example from the ITU corpus,

(86) The maritime satellite message transmission system comprises a maritime local circuit, a maritime satellite circuit, a maritime terrestrial circuit and a maritime store-and-forward unit

all of the entries for the terms which follow *comprises* would specify *maritime satellite message transmission system* in the *is_part_of* field. The *synonym* field is to

Table 2. Record sheet for *ankyrin 2.2*

Language	English
Term	ankyrin 2.2
Grammatical category	noun
Gender	
Plural	? not attested in corpus
Subject field	
Definition: a variant of erythroid ankyrin, an 'activated' molecule with the ability to bind to all brain and erythroid sites, and to additional sites of its own.	
Conceptual information	
is_a:	ankyrin
is_part_of:	family of proteins
Lexical information	
synonym:	protein 2.2?
abbreviated form:	
short_for:	
also_known_as:	protein 2.2?
Corpus attested collocations within 3 words to the left and right of the node	
Verbs:	activates, insert, lacks, identified
Nouns:	protein, ankyrin, variant
Adjectives:	erythroid
Related terms:	ankyrin, erythroid ankyrin, erythrocyte ankyrin

be used only when a synonym is clearly indicated in the corpus. The *abbreviated form* and *short for* fields are to be used for indicating either the abbreviated or full form of a term depending on the form in which it has been entered in the *term* field. For example, if *frequency division multiplexing* (ITU corpus) appears in the *term* field, *FDM* should appear in the *abbreviated form* field. The *also_known_as* field is to be used for equivalents which are not synonyms. These may be deprecated terms. The fields for corpus-attested collocations allow for the inclusion of verbs, nouns, adjectives, adverbs and prepositions which typically co-occur with the term in the corpus. The *related terms* field can be used to indicate terms which, on the basis of corpus evidence, appear to be related to the term being defined. In the case of an entry for *ankyrin*, for example, terms such as *erythroid ankyrin, erythrocyte ankyrin* would appear in the *related terms* field.

Table 2 consists of a record sheet for *ankyrin 2.2* which has been completed on the basis of evidence from a total of only five concordance lines for this term in the Nature corpus.

9.5 Conclusion

This chapter investigated the possibility of using the term as the search node for the retrieval of information about a term's meaning and usage. Concordance files were produced for two terms, *ankyrin* and *respiration*. In the case of the concordance file for *ankyrin*, the file not only yielded information about the term *ankyrin* but also yielded information about types of *ankyrin* such as *erythroid ankyrin, erythrocyte ankyrin* and *ankyrin 2.2*. The concordances for each of these were examined separately. In some instances (e.g. *ankyrin 2.2*), it was surprising to note how much information could be gleaned from only a very small number of concordance lines.

A record sheet was devised for recording the different types of information retrieved. Once record sheets have been completed on the basis of corpus evidence, they can be submitted to subject experts for validation and post-processing.

10 Summary

10.1 Summary of findings

The idea for this investigation arose out of a desire to develop a methodology for retrieving information about terms from corpora. It represents one of the first real attempts to integrate terminology studies and corpus linguistics. In the late eighties, the EU financed a number of projects in each of these areas to promote linguistic research and engineering. It recognized the need to develop systems which would be able to process large volumes of natural language text (semi-) automatically. One of the projects, namely ET10/66 Terminology and Extralinguistic Knowledge, in which this author was involved, explored the possibility of using specialized texts as a resource for building terminological knowledge structures. It was envisaged that this resource would be used to enrich machine lexica and facilitate disambiguation in natural language processing. The scope of the project meant that it was only implemented on a very small scale, i.e. on a single text, but the underlying principles appeared to be sufficiently sound to warrant implementation on a larger scale; hence the motivation for this investigation.

It was therefore envisaged that a study of a specialized corpus would be undertaken in order to identify the terms used in the corpus and to examine whether and how terms related to others. It was anticipated that the way in which terms were used in text would reveal something about the terms themselves, their meaning and their usage. It was hoped that the results of the study would make it possible to build knowledge structures similar to those advocated by theoretical terminologists. We were approaching the project from a terminological standpoint with little experience in corpus work but a firm conviction that corpora could serve as a useful resource for terminology studies.

The first question which had to be addressed was how one might design a system which would be able to discriminate automatically between words and terms in corpora. We thought that some of the principles underlying the traditional approach to terminology might provide the answer. Unfortunately, this proved not to be the case. The traditional approach to terminology tends to separate language into two categories, language for general purposes (LGP) where the lexis consists primarily of general language words, and language for special purposes (LSP) where the lexis

consists of terms. Within LSP, terms have protected status; their meaning remains fixed. What we found, however, was that there are no adequate criteria for identifying LSP texts. The criteria proposed by traditional terminologists, and sublanguage researchers too, hold for very narrowly defined text types. However, they do not account for a very large body of language where terms are also used alongside general language words and still retain their protected status. An attempt was made, therefore, to devise a text classification system which would cater for this grey area which does not fit easily into the LSP or LGP category. This led us to consider the question of communicative setting. We identified a number of broad communicative settings in which terms were likely to be used as terms and others where what appeared to be terms were not in fact functioning as such because their protected status was no longer assured. Broadly speaking, authorship, readership and text function define the communicative context. We identified three communicative settings which were likely to use terms. The first of these concerns expert-expert communication where a high density of terms is to be expected. The second concerns communication between experts and people working in the same field, but with a lower level of expertise. Here, we are likely to encounter a fairly high density of terms but also a large number of general language words. The third communicative context concerns communication between experts and the uninitiated (e.g. teacher-pupil communication) where the ratio of general language words to terms is very much higher than in the previous two contexts but where terms still retain their protected status. This approach allowed us to consider the inclusion of some text types, e.g. school textbooks, which would not normally be considered as suitable material for the identification of terms.

Having identified three communicative settings, we recognized that not all texts produced within these settings would be suitable for our purposes. This led us to consider text selection or corpus design criteria. We found that little previous research had been carried out on the design of special purpose corpora, largely because those who have used special purpose corpora in the past have generally had to use whatever material was available. To our knowledge, therefore, the design criteria which we have devised represent the first attempt to draw up general design criteria for the construction of special purpose corpora. These criteria enabled us to select only those texts which were suitable for the investigation which we were undertaking.

Using the corpus design criteria, we selected three corpora (ITU, GCSE and Nature corpora) for our analysis. Each of these corpora corresponds to one of the three prescribed communicative settings and they all meet the design criteria for special purpose corpora. The first step in our research involved producing specifications for a TermHunter, a system which would identify and retrieve terms from the corpora. An initial analysis of the corpora showed that all three contained general

language words as well as terms. We therefore had to find some means of distinguishing between words and terms in the retrieval process. We used an approach which was carried out in two stages. In the first instance, a set of term formation patterns was produced for each of the corpora. These sets were then input into a pattern match program which was used to retrieve all term candidates. We then took the process one step further by specifying certain co-occurrence restrictions in order to refine the output from the first stage. To be eligible for consideration as a term, a term candidate had to meet what we called the generic reference criterion at least once in the corpus under investigation. It also had to co-occur at least once with one of a specified set of linguistic signals. Term candidates which did not meet these criteria at least once in the corpus under investigation were immediately eliminated from consideration. The refinements proposed greatly improved the term retrieval process. Given that the notion of generic reference is a key concept in traditional terminology, it is surprising that it has not been used previously in term identification systems. We would suggest that it is a refinement which could usefully be added to existing and/or future term identification systems. The TermHunter which we have described here could also be used for purposes other than term retrieval, and these are discussed in Section 10.2.2 below.

Our original intention had been to take the sets of terms retrieved from each of the corpora as our starting point for the next stage of our analysis and to investigate whether and how terms related to each other by examining co-occurrence patterns. In other words, we were hoping to implement the ET10/66 approach on a much larger scale. While attempting to devise a term-centred approach, we discovered some interesting facts about the use of language in all three corpora. We were surprised to observe a high frequency of patterns which were being used to provide information about the terms in the corpora. These were metalanguage patterns used by authors to explain or define terms. Instead of using language as the basis for our study, we found ourselves studying metalanguage. With hindsight, it should perhaps have been obvious from the outset, given the function of the corpora, that they would contain both language and metalanguage statements. However, LSP studies have previously focused mainly on the special lexis, and grammar, of LSP rather than on metalanguage patterns in LSP, Harris and Flowerdew being two notable exceptions.

We decided to concentrate on metalanguage patterns and to devise a methodology for retrieving these patterns from the corpora. We began to draw up a list of these patterns and found that, with relatively little computational effort, many of them could be used as input for the formulation of a terminological definition and some were actually functioning as complete and adequate definitions. We identified patterns where terms were simply explained by means of simple substitution, whereby an analogous general language word which was likely to be known to the

reader was provided (e.g. a tiny hole called the micropyle, example from the GCSE corpus). These were particularly prevalent in the GCSE corpus. The same patterns were also used to indicate synonymous terms, particularly in the ITU and Nature corpora. When we started to investigate further, we found patterns which indicated that terms were explained by means of paraphrasing and substitution. Phrases or clauses which were in some way equivalent to the term which had been introduced were inserted to explain the term. On analysing the patterns further, we noted that there were many instances where the metalanguage patterns corresponded to what Trimble calls formal and semi-formal definitions. We found that, if certain selection restrictions were specified, it would even be possible to retrieve statements which were functioning as semi-formal and formal defining expositives in the corpora.

In the corpora, metalanguage statements fall into two categories, specific and generic. Specific metalanguage statements are statements which are qualified in some way by the author. The author may wish to restrict the scope of the statement to the particular text segment in which the statement appears and will use hedges such as *in this context, here* to stipulate that the scope of the metalanguage statement is restricted. Generic metalanguage statements, on the other hand, are, as the name suggests, statements which have general applicability. We devised a methodology for retrieving metalanguage statements which could be classified as generic if they met certain felicity conditions. The methodology involved using a combination of the TermHunter criteria and criteria relating to the structure of the metalanguage patterns. To have used the criteria for the structure of the metalanguage patterns alone would not have been sufficient because many of these patterns only function as metalanguage statements iff they co-occur with terms. When they co-occur with general language words, they are simply functioning as ordinary language.

In addition to the retrieval of simple substitution relations and genus-species relations, our methodology enabled us to locate instances of complete and partial definition statements in the corpora. We found that metalanguage statements were much more common in the ITU and GCSE corpora than in the Nature corpus. This is not surprising as there is an assumption in the Nature texts that the author and reader share a similar level of expertise.

We then investigated the possibility of using a term-centred approach for the retrieval of information about the meaning and usage of terms. This approach is analogous to that used by Cobuild for the compilation of general language dictionaries. We examined the concordances of a selection of terms and found that even in instances where a term occurs fairly infrequently in a corpus (i.e. 5 times or less), this approach is quite productive. A record sheet was devised for recording the information retrieved. It is envisaged that record sheets, once completed, would be submitted to subject experts for validation and post-editing.

Both approaches presented here for the retrieval of information about terms demonstrate that it is indeed possible to retrieve from corpora at least some of the information which would previously have been gathered through consultation with subject experts.

10.2 Implications for future research

We believe that the investigation described here will have an impact on four areas: term retrieval, corpus design and text evaluation, terminography, teaching of LSP.

10.2.1 Term retrieval

As noted, previous term identification systems have tended to use a combination of syntactic patterns and frequency to identify and retrieve term candidates. We would argue that term retrieval techniques which rely only on morpho--syntactic and statistical criteria result in output which requires greater post-processing than those which use an add-on module such as the one proposed here. The proposed TermHunter system allows for considerable refinement of the initial output, leading to the elimination of many non-terms which would have been retrieved by the methods used previously. The refinements are relatively simple to implement, and we believe that the criterion of generic reference, in particular, will prove to be a powerful device. It would be necessary to adapt the CLG tagger, and perhaps other taggers too, to ensure that it could discriminate between different types of determiner. There are of course other ways of retrieving only term candidates preceded by the indefinite article or no article at all but modification of the tagset seems to be the most straightforward. It would also be useful to devise a mechanism, perhaps a stoplist, for preventing some common adjectives such as *many, different, others, some* from combining with nouns to form term candidates in an adj + noun pattern. Again, this should be relatively simple to implement.

10.2.2 Text evaluation and corpus design

TermHunter can be used not only as a means of identifying terms but also as a means of evaluating whether or not a particular text or collection of texts is suitable for inclusion in a corpus about a particular subject domain. In the construction of, for example, a corpus of engineering texts, it might be used to evaluate whether or not a particular text or collections of texts contained a sufficiently high density of engineering terms to warrant inclusion in such a corpus. It could thus be used as an

internal criterion for assessing whether or not a text was a 'good' source for engineering terms.

Similarly, the set of metalanguage patterns could be used to assess whether a text or texts were 'good' sources of information about the meaning of terms. We noted, for example, that the GCSE and ITU corpora were good sources for information about terms whereas the Nature corpus would perhaps be more suitable as a source for terms.

Both TermHunter and the metalanguage patterns could therefore be used as internal evaluation criteria in the construction of special purpose corpora for use in terminography.

The study of specialized corpora is still very much in its infancy but we believe that this is an area which will expand and receive a great deal more attention when specialized texts become more readily available to researchers. We have endeavoured to devise a set of text collection criteria which can be used as a basis for the compilation of specialized corpora. We have shown that many of the criteria devised for the collection of general language texts are also valid for the collection of specialized texts but that there are others which are less important and perhaps even irrelevant. The external design criteria proposed for this investigation can therefore be used in the compilation of special purpose corpora.

10.2.3 Terminography

This is the area in which our findings could have the greatest impact. In the past, when terminologists have used texts for terminography, it has been mainly for the purpose of identifying the terms of a domain and also for retrieving contextual fragments. What has been proposed here both in the specifications for retrieving metalanguage patterns and in the term-centred approach is a means of using a semi-automatic method for retrieving information about terms. The fact that it is possible to identify genus-species relations and synonymous terms means that this information can be used to enrich machine and other lexica without great human effort. The genus-species and synonymous relations can be extracted and used for the purpose of disambiguation in natural language processing systems. This is the first step towards building the type of knowledge structure which is necessary in order to establish how terms relate to each other. The fact that it is also possible to retrieve formal defining expositives means that it is possible not only to identify genus-species relations but also to ascertain what distinguishes a term from its superordinate. Further investigation will be required to establish whether it is possible to predict what type of distinguishing characteristic will be specified, e.g. whether a particular class of term is more likely to be explained in terms of its purpose or its properties, and how one might distinguish between different characteristics.

In the compilation of specialized glossaries and dictionaries and the construction of termbanks, terminographers could use the method proposed here as a starting point in the terminography process. Metalanguage patterns could be retrieved and concordances produced and presented to subject experts for validation, thereby considerably reducing the human effort previously required for such work. We would suggest that there may be no need to transform some of the metalanguage statements into conventional dictionary type entries; they should rather retain their general original form so that they look more like Cobuild definitions.

While this investigation has focused only on metalanguage patterns in English, such patterns must also exist in other languages. As the ITU corpus is also available in French and Spanish, the specifications provided here could serve as a basis for identifying equivalent metalanguage patterns in these languages. The same would apply to comparable multilingual corpora compiled using the special purpose design criteria.

10.2.4 Using the approach for teaching LSP

While the approach described for the retrieval of information about terms is designed for use on corpora, it can also be adapted for use in the LSP classroom. In fact, this approach has already been successfully implemented in an environment involving the teaching of specialized translation (Pearson 1996). Students who need to be able to understand specialized texts in languages other than their native language can be shown how to screen texts for clues about the meaning of terms. They can search for metalanguage patterns which will help them to understand the terminology being used. By examining all occurrences of a particular term in a text, they can also discover interesting facts about the way in which the term is used. This is particularly useful in the case of recently coined terms which may not yet be listed in any dictionary or when students simply do not have access to appropriate specialized dictionaries.

References

Aarts, J. and W. Meijs (eds). 1991. *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam: Rodopi B.V.

Ahmad, K. 1994. ''Pragmatics of Specialist Terms and Terminology Management.'' In P. Steffens (ed.). *Proceedings of the European Association of Machine Translation Symposium*. Heidelberg: Springer-Verlag, 51–76.

Ahmad, K., A. Davies, H. Fulford and M. Rogers. 1994. ''The elaboration of special language terms; the role of contextual examples, representative samples and normative requirements.'' *Euralex '92 Proceedings I*. Tampere: Studia Translatologica, 139–150.

Ahmad, K., A. Davies, H. Fulford and M. Rogers. 1994. ''What is a term? The semi-automatic extraction of terms from text.'' In M. Snell-Hornby, F. Pöchhacker and K. Kaindl (eds), *Translation Studies: an Interdiscipline*. Amsterdam: John Benjamins Publishing Company, 267–277.

Aijmer, K. and B. Altenberg. (eds) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Allen, J.P.B. and H.G. Widdowson. 1974. ''Teaching the Communicative Use of English.'' *International Review of Applied Linguistics*, (12)1: 1–20.

Anthony, E. 1976. ''English for Special Purposes – A Lexical Context.'' In J.C. Richards (ed.), *Teaching English for Science and Technology*. Singapore: Singapore University Press.

Apresyan, Y, I.A. Mel'čuk and A.K. Zolkowsky. 1970. ''Semantics and Lexicography: Towards a new type of Unilingual Dictionary.'' In F. Kiefer (ed.), *Studies in Syntax and Semantics*. Dordrecht: Reidel, 1–33.

Arntz, R. 1988. ''Steps towards a translation-oriented typology of technical texts.'' *Meta* (23)4: 468–471.

(–) 1977. *Astronomy: A Dictionary of Space and the Universe*. London: Arrow.

Atkins, S., J. Clear and N. Ostler. 1992. ''Corpus Design Criteria.'' *Literary and Linguistic Computing*, (7)1. Oxford: Oxford University Press, 1–16.

Austin, J.L. 1962. *How to do things with words*. Oxford: Oxford University Press.

Bachrach, A.J. 1973. ''The Relative Value of Definitions in Terminological References.'' *Meta* (18)1: 161–170.

Baker, M. 1993. ''Corpus Linguistics and Translation Studies.'' In M. Baker et al. (eds), 233–250.

Baker, M. 1995 ''Corpora in Translation Studies: An Overview and some Suggestions for future Research.'' *Target* (7)2: 223–243.

Baker, M., G. Francis and E. Tognini-Bonelli (eds). 1993. *Text and Technology*. Amsterdam: John Benjamins Publishing Company.

Barlow, M. 1996. ''Parallel Texts in Language Teaching.'' In S. Botley et al. (eds), 45–56.

Barnbrook, G., 1993. ''The Automatic Analysis of Definitions.'' In M. Baker et al. (eds), 312–331.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Barnbrook, G. and J. Sinclair. 1995. ''Parsing Cobuild Entries.'' In J. Sinclair, M. Hoelter, C. Peters (eds), *The Languages of Definition: The Formalization of Dictionary Definitions*

for Natural Language Processing. Luxembourg: Office for Official Publication of the European Communities, 13–58.

Beaugrande, R. de. 1985. "Text Linguistics in Discourse Studies." In T.A. Van Dijk (ed.), *Handbook of Discourse Analysis, Vol. I*. London: Academic Press, 41–70.

Beaugrande, R. de. and Dressler, W. 1983. *Introduction to Text Linguistics*. London and New York: Longman.

Beedham, C. and M. Bloor. 1989. "English for Computer Science and the Formal Realization of Communicative Functions." *Fachsprache* (11)1–2: 13–24.

Bhatia, V. 1993. *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1989. "A Typology of English Texts." *Linguistics* 27: 3–43.

Biber, D. 1993. "Representativeness in corpus design." *Literary and Linguistic Computing*, 8(4). Oxford: Oxford University Press, 243–257.

Bloomfield, L. 1939. "Linguistic Aspects of Science." *International Encyclopedia of Unified Science* (1)4, 1–57.

Bolinger, D. 1965. "The Atomization of Meaning." *Language* (41)4: 555–573.

Botley, S., J. Glass, T. McEnery and A. Wilson (eds). 1996. *Proceedings of Teaching and Language Corpora 1996*. Lancaster: University Centre for Computer Corpus Research on Language Technical Papers 9 (Special Issue).

Bourigault, D., I. Gonzalez-Mullier and C. Gros. 1996. "LEXTER, a Natural Language Processing Tool for Terminology Extraction." *Euralex '96 Proceedings II*, Göteborg: Göteborg University, 771–780.

Boutin-Quesnel, R., N. Bélanger, N. Kerpan and L.J. Rousseau. 1985. *Vocabulaire systématique de la terminologie*. Quebec: Office de la Langue Française.

Braasch, A. 1992. "Text based dictionary work for a domain-specific language." In F. Kiefer et al. (eds), 71–79.

Bowker, L. 1995. *A Multidimensional Approach to Classification in Terminology: Working within a Computational Framework*. PhD Thesis. University of Manchester, United Kingdom.

Brown, G. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Butler, C. 1985. *Statistics in Linguistics*. London: Basil Blackwell

Calzolari, N. 1988. "The Dictionary and the Thesaurus can be combined." In M. Walton Evens (ed.), *Studies in NLP: Relational Models of the Lexicon. Representing Knowledge in Semantic Networks*. Cambridge: Cambridge University Press, 75–95.

Carter, R. 1987. *Vocabulary*. London and New York: Routledge.

Channell, J. 1993. *Vague Language*. Oxford: Oxford University Press.

Chargaff, E. 1986. "How Scientific Papers are Written." *Fachsprache* (8)3–4: 106–109.

Cheong, L.K. 1978. *The Syntax of Scientific English*. Singapore: Singapore University Press.

Clear, J. 1987. "Computing: Overview of the Role of Computing in Cobuild." In J.M. Sinclair (ed.), 41–61.

Clear, J. 1992. ''Corpus Sampling.'' In G. Leitner (ed.), *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter. 21–31.
Clear, J. 1993. ''From Firth Principles: Computational Tools for the Study of Collocation.'' In M. Baker et al. (eds.), 271–292.
(–) *Collins Cobuild Bridge Bilingual Portugese*. 1995. London: HarperCollins.
(–) *Collins Cobuild English Language Dictionary*. 1987. London and Glasgow: Collins.
(–) *Collins Cobuild English Dictionary*. 1995. London: HarperCollins
(–) *Collins English Dictionary*. 1991. Glasgow: HarperCollins.
(–) *Collins Cobuild English Grammar*. 1991. London and Glasgow: Collins. 19–21.
(–) *Compact Edition of the Oxford English Dictionary*. 1971. (Complete Text of OED reproduced micrographically) Oxford: Oxford University Press.
Cook, J. 1989. *Discourse*. Oxford: Oxford University Press.
Crofts, J.N. 1981. ''Subjects and Objects in ESP teaching materials.'' In L. Selinker et al. (eds), 23–39.
Cruse, D.A. 1986. *Lexical Semantics*. New York: Cambridge University Press.
Daille, B. 1994. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. PhD Thesis. Université Paris VII.
Darien, S. 1981. ''The role of definitions in scientific and technical writing: forms, functions and properties.'' *English Language Research Journal* 2: 41–56.
Davidson, J. and M. Sitarski. 1983. ''Preparing Terminology Standards: The Importance of Providing Essential Linkages and Usage Information.'' In C.G. Interrante and F.J. Heymann (eds), 104–110.
Duquet-Picard, D. ''What happens to terms that have failed Standardization? They live on and proliferate in dictionaries.'' In C.G. Interrante and F.J. Heymann (eds), 15–25.
Elnitsky, L. 1984. ''Présentation d'un article de dictionnaire (lexème) et d'un superarticle (vocable).'' In I. Mel'čuk (ed.), *Dictionnaire explicatif et combinatoire du français contemporain*. Montreal: Presses de l'Université de Montréal, 17–25.
Fang, C.Y., 1991. ''Building a corpus of the English of computer science.'' In J. Aarts, P. de Haan and N. Oostdijk (eds), *English Language Corpora*. Amsterdam: Rodopi, 73–78.
Felber, H. 1984. ''Basic Principles and Methods for the Preparation of Terminology Standards'', in C.G. Interrante and F.J.Heymann (eds).
Fillmore, C.J. 1993. '' 'Corpus Linguistics' or 'Computer-aided armchair linguistics'.'' In J. Svartvik (ed.), 35–60.
Firth, J.R. 1958a. ''The Technique of Semantics.'' *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press, 7–33.
Firth, J.R. 1958b. ''The Statistics of Linguistic Science.'' *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press, 139–147.
Firth, J.R. 1958. ''Modes of Meaning.'' *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press, 192–215.
Flowerdew, J. 1991. ''Pragmatic modifications on the 'representative' speech act of defining.'' *Journal of Pragmatics* (15)3: 253–264.
Flowerdew, J. 1992a. ''Definitions in Science Lectures.'' *Applied Linguistics* (13)2: 202–221.

Flowerdew, J. 1992b. ''Salience in the Performance of One Speech Act: The Case of Definitions.'' *Discourse Processes* 15: 165–181.
Flowerdew, J. 1993. ''Concordancing as a tool in Course Design.'' *System* (21)2. Exeter: Pergamon Press, 231–244.
Fodor, J.A. and J.J. Katz (eds) 1994. *The Structure of Language: Readings in the Philosophy of Science*. Englewood Cliffs, New Jersey: Prentice-Hall.
Francis, G., 1994. ''A corpus-driven approach to grammar.'' In M. Baker et al. (eds.), 137–156.
Francis, W.N. 1992. ''Language Corpora B.C.'' In J. Svartvik (ed.), 17–32.
Frawley, W. 1980/1981. ''Lexicography and the Philosophy of Science.'' *Dictionaries*, 3: 18–27.
Frege, G., 1984. *Collected Papers on Mathematics, Logic and Philosophy*. Oxford: Blackwell.
Gläser, R. 1982. ''The Problem of Style Classification in LSP (ESP).'' *Proceedings of the 3rd European Symposium on LSP*. Paris: Unesco, 69–81.
Gläser, R. 1992. ''A Multi-level Model for a Typology of LSP Genres.'' *Fachsprache* (14) 1–2: 18–25.
Godman, A. and E.M.F. Payne. 1979. *Longman Dictionary of Scientific Usage*. London: Longman.
Godman, A. and E.M.F. Payne. 1981. ''A taxonomic approach to the lexis of science.'' In L. Selinker et al. (eds), 23–39.
Gonsalves, R.J., 1984. *The Validity of Definitions*. PhD Thesis. City University of New York.
Gopnik, M. 1972. *Linguistic Structures in Scientific Text*. The Hague: Mouton.
Grishman, R., L. Hirschman and N. Thank Nhan. 1986. ''Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments.'' *Computational Linguistics* (12)3: 205–214.
Grishman, R. and R. Kittredge (eds). 1986. *Analyzing Language in Restricted Domains*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Haan, P. de. 1992. ''The optimum corpus sample size?'' In G. Leitner (ed.), *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter. 3–19.
Hanks, P. 1987. ''Definitions and Explanations.'' In J.M. Sinclair (ed.), 116–136
Halliday, M.A.K. 1966. ''Lexis as a linguistic level.'' In C.E. Bazell, J.C. Catford, M.A.K. Halliday (eds), *In Memory of J.R. Firth*. London: Longmans Linguistic Library, 148–162.
Halliday, M.A.K. and Hasan, R. 1989. *Language, Context and texts: aspects of language in a social-semiotic perspective*. Oxford, New York: Oxford University Press.
Halliday, M.A.K. and Martin, J.R. 1993. *Writing Science*. Pittsburgh: University of Pittsburgh Press.
Harris, Z. 1968. *Mathematical Structures of Languages*. New York: John Wiley & Sons.
Harris, Z. 1976. ''On a Theory of Language.'' *The Journal of Philosophy* (73)10, 253–277.
Harris, Z. 1988. *Language and Information*. New York: Columbia University Press.
Hartenstein, K. 1986. ''La Notion de 'fonction lexicale' dans la conception lexicologique d'Igor Mel'čuk.'' *Actes du XVIIIe Congrès International de linguistique et de Philologie Romanes*, Université de Trèves. Tübingen: Max Niemeyer Verlag, 153–163.

Hartmann, R.R.K. 1980. *Contrastive Textology: Comparative Discourse Analysis in Applied Linguistics*. Heidelberg: Julius Groos Verlag.

Hartmann, R.R.K. 1983. *Lexicography: Principles and Practice*. London, New York: Academic Press.

Hartmann, R.R.K. (ed.) 1984. *Lexeter '83 Proceedings, Papers from the International Conference on Lexicography at Exeter, 9–12 September 1983*. Tübingen: Max Niemeyer Verlag.

Heid, U. and J. McNaught,(eds). 1991. *Eurotra-7 Study: Feasibility and Project Definition Study on the Reusability of lexical and terminological resources in computerized applications*. Final Report. CEC, Luxembourg.

Heid, U. and G. Freibott. 1991. ''Collocations dans une base de données terminologique et lexicale.'' *Meta* (36)1: 77–91.

Heid, U., M. Heyn and O. Christ. 1992. ''Extracting Linguistic Information from machine-readable versions of traditional dictionaries - a metalexicographic method and some tools.'' In F. Kiefer, G. Kiss, and J. Pajzs (eds), 162–174.

Herbert, A.J., 1965. *The Structure of Technical English*. London: Longman.

Hirs, W. 1993. ''The Use of Terminological Principles and Methods in Medicine.'' In H.B. Sonneveld and K.L. Loening (eds), *Terminology*. Amsterdam: John Benjamins Publishing Company, 223–240.

Hirschman, L. 1986. ''Discovering Sublanguage Structures.'' In R. Grishman and R. Kittredge (eds), 211–234.

Hirschman, L. and Sager, N. 1982. ''Automatic Information formatting of a Medical Sublanguage.'' In R. Kittredge and J. Lehrberger (eds), 27–80.

Hobbs, J.R., 1986. ''Sublanguage and Knowledge.'' In R. Grishman and R. Kittredge (eds), 53–68.

Hockey, S. and D. Walker. 1993. ''Developing Effective Resources for Research in Texts: Collecting Texts, Tagging Texts, Cataloguing Texts, Using Texts and Putting Texts in Context'' *Literary and Linguistic Computing*, 8(4). Oxford: Oxford University Press, 235–242.

Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoffmann, L. 1985. *Kommunikationsmittel Fachsprache*. Tübingen: Gunter Narr Verlag.

Horsella, M. and F. Pérez. 1991. ''Nominal Compounds in Chemical English Literature: Towards an approach to Text Typology.'' *English for Specific Purposes* 10: 125–138.

Hunston, S., 1993. ''Professional Conflict: Disagreement in Academic Discourse.'' In M. Baker et al. (eds), 115–133.

Hutchins, W.J., 1977. ''On the Problem of 'Aboutness' in Document Analysis.'' *Journal of Informatics* (1)1: 17–35.

Ilson, R. 1984. ''The Communicative Significance of some lexicographic Conventions.'' In R.R.K. Hartmann (ed.) 80–86.

Ilson, R. 1986a. ''General English Dictionaries for Foreign Learners: Explanatory Techniques in Dictionaries.'' *Lexicographica*, 2, 214–222.

Ilson, R. 1986b. ''Towards a Taxonomy of Dictionary Definitions.'' *Polyglot* 7:F1–F14.

(–) 1968. *ISO R704 Naming Principles*.

(–) 1968. *ISO R 860 International Unification of Concepts and terms*.
(–) 1969. *ISO R 919 Guide for the Preparation of Classified Vocabularies*.
(–) 1990. *ISO 1087 Terminology - Vocabulary*.
(–) 1969. *ISO R 1149 Layout of multilingual classified vocabularies*.
(–) 1973. *ISO 1951 Lexicographical Symbols particularly for use in classified defining vocabularies*.
(–) 1974. *ISO 2788 Documentation - Guidelines for the establishment and development of monolingual thesauri*.
(–) 1985. *ISO 5964 Documentation and Guidelines for the establishment and development of multilingual thesaur*
(–) 1992. *ISO WD 12 200 Computational aids in terminology – Terminology interchange format (TIF) – An SGML Application*.
Interrante, C.G. and F.J. Heymann (eds). 1983. *Standardization of Technical Terminology*. West Conshohocken, PA: ASTM Special Technical Publication.
Jacobi, D. 1990. ''Lexique et Reformulation intradiscursive dans les documents de vulgarisation scientifique.'' In D. Candel (ed.), *Français Scientifique et Technique et dictionnaires de langue*. Paris: Didier Erudition, 77–91.
Jacquemin, C. and J. Royaute. 1994. ''Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework.'' *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. New York, Heidelberg: Springer-Verlag, 132–141.
Jahr, S. 1992. ''Zum Verhältnis von Bedeutung, Begriff und Wissen bei Fachtermini.'' *Fachsprache* (14)1–2: 38–44.
James, G. with R. Davison, H.A. Cheung and S. Deerwester. 1994. *English in Computer Science: A corpus-based lexical approach*. Asia: Longman.
Katz, J.J. 1964. ''Analyticity and Contradiction in Natural Language.'' In J.A. Fodor and J.J. Katz (eds), 519–539.
Katz, J.J. 1966. *The Philosophy of Language*. New York and London: Harper and Row.
Katz, J.J. and J.A. Fodor. 1964. ''The Structure of a Semantic Theory.'' J.A. Fodor and J.J. Katz (eds), 480–518.
Kenny, D. (forthcoming). ''Corpora in Translation Studies.'' In M. Baker (ed.), *The Translation Studies Encyclopedia*. London, New York: Routledge.
Kiefer, F., G. Kiss and J. Paizs (eds). 1992. *Papers in Computational Lexicography Complex '92*. Budapest: Linguistics Institute Hungarian Academy of Sciences.
Kircz, J.G. 1991. ''Rhetorical Structure of Scientific Articles: The case for argumentational analysis in information retrieval.'' *Journal of Documentation* (47)4: 354–372.
Kittredge, R. 1981. ''Cohesive Text Structure in Sublanguages.'' In B. Rieger (ed.), *Empirical Semantics*. Bochum: Brockmeyer. 447–466.
Kittredge, R. and J. Lehrberger. (eds). 1982. *Sublanguage: Studies of Language in Restricted Domains*. Berlin, New York: Walter de Gruyter.
Kleinedam, H. 1986. ''La notion de 'fonction lexicale' et son application lexicographique dans le dictionnaire explicatif et combinatoire du français contemporain d'I.A. Mel'čuk.'' *Actes du XVIIIe Congrès International de linguistique et de Philologie Romanes*, Université de Trèves. Tübingen: Max Niemeyer Verlag, 165–177.

Kocourek, R. 1982. *La Langue Française de la Technique et de la Science*. Wiesbaden: Brandstetter Verlag.

Koeppel, R. 1994. ''Satzbezogene Verweisformen in schriftlichen Fachtexten. Zu ihren Typen und Formen.'' *Fachsprache* (16)3–4: 103–115.

Kuang-hua, C. and C. Hsin-Hsi. 1995. ''Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and its Automatic Evaluation.'' Obtained from Internet: hh_chen @csie.ntu.edu.tw.

Kruyt, J.G. and E. Putter. 1992. Corpus Design Criteria. *INL Working Papers 92–11*.

Lackstrom, J., E. Selinker and L. Trimble. 1972. ''Grammar and Technical English.'' In J. Swales (ed.), *Episodes in ESP*. Exeter: Pergamon Press, 60–66. (Originally published in English Teaching Forum, X (5) 1972).

Landau, S.I. 1989. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Landheer, R. 1989. ''L'importance des relations hyponymiques dans la description lexicographique.'' *Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes*, Université de Trèves. Tübingen: Max Niemeyer Verlag, 139–151.

Leech, G.N. 1983. *Principles of Pragmatics*. London and New York: Longman

Leech, G.N. 1993. ''Corpus Annotation Schemes.'' In *Literary and Linguistic Computing*, 8(4). Oxford : Oxford University Press, 275–281.

Lehrberger, J. 1982. ''Automatic Translation and the Concept of Sublanguage.'' In R. Kittredge and J. Lehrberger (eds), 81–106.

Lehrberger, J. 1986. ''Sublanguage Analysis.'' R. Grishman and R. Kittredge (eds), 19–38.

Lehrer, A. 1974. *Semantic Fields and Lexical Structure*. Amsterdam: North-Holland Publishing Company.

Leitner, G. (ed.). 1992. *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter.

Lérat, P. 1986. ''Pour une étude comparée du vocabulaire des institutions dans les langues romanes.'' *Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes*, Université de Trèves. Tübingen: Max Niemeyer Verlag, 201–207.

Lérat, P. 1990. ''L'Hyperonymie dans la structuration des terminologies'' *Langages*, June 1990: 79–86.

Lérat, P. 1994. ''Terminologie vs. Lexicographie.'' In D. Candel (ed.), *Français Scientifique et Technique et dictionnaires de langue*. Paris: Didier Erudition. 27–36.

Lipka, L. 1990. *An Outline of English Lexicology*. Tübingen: Max Niemayer Verlag.

Loffler-Laurian, A-M. 1994. ''Les définitions dans la vulgarisation scientifique.'' In D. Candel (ed.) *Français Scientifique et Technique et dictionnaires de langue*. Paris: Didier Erudition. 93–111.

Lyons, J. 1963. ''Meaning.'' In *Structural Semantics*. Oxford: Blackwell, 51–90.

Lyons, J. 1966. ''Firth's Theory of Meaning.'' C.E. Bazell, J.C. Catford, M.A.K. Halliday (eds), *In Memory of J.R. Firth*. London: Longmans Linguistic Library. 288–302.

Lyons, J. 1977. *Semantics, 1–2*. Cambridge: Cambridge University Press.

McEnery, T. and A. Wilson. 1993. ''The Role of Corpora in Computer-Assisted Language Learning.'' *Call* (6)3, 233–248.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MacKenzie, I. 1990. ''The Vocable ESCAPE: A Family of Lexical Entries for an Explanatory Combinatorial Dictonary of English.'' In J. Steele (ed.), 95–130.

Macleod, C., S. Chen and J.M. Clifford. 1987. ''Parsing Unedited Medical Narrative.'' In N. Sager, C. Friedman and M.S. Lyman (eds), *Medical Language Processing*. Reading, Massachusetts: Addison-Wesley, 163–174.

McNaught, J. 1993. ''User Needs for Textual Corpora in Natural Language Processing.'' *Literary and Linguistic Computing*, 8(4). Oxford: Oxford University Press, 227–234.

Meijs, W. 1993. ''Analysing nominal compounds with the help of a computerized knowledge system.'' In J.Aarts, P.de Haan and N. Oostdijk (eds), *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi. 299–312.

Mel'čuk, I. (ed.). 1984. *Dictionnaire explicatif et combinatoire du français contemporain*. Montréal: Presses de l'Université de Montréal.

Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Meyer, I. 1990 ''Interlingual meaning-text lexicography: Towards a new type of dictionary for translation.'' In J. Steele (ed.), 175–197.

Meyer, I. and Steele, J. 1990. ''The Presentation of an Entry and Super-Entry in an Explanatory Combinatorial Dictionary of English.'' In J. Steele (ed.), 62–90.

Mortureux, M-F. 1994. ''L'analyse du discours de la vulgarisation scientifique et le dictionnaire de la langue scientifique.'' In D. Candel, (ed.) *Français scientifique et technique et dictionnaire de langue*. Paris: Didier Erudition, 63–75.

Nakhimovsky, A. 1990. ''Word Meaning and Syntactic Structure: Some Comparative Notes.'' In J. Steele (ed.), 3–17.

(–) 1993. *New Shorter Oxford English Dictionary*. Oxford: Oxford University Press

Nichols, J. 1990. ''The Meeting of East and West: Confrontation and Convergence in Contemporary Linguistics.'' In J.Steele (ed.), 18–37.

Nkwenti-Azeh, B. 1989. *An Investigation into the Structure of the Terminological Information contained in Special Language Definitions*. PhD Thesis. Manchester: UMIST

Nkwenti-Azeh, B. 1992. *Positional and Combinational Characteristics of Satellite Communications Terms*. Final Report, Eurotra Project. UK-CCL-UMIST.

Nutter, J.T. 1989. ''Representing Knowledge about Words.'' *Proceedings of 2nd Irish Conference on Artificial Intelligence and Cognitive Science*. Heidelberg, New York: Springer Verlag Workshops in Computing, 313–328.

O'Brien, S. 1993. *Sublanguage, text type and machine translation*. MA Thesis. Dublin City University.

O'Brien, S., M. Creed and E. Quinlan. 1992. *Sublanguage and Text Types*. Final Report. Eurotra Project. CEC, Luxembourg.

(–) 1888–1928. *OED: The Oxford English Dictionary*; originally entitled *A New English Dictionary on Historical Principles*, edited by J.A.H. Murray, H. Bradley, A. Craigie, C.T. Onions, and others. Oxford: Clarendon Press.

Opitz, K. 1982. ''LSP versus common language: the muddle of definiens and definiendum.'' *Proceedings of the 3rd European Symposium on LSP*. Paris: UNESCO. 185–196.

Opitz, K. 1983. ''The terminological standardised dictionary.'' In R.R.K. Hartmann (ed.), 163–180.

Palmer, J.R., 1968. *Selected Papers of J.R. Firth 1952–1959*. London: Longmans Linguistic Library.

Pearson, J. (ed.). 1992. *ET10/66 - Terminology and Extra-Linguistic Knowledge, Report 1*, CEC, DG-XIII, Luxembourg.

Pearson, J. (ed.). 1996. ''Electronic texts and concordances in the translation classroom.'' *Teanga 16*. Dublin: IRAAL. 86–96

Pearson, J. and D. Kenny. 1991. ''Terminology in Eurotra.'' In C. Copeland, J. Durand, S. Krauwer and B. Maegaard (eds), *Studies in Machine Translation and Natural Language Processing, 1: The Eurotra Linguistic Specifications*. Luxembourg: Office for Official Publications of the Commission of the European Communities, 161–163.

Peters, C., E. Picchi and L. Biagini. 1996. ''Parallel and comparable bilingual corpora in language teaching and learning.'' In S. Botley et al. (eds), 68–82.

Phillips, M. 1983. *Lexical Macrostructure in Science Text*. PhD Thesis. University of Birmingham.

Phillips, M. 1985. *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam: North-Holland Publishing Company, 197–217.

Phillips, M. 1989. *Lexical Structure of Text. Discourse analysis monograph, No.12*. University of Birmingham.

Picht, H. 1992. ''Terminologie, ein Trans- und interdisziplinäres Wissensgebiet. Die Entwicklung nach Eugen Wüster.'' *Fachsprache* (14)1–2: 2–18.

Picht, H. and J. Draskau. 1985. *Terminology: an Introduction*. Guildford: University of Surrey.

(–) 1996. *Pointer Final Report*. Obtained via Internet from www.mcs.surrey.ac.uk

Putnam, H. 1962. ''The Analytic and the Synthetic.'' In H. Feigl and G. Maxwell (eds), *Minnesota Studies in the Philosophy of Science*, nonloc: University of Minnesota Press, 358–397.

Quine, W.V. 1964. ''Speaking of Objects.'' In J.A. Fodor and J.J. Katz (eds), 446–459.

Quine, W.V. 1964. ''Meaning and Translation.'' In J.A. Fodor and J.J. Katz (eds), 461–478.

Quirk, R. 1982. ''On Corpus Principles and Design.'' In J.Svartvik (ed.), 457–469.

Quist, C. 1991. ''Semantic Features of Scientific and Technological Languages.'' *Copenhagen Studies of Language* 14: 24–38.

Reddick, A. 1990. *The Making of Johnson's Dictionary*. Cambridge: Cambridge University Press.

Rey, A. 1975. ''Terminologies et 'terminographie'.'' *Banque des Mots* 10: 145–154.

Rey, A. 1979. *La Terminologie, Noms et Notions*. Paris: Presses Universitaires de France.

Riegel, M. 1987. ''Définition directe et indirecte dans le langage ordinaire: les énoncés définitoires copulatifs.'' *Langue Française* 73: 29–53.

Roche, E. 1992. ''Looking for syntactic patterns in texts.'' In F. Kiefer, G. Kiss and J. Paizs, 279–287.

Roe, P.J. 1977. *The Notion of Difficulty in Scientific Text*. PhD Thesis. University of Birmingham.

Rondeau,G. 1984. *Introduction à la terminologie*. Québec: Gaëtan Morin.

Sager, J. 1984. ''Terminology and the Technical Dictionary.'' In R.R.K. Hartmann (ed.), 315–326.

Sager, J. 1990 *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins Publishing Company.

Sager, J., D. Dungworth and P.F. McDonald. 1980. *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

Sager, N. 1982. ''Syntactic formatting of Science Information.'' In R. Kittredge and J. Lehrberger (eds), 9–26. (Article reprinted from *AFIPS Conference Proceedings 41*. New Jersey: AFIPS Press, 791–800).

Saussure, F. de 1978. *Cours de Linguistique Générale*. Edition critique préparée par Tullio de Mauro. Paris: Payot.

Schneider, F. 1994. ''Semantische Vernetzung als konstitutives Prinzip der Makrostruktur eines aktiven Kontextfachwörterbuchs.'' In *Fachsprache* (16)3–4: 116–128.

Searle, J.R. 1970. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Selinker, L., R.M. Trimble, Todd and L. Trimble. 1976. ''On Reading English for Science and Technology: Presuppositional Rhetorical Information in the Discourse.'' In J.C. Richards (ed.), *Teaching English for Science and Technology*. Singapore: Singapore University Press, 37–67.

Selinker, L., E. Tarone, and V. Hanzeli (eds). 1981. *English for Academic and Technical Purposes: Studies in Honour of Louis Trimble*. Rowley MA: Newbury House.

Sinclair, J. 1966. ''Beginning the Study of Lexis.'' In C.E. Bazell, J.C. Catford and M.A.K. Halliday (eds), *In Memory of J.R. Firth*. London: Longmans Linguistic Library, 410–430.

Sinclair, J. 1987a. ''Introduction.'' In *Collins Cobuild English Language Dictionary*. Collins: London and Glasgow, xv–xxi.

Sinclair, J.M. (ed.) 1987b. *Looking Up*. London and Glasgow: Collins.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. 1992. ''The Automatic Analysis of Corpora.'' In J. Svartvik (ed.), 379–397.

Sinclair, J. 1994a. ''Corpus Typology: A Framework for Classification'' EAGLES document 1–18, now published as J. Sinclair 1995. ''Corpus Typology: A Framework for Classification.'' G. Melchers and B. Warren (eds), *Studies in Anglistics*. Stockhom: Almquist and Wiksell International, 17–34.

Sinclair, J. 1994b. ''Prospects for Automatic Lexicography'', The Otto Jaspersen Memorial Lecture. Copenhagen.

Sinclair, J. 1995. ''Introduction.'' In *Collins Cobuild English Dictionary*. London: HarperCollins. viii–xi.

Sinclair, J. 1996a. ''The Empty Lexicon.'' *International Journal of Corpus Linguistics* (1)1: 99–120.

Sinclair, J. 1996b. ''The Search for Units of Meaning.'' *Textus* 9: 75–106.

Sinclair, J. and J. Ball. 1995. *Eagles Text Typology*. Internal Working Document; available via ftp: ilc.pi.cnr.it.

Sinclair, J. and D.M. Kirby. 1990. ''Progress in Computational Lexicography.'' *Computational Lexicology and Lexicography*, Vol. VII. Pisa: Giardini Editori. 233–257.

Steele, J. (ed.) 1990. *Meaning-Text Theory*. Ottawa: University of Ottawa Press.

Steele, J. and I. Meyer. 1990. "Lexical Functions in an Explanatory Combinatorial Dictionary." In J. Steele (ed.), 41–61.

Strehlow, R.A. 1983. "Terminology and the Well-formed Definition." In C.G. Interrante and F.J. Heymann (eds), 15–25.

Strehlow, R.A. 1983. "The Varieties of Compound Terms and their Treatment." In C.G. Interrante and F.J. Heymann (eds), 26–33.

Stubbs, M. 1986. "Lexical Density: A Technique and Some Findings." In M. Coulthard (ed.), *Talking about Text*. Birmingham: English Language Research, 27–42.

Stubbs, M. 1993. "British Traditions in Text Analysis." In M. Baker et al. (eds), 1–33.

Stubbs, M. 1996. Text and Corpus Analysis. Oxford: Blackwell.

Svartvik, J. (ed.). 1992. *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82*. Berlin: Mouton de Gruyter.

Svensen, B. 1990. *Practical Lexicography*. Oxford: Oxford University Press.

Swales, J. 1971. *Writing Scientific English*. Lagos: Nelson Publishing Company.

Swales, J. 1981. "Definitions in Science and Law: Evidence for Subject-Specific Course Components?" *Fachsprache* 3: 106–112.

Swales, J. (ed.). 1985. *Episodes in ESP*. Oxford: Pergamon Press.

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Talbot, G. (ed.). 1993. *ET10/66 Terminology and Extra-Linguistic Knowledge*. Final Report. CEC, Luxembourg.

Teubert, W. 1996. "Comparable or Parallel Corpora?" *International Journal of Lexicography* (9)3: 238–264.

Temmerman, R. 1996. "Retrospect Lexicalisation: A Recurrent Phenomenon in the Lexicalisation Process of the Life Sciences." *Euralex '96 Proceedings II*. Göteborg: Göteborg University. 825–835.

Thomas, P. 1993. "Choosing Headwords from Language-for-Special Purposes (LSP) Collocations for Entry into a Terminology Data Bank (Term Bank)." In H.B. Sonneveld and K.L. Loening (eds), *Terminology*. Amsterdam: John Benjamins Publishing Company, 43–61.

Trimble, R.M.T. and L. Trimble. 1978. "The Development of EFL Materials for Occupational English: The Technical Manual." In R.M.T. Trimble, L. Trimble and K. Drobnic (eds), *English for Specific Purposes. Science and Technology*. English Language Institute, Oregon State University, 74–132.

Trimble, L. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Tognini-Bonelli, E. 1993. "Rationale and Aims of Corpus Linguistics." Internal memorandum Corpus Linguistics Group, University of Birmingham.

Van Dijk, T.A. *Handbook of Discourse Analysis*. London: Academic Press.

Varantola, K. 1992. "Words, terms and translators" *Euralex '92 Proceedings I–II*, Tampere: Studia Translatologica, 121–128.

Velardi, P. 1989. "Acquisition of Semantic Patterns from a Natural Corpus of Texts." *SIGART Newsletter: Knowledge Acquisition Special Issue*, April 1989, 115–123.

Wagner, H. "Les dictionnaires du français langue de spécialité/langue économique." *Actes*

du XVIIIe Congrès International de Linguistique et de Philologie Romanes, Université de Trèves. Tübingen: Max Niemeyer Verlag, 208–219

Walter, E. 1992. ''Semantic set-defining: benefits to the lexicographer and the user.'' *Euralex '92 Proceedings I–II*, Tampere: Studia Translatologica, 129–136.

Watson, R. 1985. ''Towards a theory of definition.'' *Journal of Child Language* (12)1: 181–197.

(–) 1991. *Webster"s New World Dictionary of American English*. New York: Prentice Hall.

Weise, G. 1992. ''Criteria for the Classification of ESP Texts.'' *Fachsprache* (14)1–2: 26–37.

Weise, G. 1994. ''Stages in the Comprehension of Scientific Texts.'' *Fachsprache* (16)3–4: 98–105.

White, J.S. 1988. ''Determination of lexical-semantic relations for multi-lingual terminology structures.'' In M. Walton Evens (ed.), *Studies in NLP: Relational Models of the Lexicon*. Cambridge: Cambridge University Press, 183–198.

Widdowson, H.G. 1979. *Explorations in Applied Linguistics*. Oxford: Oxford University Press.

Widdowson, H.G. 1979. *Reading and Thinking in English: Discovering Discourse*. Oxford: Oxford University Press.

Widdowson, H.G. and J.P.B. Allen. 1985. ''Teaching the Communicative Use of English.'' In J. Swales (ed.), *Episodes in ESP*. Oxford: Pergamon Press, 69–89. (Originally published in IRAAL XII, 1,1974, 1–20)

Winter, E. 1977. ''A clause-relational approach to English text: a study of some predictive lexical items in written discourse.'' *Journal of Instructional Science* (6)1: 1–92.

Winter, E. 1986. ''Clause Relations as Information Structure: Two Basic Text Structures in English.'' In M. Coulthard (ed.), *Talking about Text*. Birmingham: English Language Research, 88–108.

Wüster, E. 1931. *Die Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik*. Berlin: VDI-Verlag.

Wüster, E. 1968. *The Machine Tool: An Interlingual dictionary of basic concepts*. London: Technical Press.

Wüster, E. 1979. *Einführung in die allgemeine Terminologielehre und in die terminologische Lexikographie*. UNESCO ALSED LSP Network.

Yablo, S. 1993. ''Definitions Consistent and Inconsistent.'' *Philosophical Studies* 72: 147–175.

Yang, H.Z. 1986. ''A new technique for identifying scientific and technical terms and describing science texts.'' *Literary and Linguistic Computing*, 1(2). Oxford: Oxford University Press, 93–103.

Appendix A

Codes used by CLG tagger

DT = determiner
IN = preposition
JJ = adjective
NN = noun
NNS = nouns
VB = verb
VBN = past participle
VBG = present participle

Specifications for tag sequence pattern files used in Chapter 6
Specifications for ITU corpus

Tag sequence pattern file 1:
1) !JJ !DT JJ NN NN|NNS
2) !JJ !DT JJ NN|NNS

Tag sequence pattern file 2:
1) DT JJ NN NN|NNS
2) DT JJ NN|NNS

Tag sequence pattern file 3:
1) DT NN NN NN|NNS
2) DT NN NN|NNS
3) DT NN

Tag sequence pattern file 4:
1) !JJ !DT NN NN NN
2) !JJ !DT NN NN
3) !JJ !DT NN

Tag sequence pattern file 5:
DT NN IN|VBN NN !NN

Tag sequence pattern file 6:
1) DT VBN|VBG NN NN|NNS !IN
2) DT VBN|VBG NN|NNS !IN

Tag sequence pattern file 7:
DT NN NN VBG NN|NNS !IN

Tag sequence pattern file 8:
1) DT VB NN NN|NNS
2) DT VB NN|NNS

Specifications for GCSE corpus

Tag sequence pattern file 1:
1) !DT !JJ NN NN|NNS
2) !DT !JJ NN|NNS

Tag sequence pattern file 2:
1) DT VBG NN NN|NNS !IN
2) DT VBN NN|NNS !IN

Tag sequence pattern file 3:
1) DT NN NN|NNS
2) DT NN|NNS

Tag sequence pattern file 4:
1) DT JJ NN NN|NNS !IN
2) DT JJ NN|NNS

Tag sequence pattern file 5:
1) !DT JJ NN NN|NNS !IN
2) !DT JJ NN|NNS

Specifications for Nature corpus

Tag sequence pattern file 1:
1) DT NN NN NN|NNS

2) DT NN NN|NNS
3) DT NN|NNS

Tag sequence pattern file 2:
1) DT JJ JJ|NN NN NN|NNS
2) DT JJ NN NN|NNS
3) DT JJ NN|NNS

Appendix B

B.1 Concordance of ANKYRIN* from the Nature corpus

Ankyrin* occurs 85 times in the Nature corpus. The concordance provided here was produced using Wordsmith Tools and is sorted one word to the left of the node.

1 present at positions 834, 1,378 and 1,500. Ankyrin is also acylated by palmitic acid42,43, and the rapid turnover of fatty acid indicates a regulatory role43. Unfortunate

2 ankyrin called ankyrin 2.2 (or protein 2.2). **Ankyrin** 2.2 is particularly interesting because it is an 'activated' molecule with the ability to bind to all brain and erythr

3 ankyrin-like repeats function as 'built-in' **ankyrin**s and form binding sites for integral membrane proteins, tubulin or other proteins. This binding could regulate the sym

4 13,539; pI 4.1–4.2) and partially 'deactivates' **ankyrin**, muting its affinity for the erythroid anion exchanger3O. The exquisite protease sensitivity of the 55K region (V.B.,

5 structural analysis and binding studies. Also, **ankyrin** and the red-cell membrane are very well understood compared with the ankyrin homologues and the membranes of the

6 allow the number of repeats to vary among ankyrins or **ankyrin**= related proteins (see below) without altering the contact points or conformation of individual repeats. T

7 less than that of native ankyrin3O. **Ankyrin** repeats in the 89K domain The 89K domain is almost entirely composed of a repeated sequence motif. This is easily de

8 antibody to a C-terminal peptide stains both **ankyrin**s (Fig. 1e). Because the two variants only differ slightly on peptide maps4,5,29, they are probably identical except fo

9 by digestion with trypsin23,24. Brain **ankyrin** has a similar structure8. The 89K domain binds the erythroid anion exchanger25,26 and tubulin9,25. The 62K domain bind

10 be polarized to function properly. Brain-type **ankyrin**s are more widely distributed and bind to different integral proteinsl7, none of which have been identified.

11 from a variant of erythroid ankyrin called **ankyrin** 2.2 (or protein 2.2). Ankyrin 2.2 is particularly interesting because it is an 'activated' molecule with the ability t

12 (protein 2.1) but not a smaller variant called **ankyrin** 2.2 (Fig. 1e). By contrast, an antibody to a C-terminal peptide stains both ankyrins (Fig. 1e). Because the two varian

13 Complementary DNA cloning We cloned **ankyrin** from a human reticulocyte cDNA library3l to ensure isolation of the erythrocyte isoform. We obtained a single clone, p

14 GAG codon), Asp for pAnk15 (GAC codon). **Ankyrin** contains three structural domains The deduced ankyrin sequence is divided into three regions corresponding to the do

15 other tissues. Among the proteins containing **ankyrin** or ankyrin-like repeats, only ankyrin is available in amounts suitable for structural analysis and binding stud-

16 three structural domains The deduced **ankyrin** sequence is divided into three regions corresponding to the domains defined by chymotrypsin cleavage. The 89K and 62K

17 their unique biconcave shape. A different **ankyrin** has been purified from brain tissue8 and there are hints that other ankyrins exist9–11. Ankyrins are found in many oth

18 during Drosophila differentiation51,70. **Ankyrin**-like repeats could also be the site of the genetically postulated association between Notch and one of the Enhancer of

19 (7,252 bp). Verifying that the cDNAs encode **ankyrin** We verified the identity of the ankyrin clones by comparison with N-terminal amino-acid sequences of 34 peptides

20 and cytoskeletal elements. Erythrocyte **ankyrin** (protein 2.1) is the best-characterized isoform. lt attaches the spectrin skeleton (membrane skeleton) to band 3, the

21 of which have been identified. Erythrocyte **ankyrin** contains two principal chymotrypsin-resistant domains that were originally judged to have relative molecular masses of

22 Analysis of cDNA for human erythrocyte **ankyrin** indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins Samuel E. Lu

23 from proteolytic digests of erythrocyte **ankyrin** or its chymotryptic domains. Each of these sequences was found in the translatedankyrin cDNA sequence (Fig. 2, underl

24 ankyrin. The availability of erythrocyte **ankyrin** clones will also facilitate cloning of other ankyrins, analysis of ankyrin genomic structure, and detection of ankyrin

25 a more complete analysis of the erythrocyte **ankyrin** gene is required to confirm this assignment. The ankyrin repeats were presumably formed by successive gene dupli-

26 define the structural domains of erythrocyte **ankyrin**, the first-described member of the family, and identify a cluster of characteristic repeats also present in slightly a

27 using expressed segments of erythrocyte **ankyrin**. The availability of erythrocyte ankyrin clones will also facilitate cloning of other ankyrins, analysis of ankyrin ge

28 first determined the structure of erythrocyte **ankyrin**. Complementary DNA cloning We cloned ankyrin from a human reticulocyte cDNA library3l to ensure isolation of the

28 the sequence of human erythrocyte **ankyrin**. They find alternative sequences at the C terminus that probably result from differential splicing: (1) deletion of bp

29 ll-characterized ankyrinsl7. Erythrocyte-type **ankyrin**s are typically confined to specific membrane domainsl4,l7 and bind integral membrane proteins, such as the (Na+ + K+)-

30 and closely resemble the repeats in erythroid **ankyrin** (compare Figs 3b and 4b). The principal differences are: (1) residue 3 is much more polar in the ankyrin-like

31 a consensus sequence of the erythroid **ankyrin** repeats and found very similar repeats in several invertebrate, yeast and viral proteins (Fig. 4). Multiple searches w

32 of complementary DNA for human erythroid **ankyrin** indicates that the mature protein contains 1,880 amino acids comprising an N-terminal domain binding integral membrane

33 other is missing from a variant of erythroid **ankyrin** called ankyrin 2.2 (or protein 2.2). Ankyrin 2.2 is particularly interesting because it is an 'activated' molecule wit

34 the ankyrin-like repeats than it is in erythroid **ankyrin** repeats; (2) the consensus Gly-His dipeptide at positions 13–14 in ankyrin repeats is not conserved in ankyrin-like re

35 the ability to bind to all brain and erythroid **ankyrin** sites, and to additional sites of its ownl7. Ankyrins therefore facilitate and probably control interactions betwee

36 there are hints that other ankyrins exist9–11. **Ankyrin**s are found in many other, and perhaps all, cells8–17. They have been provisionally subdivided into 'erythrocyte' and '

37 alternatively spliced sequence missing from **ankyrin** variant 2.2. The N-terminal domain is almost entirely composed of 22 tandem 33-amino-acid repeats. Similar repeats ar

38 to a peptide in the insert stains full-length **ankyrin** (protein 2.1) but not a smaller variant called ankyrin 2.2 (Fig. 1e). By contrast, an antibody to a C-terminal peptide

39 ankyrin should be useful in deciphering how **ankyrin** homologues function in tissue differentiation and cell-cycle control. Note added in proof: In parallel experiments S

40 for palmitoylation has been identified. **Ankyrin** 2.2 lacks some of the regulatory domain The central third of the C-terminal 55K domain contains a 486-bp sequence th

41 with ankyrin, even though the repeats in **ankyrin** are more homologous than those in spectrin:37.0+ 7.6% (s.d.) identical to each other on average (range, 12–58%). Stru

42 of protein 2.2. The absence of this insert in **ankyrin** 2.2 'activates' the protein and enhances binding to spectrin and to sites on membranes, especially kidney membranesl7,

43 Gly-His dipeptide at positions 13–14 in **ankyrin** repeats is not conserved in ankyrin-like repeats; (3) the consensus Asn/Asp at position 29in the ankyrin repeats is le

44 in ankyrin repeats is not conserved in **ankyrin**-like repeats; (3) the consensus Asn/Asp at position 29in the ankyrin repeats is less conserved in the ankyrin-like rep

45 We postulate that the invertebrate **ankyrin**-like repeats function as 'built-in' ankyrins and form binding sites for integral membrane proteins, tubulin or other p

46 residue 1,382 are isolated from native **ankyrin** under digestion conditions in which the 89K and 62K domains are not released. Therefore the 89K domain ends somewhere

47 identical to the N-terminal sequence of native **ankyrin**. Presumably the initiator methionine is removed during synthesis exposing the underlying proline. This is a well known

48 dimers34,35, and reduces the capacity of **ankyrin** to bind band 3, the anion-exchange protein36; but it is difiicult to identify the phosphorylation sites, because the s

49 domain that regulates the binding of **ankyrin** to spectrin and the anion-exchange protein3O. The function of ankyrin is also regulated by phosphorylation34–36. In vi

50 The deduced sequence of **ankyrin** extends for 1,880 amino acids from the N-terminal proline (corresponding to an Mr of 206,144; p1 5.95), and the 5' and

51 anion-exchange protein3O. The function of **ankyrin** is also regulated by phosphorylation34–36. In vitro, up to seven Ser-Thr phosphates are added to ankyrin by the

52 ankyrin genomic structure, and detection of **ankyrin** defects in red cells and other tissues. Among the proteins containing ankyrin or ankyrin-like repeats, only ankyrin is

53 cloning of other ankyrins, analysis of **ankyrin** genomic structure, and detection of ankyrin defects in red cells and other tissues. Among the proteins containing anky

54 about the structure and function of **ankyrin** should be useful in deciphering how ankyrin homologues function in tissue differentiation and cell-cycle control. N

55 host-range proteins contain two and one **ankyrin**-like repeats respectively. The single repeat in the cowpox virus (residues 458–490; ref. 63) is 39% identical to the A

56 ankyrin or ankyrin-like repeats, only **ankyrin** is available in amounts suitable for structural analysis and binding studies. Also, ankyrin and the red-cell membrane

57 Among the proteins containing ankyrin or **ankyrin**-like repeats, only ankyrin is available in amounts suitable for structural analysis and binding studies. Also, ankyrin

58 of repeats to vary among ankyrins or **ankyrin**= related proteins (see below) without altering the contact points or conformation of individual repeats. This is not so

59 brain tissue8 and there are hints that other **ankyrin**s exist9–11. Ankyrins are found in many other, and perhaps all, cells8–17. They have been provisionally subdivided into

60 clones will also facilitate cloning of other **ankyrin**s, analysis of ankyrin genomic structure, and detection of ankyrin defects in red cells and other tissues. Among the pr

61 sites, and to additional sites of its ownl7. **Ankyrin**s therefore facilitate and probably control interactions between integral membrane proteins and cytoskeletal elements.

62 cerevisiae)6O. Each contains two separated **ankyrin**-like repeats. The recently reported SWl4 gene product is not shown in Fig. 4, but its repeats are not significantly di

63 biochemical mechanism. Finally, six or seven **ankyrin**-like repeats have recently been discovered in bcl-3, a candidate proto-oncogene on chromosome 19 that is activated

64 and a cytoplasmic domain containing six **ankyrin**-like repeats. All three proteins regulate tissue differentiation. The Notch gene product is required for correct diffe

65 and differentiation. Discussion In summary, **ankyrin**s constitute a new family of proteins that seem to function as ‘molecular brokers’; that is, they coordinate interactio

66 and microscopic data indicate that **ankyrin** is globular, except for the proteolytically sensitive regulatory domain3O. For a globular domain, isotropic subunits c

67 is observation strengthens the inference that **ankyrin**-like repeats are involved in cell growth and differentiation. Discussion In summary, ankyrins constitute a new fa

68 revealed no other significant homologies. The '**ankyrin**-like' repeats are easily detected on dot-matrix plots (Fig. 4a) and closely resemble the repeats in erythroid ankyrin

69 the consensus Asn/Asp at position 29in the **ankyrin** repeats is less conserved in the ankyrin-like repeats; (4) the homologues favour Asp/Asn at the penultimate residue

70 for these repeats. We postulate that the **ankyrin** repeats are arranged in an isotropic array, and that different repeats, singly or in combination, form binding sites f

71 This is easily detected by comparing the **ankyrin** amino-acid sequence with itself using a dot-matrix analysis (Fig. 30). About the first 40 and last 60 amino acids are

72 is required to confirm this assignment. The **ankyrin** repeats were presumably formed by successive gene duplications. We looked for patterns that would reveal some of the e

73 Multiple searches with other portions of the **ankyrin** sequence revealed no other significant homologies. The 'ankyrin-like' repeats are easily detected on dot-m

74 ankyrin We verified the identity of the **ankyrin** clones by comparison with N-terminal amino-acid sequences of 34 peptides isolated from proteolytic digests of erythroc

75 are very well understood compared with the **ankyrin** homologues and the membranes of the cells containing them. The simple discovery that the cytoplasmic repeats

76 are: (1) residue 3 is much more polar in the **ankyrin**-like repeats than it is in erythroid ankyrin repeats; (2) the consensus Gly-His dipeptide at positions 13–14 in ankyri

77 repeats prefer threonine; (5) some of the **ankyrin**-like repeats (not shown in Fig. 4b) in each of the three invertebrate proteins have extra amino acids between or after

78 the ankyrin repeats is less conserved in the **ankyrin**-like repeats; (4) the homologues favour Asp/Asn at the penultimate residue, whereas ankyrin repeats prefer threonine;

79 and glp-1 gene products are related to **ankyrin** limits their possible functions, and the evolutionary conservation of these repeats emphasizes their importance. This

80 p to seven Ser-Thr phosphates are added to **ankyrin** by the erythrocyte membrane cyclic AMP-independent casein kinase I34. This abolishes the preferential binding of unpho

81 sequences was found in the translated **ankyrin** cDNA sequence (Fig. 2, underlined amino acids). Five polymorphisms were detected by variations between the peptide

82 kb) are compatible with the sizes of the two **ankyrin** RNAs observed on lightly exposed northern blots (6.8 and 7.2 kb) (Fig. 1 d). The extra in-frame sequence encodes a pol

83 the preferential binding of unphosphorylated **ankyrin** to spectrin tetramers and oligomers rather than to spectrin dimers34,35, and reduces the capacity of ankyrin to bind b

84 sp/Asn at the penultimate residue, whereas **ankyrin** repeats prefer threonine; (5) some of the ankyrin-like repeats (not shown in Fig. 4b) in each of the three invertebrat

85 amino acids; but it was not informative with **ankyrin**, even though the repeats in ankyrin are more homologous than those in spectrin:37.0+ 7.6% (s.d.) identical to each ot

B.2 Concordance of respiration from the GCSE corpus

There were 91 occurrences of respiration in the GCSE corpus. The concordance file was produced using Wordsmith Tools and the file is sorted one word to the left of the node.

1 minerals needed for plant growth. TOPIC 13 **RESPIRATION** Every living organism needs energy to keep itself alive. Movement, growth, reproduc- tion, feeding and excre-

2 energy. More energy is set free in aerobic **respiration** than in anaerobic respiration. Too much lactic acid will soon stop muscle cells working: the sprinter

3 food by chemical reactions in cells. Aerobic **respiration** setting free the energy in food by using oxygen. Anaerobic respiration setting free the energy in food wit

4 To find out if oxygen is used up in aerobic **respiration** Put a few small animals such as woodlice into a syringe. A piece of sponge should separate the woodlice from some l

5 respiration is different from aerobic **respiration** because in anaerobic respiration:1. the energy in sugar is set free without using oxygen to do it, 2. sugar is no

6 food and enzymes. 2 Oxygen for aerobic **respiration** of seed tissue. This releases the energy needed for growth and development. 3 A suitable temperature for optimum e

7 both can happen at the same time. Aerobic **respiration** uses oxygen to set energy free. Anaerobic respiration does not use oxygen to set energy free. AEROBIC RESPIRA-

8 not use oxygen to set energy free. AEROBIC **RESPIRATION** Energy is usually set free from fat or sugar in respiration. ‘to respire’ means ‘to do respiration’ Figure 13.

9 get oxygen quickly enough for aerobic **respiration**, they change to anaerobic respiration. For the first few metres of a sprint your muscles use aerobic respiration.

10 is involved in the process it is called aerobic **respiration**. Most plant and animal cells respire aerobically and the reaction can be represented by this equation: The energy re

11 efflcient at releasing energy than aerobic **respiration**. Anaerobic respiration in yeast is used commercially in baking and brewing. Yeast breaks down glucose to obtain i

12 less energy is set free than from aerobic **respiration**. Important words Respiration the setting free of the energy in food by chemical reactions in cells. A

13 metres of a sprint your muscles use aerobic **respiration**. Anaerobic respiration is used for the rest of the race. (See figure 13.5.) Your muscles can work hard without oxyge

14 ANAEROBIC RESPIRATION In anaerobic **respiration** the energy in food is set free without using oxygen to do it. When you walk, your muscles are working slowly. Th

15 uses oxygen to set energy free. Anaerobic **respiration** does not use oxygen to set energy free. AEROBIC RESPIRATION Energy is usually set free from fat or sugar in respir

16 energy than aerobic respiration. Anaerobic **respiration** in yeast is used commercially in baking and brewing. Yeast breaks down glucose to obtain its energy and at the sa

17 both flasks must be compared. ANAEROBIC **RESPIRATION** In anaerobic respiration the energy in food is set free without using oxygen to do it. When you walk, your

18 energy in food by using oxygen. Anaerobic **respiration** setting free the energy in food without using oxygen to do it. Fermentation a kind of anaerobic re

19 to do it. Fermentation a kind of anaerobic **respiration** which makes alcohol. Things to do Bread is quick and easy to make, so bake yourself a loaf (in a kitchen, not the l

20 muscles use aerobic respiration. Anaerobic **respiration** is used for the rest of the race. (See figure 13.5.) Your muscles can work hard without oxygen for about 15 seconds.

21 from fermentation. This is a kind of anaerobic **respiration** in which alcohol is made instead of lactic acid. Figure 13.7 shows that in fermentation, organisms: USE UP

22 dough lighter and better to eat. Anaerobic **respiration** is different from aerobic respiration because in anaerobic respiration:1. the energy in sugar is set free without

23 free in aerobic respiration than in anaerobic **respiration**. Too much lactic acid will soon stop muscle cells working: the sprinter cannot carry on sprinting! At the end of

24 there is no oxygen. This is called anaerobic **respiration**. It is less efflcient at releasing energy than aerobic respiration. Anaerobic respiration in yeast is used commercia

25 respiration, they change to anaerobic **respiration**. For the first few metres of a sprint your muscles use aerobic respiration. Anaerobic respiration is used for the

26 aerobic respiration because in anaerobic **respiration**:1. the energy in sugar is set free without using oxygen to do it, 2. sugar is not completely broken down to ca

27 for humans and animals. Photosynthesis and **respiration** are very complex processes involving several steps. photosynthesis is the reverse of respiration. Respiration is

28 with the air during photosynthesis and **respiration**, 3. they give off water during transpiration. The petiole carries water and mineral salts from the stem in

29 in Fig. 5.5 suggests a balance between **respiration** and photo- synthesis. When the sun is shining, plants can use the carbon dioxide about as fast

30 (It does not really show that it is caused by **respiration** however. It could be caused by another chemical reaction.) If the temperature of the control flask changes, the chan

31 from carbon compounds in their food, by **respiration**, because they need energy to stay alive. The main energy-giving foods which contain carbon compounds are carbohydrat

32 for releasing energy from food by **respiration**. lt is surrounded by a membrane. It is often called the power house' of the cell. The number of mitochondria each ce

33 release carbon dioxide (in a process called **respiration**), and it is also formed when fuels burn. Carbon dioxide is used up when plants photosyntesize (see section I). Al

34 in our bodies by a chemical process called **respiration**. During respiration, foods react with oxygen forming carbon dioxide and water. This is why we breathe out carbon

35 which happen in all living cells. This is called **respiration**. It is one of the most important chemical reactions that happens in cells. There are two kinds of respiration: aerob

36 set free the energy in its food. This is called **respiration**. Oxygen diffuses into Amoeba from the higher concentration of oxygen in the water (figure 3.4a). An A

37 place at the cellular level. It is called cellular **respiration** and is a form of catabolism. In cellular respiration, glucose is broken down step by step. If oxygen is involved

38 and is a form of catabolism. In cellular **respiration**, glucose is broken down step by step. If oxygen is involved in the process it is called aerobic respiration. Most pl

39 sugar in respiration. 'to respire' means 'to do **respiration**' Figure 13. 1 shows that when organisms respire aerobically they: USE UP oxygen and sugar (or fat), GIVE

40 circulates more quickly. This happens during **respiration** when carbon leaves the organism, as carbon dioxide gas, and energy is set free. CARBON AND FUEL Coal, oil,

41 C and H) dioxide The energy given out during **respiration** can be used as: * heat to keep us warm; * mechanical energy in our muscles to help us to move around and keep

42 a chemical process called respiration. During **respiration**, foods react with oxygen forming carbon dioxide and water. This is why we breathe out carbon dioxide and water

43 + oxygen – carbon dioxide + water During **respiration**, foods containing carbon and hydrogen react with oxygen to form carbon dioxide and water: food + ox

44 are waste substances made by cells during **respiration**. This is the main chemical reaction which happens in cells. HOW UREA IS MADE Protein in food is digested in your

45 animals (including mice) breathe out during **respiration**. We have not, as yet, demonstrated that carbon dioxide is necessary to produce a healthy plant. In order to

46 is given out by living organisms during **respiration**. When work is done, or energy changed from one form to another, some energy is lost as heat. Activity 16.3 provid

47 during photosynthesis and used up during **respiration**. Starch is a complex carbohydrate. lts relative molecular mass is about 1 00 000. Plants store most of their carbohy

48 O6 + 602 ——=== 6CO2 + 6H2O + energy **Respiration** happens in both plants and animals, but photosynthesis can only happen in plants. Respiration and photosynthesis

49 time. The lost energy has been used for **respiration** and excretion and is not available for the next stage. This goes some way towards explaining why plant products, s

50 all living things must respire all the time. For **respiration** they must take in oxygen and get rid of carbon dioxide. A word equation for respiration may be written as follows:

51 supplies them with glucose and oxygen for **respiration** and removes the carbon dioxide and water produced. Wounds and cuts Phagocytes are carried by the blood to the site

52 rid of carbon dioxide. A word equation for **respiration** may be written as follows: Food + Oxygen – Energy + Carbon dioxide + Water The reverse of this i

53 oxygen in them. The cells use the oxygen for **respiration**. 3. As it gives up its oxygen, oxyhaemoglobin changes back to haemoglobin. 4. Haemoglobin picks up more

54 every day. 200cm3 of this is obtained from **respiration**, but 2300 cm 3 comes from the food and drink taken into the body.Table 9.2 shows the adult daily balance of water in

55 the water they get from their food and from **respiration**. They produce very little urine. Because of this they have very little odour, and people like them as pets. 5.2 Th

56 fossil fuel savings run out. 7 Foods as Fuels **Respiration** Foods are broken down in our bodies by a chemical process called respiration. During respiration, foods react with o

57 into your cells and they use it for growth, **respiration** etc. Hormones are chemicals which affect all the cells in your body. They are made in glands and carried in your blo

58 or exchanging carbon dioxide and oxygen in **respiration** and photosynthesis. Figure 5.8 shows the different parts of a leaf, listed here. 1. The cuticle is a laye

59 made by plants in photosynthesis is used in **respiration** to set free energy. Plants need mineral salts as well as carbon dioxide and water. The mineral salts are used to

60 the carbon dioxide given off by your cells in **respiration** and this turns the limewater in tube B milky. To find out if energy is set free by germinating seeds If living orga

61 seeds If living organisms give off heat in **respiration** this means that energy is also being set free. This is because heat is one kind of energy. Soak some wheat grain

62 is used. Carbohydrates and fats are used in **respiration** to give energy for growth and other activities such as the formation of new cell walls. Amino acids are used for

63 old ones. Fats are used by your cells:1. in **respiration** to set free energy, 2. to store energy. Much more fat can be stored in your body than glycogen. Digestion

64 22. 10.) Glucose is used by your cells:1. in **respiration** to set free energy, 2. to store energy as glycogen. Glycogen stored in your liver and muscles can be changed bac

65 of ponds in the heat of the summer. In **respiration**, oxygen is used up and carbon dioxide is produced. Fish and the plankton in the water consume oxygen. So do the deca

66 other parts of the plant and stored or used in **respiration**. The mid-rib supports the leaf. It has vascular bundles inside it which carry water and solutions of mineral sal

67 is usually set free from fat or sugar in **respiration**. 'to respire' means 'to do respiration' Figure 13. 1 shows that when organisms respire aerobically they: USE UP

68 important simple carbon compound. lt links **respiration** and photosynthesis and it is produced when carbon compounds burn. Carbon dioxide also has some important uses.

69 They are move- ment, sensitivity, nutrition, **respiration**, excretion, growth, and reproduction. There are three other things that can be said about all living organisms,

70 linked together in metabolism are nutrition, **respiration**, and excretion. 6 Growth All organisms grow by adding new material from within themselves. Some non- living

71 oxygen are connected in the processes of **respiration** and photosynthesis we shall deal with them as a single cycle. Besides water, respiration has another product, carbon

72 fuel, which is burned during the process of **respiration** in order to supply the energy needed to drive our living processes. There are two kinds of carbohydrate, sugars and

73 that energy is involved. When the rate of **respiration** is inhibited the rate of salt uptake is also inhibited. When respiration rate increases so does salt uptake. 24.6

74 cycle (Fig. 5.4). One of the products of **respiration** is water: Some animals, for example gerbils. are adapted to living in a desert. They may never drink water,

75 dioxide and water, produced as a result of **respiration**, and the nitrogenous compounds

76 in particular the chemical waste products of **respiration**, such as carbon dioxide and urea. . Have a nutrition system. This means being able to collect, absorb or eat foo

77 these foods for energy in the process of **respiration**. In this way, all organisms enter the water cycle (Fig. 5.4). One of the products of respiration is water: S

78 is released from food by the process of **respiration**. You possess threc different types of muscle in your body, each of which performs specific functions for you. Car

79 the energy from their food by the process of **respiration**. The waste products of metabolism are got rid of from the body by excretion. So the three characteristics linked tog

80 of decay by the decomposers is a form of **respiration**. Extending the cycle The cycle shown in Fig. 5.5 suggests a balance between respiration and photo- synthesis. When t

81 to break down food in the process of **respiration**. ln this process. the bacteria produce water, carbon dioxide, and energy. Other bacteria can grow only in the abse

82 steps. photosynthesis is the reverse of **respiration**. Respiration is exothermic. lt uses up food and oxygen and produces carbon dioxide, water and energy. ln contrast ph

83 that happens in cells. There are two kinds of **respiration**: aerobic and anaerobic. In most cells both can happen at the same time. Aerobic respiration uses oxygen to set ene

84 photosynthesis can only happen in plants. **Respiration** and photosynthesis are important in the carbon cycle (figure 2). This shows how carbon, in carbon dioxide and in oth

85 Figure 5.5 shows these opposite processes, **respiration** and photosynthesis, in a sirnple cycle. The action of decay by the decomposers is a form of respiration. Extending t

86 photosynthesis is the reverse of respiration. **Respiration** is exothermic. lt uses up food and oxygen and produces carbon dioxide, water and energy. ln contrast photosynthesis

87 They can get plenty of oxygen and their **respiration** is aerobic. When you sprint, your muscles work hard and need a lot of oxygen. When they cannot get oxygen quickly

88 with them as a single cycle. Besides water, **respiration** has another product, carbon dioxide. In order to get the energy from food, all living things must respire all the ti

89 vapour and why urine is mainly water. **Respiration** is also exothermic and energy is given out. Because of this, foods are sometimes described as ‘biological fuels’.

90 the rate of salt uptake is also inhibited. When **respiration** rate increases so does salt uptake. 24.6 Plant transport – still some unanswered questions! You have discovered

91 from aerobic respiration. Important words **Respiration** the setting free of the energy in food by chemical reactions in cells. Aerobic respiration sett

B.3 Concordance of respire* in the GCSE corpus

There were 13 occurrences of respire* in the GCSE corpus.

1 The cell feeds, reproduces, excretes and **respire**s. It is also sensitive and may move. The cells of multicellular organisms live together with other cells which they ne

2 part of a raw egg. Amoeba moves, feeds, **respire**s, reproduces, excretes, grows and is sensitive. Movement A bulge called a pseudopodium pushes out and the cytoplasm flo

3 set free from fat or sugar in respiration. 'to **respire**' means 'to do respiration' Figure 13. 1 shows that when organisms respire aerobically they: USE UP oxygen an

4 on' Figure 13. 1 shows that when organisms **respire** aerobically they: USE UP oxygen and sugar (or fat), GIVE OFF carbon dioxide gas and wat

5 to water. (See figure 13.2.) As the woodlice **respire** they use up oxygen. The carbon dioxide they give off is absorbed by the soda lime. The amount of air in the syringe ther

6 seconds. Figure 13.6 shows that when cells **respire** anaerobically you: USE UP sugar, MAKE lactic acid, SET FREE energy. More energy is set fr

7 in animals and plants. HOW BACTERIA **RESPIRE** A few bacteria can be either aerobic or anaerobic, depending upon whether they can get oxygen. Some purely anaerobic ba

8 the energy from food, all living things must **respire** all the time. For respiration they must take in oxygen and get rid of carbon dioxide. A word equation for respiration ma

9 respiration. Most plant and animal cells **respire** aerobically and the reaction can be represented by this equation: The energy released is used by the body to carry out a

10 Some organisms, for example yeast, can **respire** when there is no oxygen. This is called anaerobic respiration. It is less efflcient at releasing energy than aerobic res

11 things need to burn, and living things need to **respire**; carbon dioxide is the food that plants use to grow; nitrogen is rather slow to do anything, chemically, so that the

12 to face the sun, which most plants do. . **Respire**. This means more than just breathing; it is what happens inside muscles and cells where food substances combine w

13 colour. The lungs All the cells of the body **respire**. They use oxygen to release the energy from food. In this process water and carbon dioxide are released as waste products.

Index

In the series *Studies in Corpus Linguistics (SCL)* the following titles have been published thus far or are scheduled for publication:

22 **SCOTT, Mike and Christopher TRIBBLE:** Textual Patterns. Key words and corpus analysis in language education. *Expected April 2006*

21 **GAVIOLI, Laura:** Exploring Corpora for ESP Learning. 2005. xi, 176 pp.

20 **MAHLBERG, Michaela:** English General Nouns. A corpus theoretical approach. 2005. x, 206 pp.

19 **TOGNINI-BONELLI, Elena and Gabriella DEL LUNGO CAMICIOTTI (eds.):** Strategies in Academic Discourse. 2005. xii, 212 pp.

18 **RÖMER, Ute:** Progressives, Patterns, Pedagogy. A corpus-driven approach to English progressive forms, functions, contexts and didactics. 2005. xiv + 328 pp.

17 **ASTON, Guy, Silvia BERNARDINI and Dominic STEWART (eds.):** Corpora and Language Learners. 2004. vi, 312 pp.

16 **CONNOR, Ulla and Thomas A. UPTON (eds.):** Discourse in the Professions. Perspectives from corpus linguistics. 2004. vi, 334 pp.

15 **CRESTI, Emanuela and Massimo MONEGLIA (eds.):** C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages. 2005. xviii, 304 pp. (incl. DVD).

14 **NESSELHAUF, Nadja:** Collocations in a Learner Corpus. 2005. xii, 332 pp.

13 **LINDQUIST, Hans and Christian MAIR (eds.):** Corpus Approaches to Grammaticalization in English. 2004. xiv, 265 pp.

12 **SINCLAIR, John McH. (ed.):** How to Use Corpora in Language Teaching. 2004. viii, 308 pp.

11 **BARNBROOK, Geoff:** Defining Language. A local grammar of definition sentences. 2002. xvi, 281 pp.

10 **AIJMER, Karin:** English Discourse Particles. Evidence from a corpus. 2002. xvi, 299 pp.

9 **REPPEN, Randi, Susan M. FITZMAURICE and Douglas BIBER (eds.):** Using Corpora to Explore Linguistic Variation. 2002. xii, 275 pp.

8 **STENSTRÖM, Anna-Brita, Gisle ANDERSEN and Ingrid Kristine HASUND:** Trends in Teenage Talk. Corpus compilation, analysis and findings. 2002. xii, 229 pp.

7 **ALTENBERG, Bengt and Sylviane GRANGER (eds.):** Lexis in Contrast. Corpus-based approaches. 2002. x, 339 pp.

6 **TOGNINI-BONELLI, Elena:** Corpus Linguistics at Work. 2001. xii, 224 pp.

5 **GHADESSY, Mohsen, Alex HENRY and Robert L. ROSEBERRY (eds.):** Small Corpus Studies and ELT. Theory and practice. 2001. xxiv, 420 pp.

4 **HUNSTON, Susan and Gill FRANCIS:** Pattern Grammar. A corpus-driven approach to the lexical grammar of English. 2000. xiv, 288 pp.

3 **BOTLEY, Simon Philip and Tony McENERY (eds.):** Corpus-based and Computational Approaches to Discourse Anaphora. 2000. vi, 258 pp.

2 **PARTINGTON, Alan:** Patterns and Meanings. Using corpora for English language research and teaching. 1998. x, 158 pp.

1 **PEARSON, Jennifer:** Terms in Context. 1998. xii, 246 pp.