TERM RECOGNITION USING COMBINED KNOWLEDGE SOURCES

A thesis submitted to the Manchester Metropolitan University
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

June 2000

By
Diana Gabrielle Maynard
Department of Computing and Mathematics

Contents

Copyright **14**

Abstract **15**

Declaration **16**

Acknowledgements **17**

1 Introduction **18**

1.1 Term Recognition Using Combined Knowledge Sources 19

1.2 Organisation of this Work 21

2 Theoretical Terminology **22**

2.1 Overview 22

2.2 Terms and Concepts 23

2.2.1 Relation between term and concept 24

2.2.2 Terms and termforms 27

2.2.3 Contexts 28

2.3 The Cognitive Dimension 29

2.3.1 Labelling 30

2.3.2 Packaging 31

2.3.3 Network building 33

2.3.4 Recognising words and terms 34
2.4 Term Formation 36
2.4.1 Use of existing resources 37
2.4.2 Modification of existing resources 37
2.4.3 Creation of new linguistic entities 37
2.4.4 Primary and secondary term formation 38
2.4.5 Summary of term formation 38
2.5 Multi-word Terms 39
2.5.1 Formation of multi-word terms 39
2.5.2 Interpretation of multi-word terms 40
2.6 Ambiguity in Terms 41
2.6.1 Types of ambiguity 44
2.6.2 Term ambiguity in the field of medicine 53
2.7 Term Variation 56
2.7.1 Formation of variants 59
2.7.2 The importance of variants 63
2.8 Summary 64

3 Practical Approaches to Terminology **65**
3.1 Introduction 65
3.2 Term Recognition 65
3.2.1 Term recognition and term interpretation 66
3.2.2 Approaches using linguistic information 71
3.2.3 Approaches using statistical information 73
3.2.4 Hybrid techniques 76
3.2.5 Lexical atoms and collocations 77
3.3 Recognition of Term Variants 78
3.4 Disambiguation Techniques 80

3.4.1 Levels of ambiguity 80
3.4.2 Approaches to disambiguation 82
3.4.3 Using semantic information 83
3.4.4 Acquiring semantic knowledge 84
3.4.5 Word sense and term sense disambiguation 85
3.5 Summary 85

4 Extracting Contextual Information 87
4.1 Use of Contextual Information 88
4.1.1 Terminography 88
4.1.2 Term extraction 88
4.1.3 Semantic information 89
4.2 Identifying Relevant Contextual Information 93
4.2.1 Syntactic information 96
4.2.2 Terminological information 97
4.2.3 Semantic information 98
4.3 Obtaining Contextual Information 99
4.3.1 Syntactic weight 99
4.3.2 Context term weight 103
4.3.3 Similarity weight 104
4.4 Summary 105

5 Measuring Similarity 106
5.1 Introduction 106
5.2 Methods for Extracting Semantic Information 106
5.2.1 Corpus-based approaches 107
5.2.2 Dictionary-based approaches 108
5.3 Ontologies, Dictionaries and Thesauri 109

5.3.1 Theoretical notions of classification systems 109
5.3.2 General vs. domain-specific ontologies 112
5.3.3 Lexicalisation 115
5.3.4 General language resources 116
5.3.5 Specialised resources 118
5.4 Using Semantic Information from UMLS 122
5.5 Methods for Measuring Similarity 128
5.5.1 EBMT-based methods 129
5.5.2 IR-based methods 131
5.5.3 Thesaurus-based methods 133
5.5.4 Summary of methods 136
5.6 Similarity Measurement in the TRUCKS System 137
5.6.1 The semantic hierarchy 140
5.6.2 An example 141
5.7 Using Similarity for Term Disambiguation 142
5.7.1 Disambiguation of candidate terms 143
5.7.2 Some examples 144
5.7.3 Disambiguation of context terms 146
5.8 Conclusion 147

6 The SNC-Value 149

6.1 Introduction 149
6.2 The Statistical Approach: C-Value 149
6.2.1 The linguistic module 150
6.2.2 The statistical module 151
6.3 Adding Contextual Information: the NC-Value 153
6.4 Adding Deeper Information: the SNC-Value 155
6.4.1 The Importance Weight 156

6.4.2 Calculating the SNC-Value 157
6.5 Applying the SNC-Value to Term Extraction 157
6.5.1 The data 158
6.5.2 Stage 1: the C-Value 158
6.5.3 Stage 2: the NC-Value 160
6.5.4 Stage 3: the SNC-Value 162
6.6 Summary 173

7 Clustering Contexts **174**
7.1 Introduction 174
7.2 Aims of Context Clustering 174
7.2.1 Disambiguation 175
7.2.2 Knowledge acquisition 176
7.2.3 Tuning of ontologies 177
7.3 Approaches to Clustering 178
7.3.1 Creating semantic frames 178
7.3.2 Lexical vs. structural pattern matching 180
7.3.3 Information formatting in the medical field 183
7.4 Clustering Contexts Using Semantic Frames 185
7.4.1 Acquiring syntactic information 185
7.4.2 Identifying syntactic patterns 189
7.4.3 Identifying semantic patterns 192
7.4.4 Mapping from syntax to semantics 196
7.4.5 Creating semantic frames 199
7.4.6 An example 201
7.5 Discussion 202
7.5.1 Clustering errors 203
7.5.2 Granularity of clustering 206

7.5.3 Further expansion 206

8 Evaluation **208**

8.1 Evaluation of ATR 208

8.1.1 Comparison of ATR methods 209

8.1.2 Precision and recall 210

8.1.3 Manual evaluation of terms 211

8.1.4 Evaluating terms against the UMLS Metathesaurus 212

8.1.5 Comparing the manual evaluation with the UMLS 213

8.1.6 Comparison of the NC-Value and SNC-Value rankings against the UMLS 215

8.2 Similarity Weight 217

8.2.1 Expected theoretical performance 217

8.2.2 Comparison with statistical methods of measuring similarity 219

8.2.3 Practical performance 223

8.3 The Context Term Weight 224

8.3.1 Comparing context terms with context words 224

8.3.2 The distribution of context terms in the corpus 227

8.4 Syntactic Weight 228

8.4.1 Average syntactic weights for terms and non-terms 228

8.4.2 Evaluation of the addition of the syntactic weight to the SNC-Value 229

8.5 Overall Evaluation of the SNC-Value 233

8.5.1 Comparing results: upward and downward trends 233

8.5.2 Precision of the SNC and NC-Values 235

8.5.3 Accumulative precision of the SNC and NC-Values 237

8.5.4 Detailed precision at the top of the list 238

8.5.5 Accumulative precision of the top 250 terms 240

8.5.6 Distribution 241
8.5.7 Distribution of the top 250 terms 242
8.6 Conclusions 243

9 Conclusions **245**
9.1 Summary 245
9.2 Further Work 247

A Stop List for Term Extraction **251**

B UMLS Semantic Hierarchy **260**

C Frequency List for Verbs **266**

D Part-of-Speech Tags **270**

E List of SuperTags **272**

F Variations of the term *epilepsy* **286**

G Results of SNC-Value **295**

List of Tables

2.1 Examples of the composition of multi-word medical terms 40

2.2 Types of variants of the term *enzyme activity* 58

4.1 Distribution of boundary words by category for terms and non-terms 101

5.1 Advantages and disadvantages of corpus-based approaches for acquiring semantic information 108

5.2 Advantages and disadvantages of dictionary-based approaches to semantic knowledge acquisition 109

5.3 Relationship between *auricular fibrillation* and *atrial fibrillation* in the UMLS 128

5.4 Similarity between candidate terms and ambiguous context terms 146

6.1 Example of context words and weights occurring with the term *proliferative_retinopathy* 161

6.2 Syntactic weights for boundary words 165

6.3 Syntactic weights for *descemet's membrane* 166

6.4 Syntactic weights for *macular dystrophy* 166

6.5 Syntactic weights for *acanthotic hyperkeratotic* 166

6.6 Top 20 results for the SNC-Value and NC-Value 172

7.1 Example of an information format for the class *Sign or Symptom* 179

7.2 Example of Rocha et al.'s Event Definition model 180

7.3 Examples of syntactic frames 192
7.4 Distribution of verbs appearing in syntactic frames 193
7.5 Distribution of most frequent verbs in syntactic frames 193
7.6 Verb classes 194
7.7 Noun classes 195
7.8 Strings matching the pattern V(act)<obs> N Prep N 197
7.9 Mappings from syntactic to semantic patterns 198
7.10 Examples of semantic frames with fillers 200
7.11 Contexts and their normalised forms 203
7.12 Examples of final clustered contexts 204

8.1 Comparison of manual evaluation and UMLS 213
8.2 Evaluation of NC-Value extracted terms compared with UMLS terms 215
8.3 Evaluation of NC-Value extracted terms compared with manually judged terms 215
8.4 Manually evaluated and UMLS-evaluated terms per list section . . 216
8.5 Evaluation of SNC-Value extracted terms compared with UMLS terms 217
8.6 Evaluation of SNC-Value extracted terms compared with manually judged terms 217
8.7 Semantic weights of terms and non-terms 218
8.8 Experiment 1: semantic weights of terms and non-terms 222
8.9 Experiment 2: semantic weights of terms and non-terms 222
8.10 Evaluation of context words relating to terms and non-terms . . . 225
8.11 Evaluation of context terms relating to terms and non-terms . . . 225
8.12 Occurrences of context terms with terms and non-terms 227
8.13 Occurrences of context terms with terms from each part of the list 228
8.14 Comparison of average syntactic weight for terms and non-terms . 229

8.15 SNC-Value with the syntactic weight 230
8.16 SNC-Value without the syntactic weight 230
8.17 Precision of SNC-Value with and without syntactic weight added . 231
8.18 Precision of top 250 terms with and without the syntactic weight 232
8.19 Precision of SNC-Value and NC-Value 236
8.20 Accumulative precision of SNC-Value and NC-Value 237
8.21 Precision of SNC-Value and NC-Value for top 250 terms 239
8.22 Accumulative precision of SNC-Value and NC-Value for top 250 terms 240
8.23 Distribution of SNC-Value and NC-Value 242
8.24 Distribution of SNC-Value and NC-Value for top 250 terms 242

List of Figures

2.1 Ideal relation between terms and concepts 25
2.2 Relation between ambiguous terms and concepts 25
2.3 Relation between synonymous terms and concepts 25
2.4 Actual relation between terms and concepts 26
2.5 Relation between term, context and concept 29
2.6 Example of multi-inheritance in a network 50
2.7 Multi-inheritance represented in set theory 51
2.8 Section of a combinatory system (taken from (Felber, 1984)) . . . 52

5.1 Lawson's primary concepts for the field of patenting 114
5.2 UMLS Semantic Network's primary concepts for the field of medicine 114
5.3 Fragment of the UMLS Semantic Network 120
5.4 The Ogden-Richards meaning triangle (taken from the original 1923 illustration) 124
5.5 Understanding the semantics of "aspirin" using the UMLS and the Ogden-Richards framework (adapted from (Campbell et al., 1998)) 125
5.6 Relations between strings, terms and concepts in the UMLS . . . 126
5.7 Ambiguity and variation in the UMLS 127
5.8 Fragment of a food network 138
5.9 Fragment of the Semantic Network 139

8.1 Influence of statistical methods on the similarity measurement . . 223

8.2 Comparison of context words and context terms 226
8.3 Precision of SNC-Value with and without syntactic weight 231
8.4 Precision of top 250 terms with and without the syntactic weight 232
8.5 Examples of "perfect" results for precision 233
8.6 Examples of upward and downward trends 234
8.7 Precision of SNC-Value and NC-Value 236
8.8 Accumulative precision of SNC-Value and NC-Value 238
8.9 Precision of SNC-Value and NC-Value for top 250 terms 239
8.10 Accumulative precision of SNC-Value and NC-Value for top 250 terms 241

Copyright

Abstract

Increasing volumes of electronic text have given rise to a need for new methods of collecting, organising, storing and retrieving terminological data. Coupled with the laborious nature of such tasks, and the constant dynamism of technical fields, applications such as term extraction and disambiguation benefit greatly from automation. Methods for automating terminological procedures have traditionally been statistical, but these are not always sufficiently accurate, particularly for smaller corpora. Recently, hybrid approaches incorporating linguistic information have been developed, but these still rely largely on statistical information and use only very shallow syntactic knowledge. We examine the use of linguistic techniques in the area of automatic term recognition. We develop the TRUCKS system, which makes use of different types of contextual information - syntactic, semantic, terminological and statistical - seeking particularly to identify those parts of the context which are most relevant to terms. From an initial corpus of sublanguage texts, this identifies, disambiguates and ranks candidate terms. The system is evaluated with respect to the statistical approach on which it is built, and with respect to its expected theoretical performance. We show that by using deeper forms of contextual information, we can improve on the extraction of multi-word terms.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Acknowledgements

I would like to thank Dr. Sophia Ananiadou for all her help and advice; Prof. Michael Fisher for his support and comments on this thesis; Peter Quigley for his comments.

I am grateful to the many other people with whom I have had helpful discussions, especially Dr. Katerina Frantzi, Prof. Jun-ichi Tsujii and Dr. Hideki Mima; and my colleagues at MMU, particularly Dr. Clare Dixon for all her advice and help.

I thank Dr. Katerina Frantzi for allowing me access to her programs for the NC-Value, and Prof. Aravind Joshi and Dr. Srinivas Bangalore for allowing me access to their Supertagger. I could also not have achieved this research without the technological help of the Manchester Library Visually Impaired Persons Unit, and I am very grateful to its staff for all their support.

Personal thanks must also go to numerous people, but especially to Paul Silverwood, Chris Martin, Nicola Hunter, Ian Radford and John O'Neill for their constant support and encouragement, and for helping me keep things in perspective.

Chapter 1

Introduction

In recent years, there has been an increasing demand for advances in the establishment and management of terminological resources. This is largely caused by the vast quantities of new electronic material appearing. In particular, new technical terms are constantly emerging, while others are undergoing changes in meaning. This means that the need for automatic methods of processing terminology is becoming more evident.

The growing availability of large-scale electronic data is also leading to new approaches in computational terminology. Corpus-based methods are becoming more widespread in areas such as term recognition, although attention has been largely focused on statistical methods. The increasing development of electronic lexical resources, coupled with methods for automatically creating and fine-tuning them from corpora, has paved the way for a more dominant appearance of natural language processing (NLP) techniques in the field of terminology.

The use of NLP in terminology has several advantages. First, it is complementary to the use of other methods such as statistical approaches, since the two can be used simultaneously. Second, it can offset problems such as statistical skewing in small or unusual corpora. Third, advantage can be taken of reusable resources, thus making the procedure more efficient. Fourth, it can be used to further our understanding of terminology, which then influences techniques for its processing and manipulation.

1.1 Term Recognition Using Combined Knowledge Sources

Term recognition is an important procedure for many applications, such as machine translation, automatic indexing, hypertext linking, information retrieval and information extraction. Manual term recognition is a time-consuming, labour-intensive task, which is also marred by the subjectivity of the terminologist. Automated procedures for term extraction are therefore in great demand. Existing methods for multi-word term extraction are largely statistical. Where linguistic information is used, it tends to be very shallow, consisting of little more than a filter restricting the syntactic categories of the terms extracted.

The main objective of this thesis is *to investigate ways in which methods for automatic multi-word term recognition can be enhanced by the addition of linguistic and contextual information.* To this end, we have five main aims.

1. To propose a method for extracting relevant contextual information from specialised texts.

2. To enhance existing statistical methods of term recognition with linguistic knowledge.

3. To disambiguate candidate terms and context words.

4. To investigate ways in which contextual information might be used for other terminological procedures.

5. To evaluate the effect of adding deeper forms of linguistic information to a system for term recognition.

The system for multi-word term extraction described in this thesis is called TRUCKS. Not only is this an acronym for Term Recognition Using Combined Knowledge Sources, but it also evokes the structure of the system. This can be viewed as an engine pulling a set of trucks, each containing a particular type of knowledge (such as semantic, terminological, statistical and contextual). It is knowledge-rich in that it combines information from different sources: an external thesaurus and semantic network providing semantic and classificatory information, and the corpus providing statistical and syntactic information. The "trucks", or modules, are semi-independent in that each can be altered without affecting the others, but each plays a role in the assignment of weights which are used to determine the status of the candidate terms and their ranking.

The TRUCKS system is based on the NC-Value approach to term extraction (Frantzi, 1998; Frantzi and Ananiadou, 1999). Although the NC-Value is a hybrid method, incorporating both statistics and linguistics, the linguistic information used is very shallow. The NC-Value is extended to form the SNC-Value weighting system for extracted terms, which takes into account much more detailed contextual information, by calculating an Importance Weight for each candidate term. The semantic component of this is also used to perform term and context disambiguation, and to assist in clustering procedures.

The system is tested on a corpus of texts in the domain of eye pathology. From this, it identifies, disambiguates and ranks candidate terms. Evaluation shows improvements on the NC-Value method in terms of precision and distribution. The results also show that there is much potential for the future development of knowledge-intensive methods in terminology.

1.2 Organisation of this Work

This thesis can be divided into three main parts: the first, consisting of chapters 1–3, provides background information; the second, consisting of chapters 4–6, presents the details of our approach; and the final part, chapters 7–9, discusses its practical usage, evaluation and extensions.

The following two chapters therefore describe the background to this work, and the fundamental principles on which it is based. Chapter 2 investigates some of the most important theoretical foundations to terminology and term extraction. In particular, it draws together knowledge from a variety of fields such as linguistics, psychology, terminology and natural language processing. Chapter 3 then focuses on specific applications of these, with respect to term extraction, term disambiguation and term variation.

The second part describes the main body of the work, the TRUCKS system. Chapter 4 discusses the use of contextual information in terminology, and the different knowledge sources which we draw on to produce a set of context weights. Chapter 5 provides a description of the similarity measure we use for the semantic component of the system, and of the method we use for disambiguation. In Chapter 6, we describe how the context weights are combined to form the contextual Importance Weight, and how the SNC-Value is implemented.

The final part focuses on the evaluation, applications and extensions to our work. Chapter 7 investigates a proposed technique for clustering contexts, using the information we have already obtained, and shows how both the SNC-Value and other terminological applications might be improved. Chapter 8 provides an evaluation of the SNC-Value, by comparing it with the NC-Value. This enables us to measure the significance of incorporating deeper forms of information into a largely statistical measure. Finally, Chapter 9 summarises our work and suggests further improvements and extensions.

Chapter 2

Theoretical Terminology

2.1 Overview

In this chapter, we introduce some theoretical foundations for the field of terminology. In particular, we focus on the relationship between terms and concepts, which is of essential importance to this field both theoretically and practically. To understand something about the nature of terms, we cannot just examine them in a vacuum, but must relate them both to the concepts they represent and to the underlying structures to which they belong. The field of terminology should also not be viewed independently, but considered in conjunction with the fields of linguistics and cognitive science at the very least.

The first section defines and examines the relationship between terms and concepts from a traditional terminological / linguistic standpoint. We look at terms both intrinsically and extrinsically, taking into consideration the contexts in which they occur and how this information should be incorporated into our theory. The second section takes a psychological viewpoint and discusses the psycholinguistic aspects of terms and concepts – in particular how they are seen and recognised by language acquirers and speakers. Section 3 takes on a linguistic aspect and explains some of the basic principles of term formation, while Section

4 examines the theoretical side of the problem of term ambiguity, focusing on how ambiguous terms and concepts are related and on the different types of ambiguity. Methods for resolving these problems are dealt with in Chapter 3. The fifth section looks at a related problem at the opposite end of the scale, namely term variation. In particular, it examines how and why term variants are created and the various forms that they take. The description of practical solutions to their recognition is again left until Chapter 3.

2.2 Terms and Concepts

Any practical application in the field of terminology must have roots in some theoretical principles concerning the notions of term and concept. There are numerous definitions of a technical term, but all agree that it is essentially related to some kind of surface realisation of a specialised concept. The International Organization for Standardization (ISO) defines a *term* as the

> " ... designation of a defined concept in a special language by a linguistic expression",

and *concepts* as

> " ... mental constructs serving to classify individual objects of the inner or outer world by way of more or less arbitrary abstraction." (ISO-704, 1987)

According to Bourigault, the primary function of a term is as an isolated representation of a concept:

> "The terminological unit plays this role of representation in the framework of a terminology which is the linguistic evidence of the organisation of a field of knowledge in the form of a network of concepts;

> the terminological unit represents a concept, uniquely and completely, taken out of any textual context. The existence of this one-to-one relationship between a linguistic expression and an extra-linguistic object is ... a situation which particularly concerns the terminologist." (Bourigault, 1992)

Bourigault claims that it is this relationship between term and concept (that exists outside any linguistic context) that distinguishes a term from a word. Ideally this should be true, but in practice there is rarely a one-to-one relationship between term and concept, as will be discussed later. However, whilst we do not agree that the context serves to define a term, in that the link between term and concept is independent of context (as Bourigault claims), this does not mean that the context has no role to play in studies of terminology. In fact we claim the opposite, that linguistic context is crucial to the understanding of terms and their relationship with corresponding concepts, and this is reinforced by the fact that terms are both homonymous and ambiguous in practice.

2.2.1 Relation between term and concept

According to the ISO guidelines, the relationship between term and concept is through *definition*.

> "Terms are linked to concepts by means of definitions. The definiendum is a term and the definition may contain other terms. Terms defined by different definitions in different subject fields are called homonyms. Terms which represent the same concept with the same definition are called synonyms." (ISO-704, 1987)

In theory, terms should be monoreferential, i.e. there should be a one-to-one relationship between term and concept (Figure 2.1). In practice, however, terms

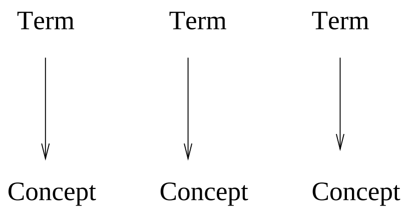

Figure 2.1: Ideal relation between terms and concepts

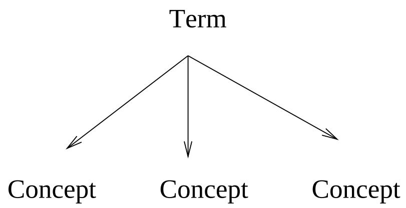

Figure 2.2: Relation between ambiguous terms and concepts

exhibit both ambiguity (Figure 2.2) and synonymy (Figure 2.3). The overall picture is therefore more likely to be some combination of the two, i.e. a many-to-many relationship (Figure 2.4).

The literature cites many different opinions regarding the "essential characteristics" of a term, such as "monosemous", "precise", "capable of forming derivations", "systematic", etc. (Schulze, 1987). However, Schulze points out that most of these are actually redundant since they follow either directly or indirectly from the basic definition of a term, and are not therefore primary or defining characteristics, but rather incidental attributes. He defines a term as

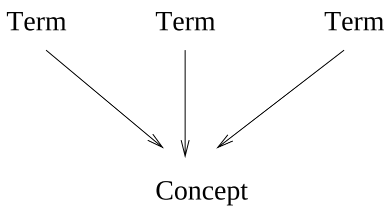

Figure 2.3: Relation between synonymous terms and concepts

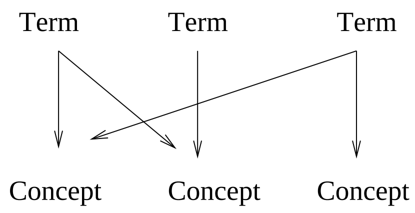

Figure 2.4: Actual relation between terms and concepts

> "a lexeme the meaning of which consists of the essential characteristics of a concept and which expresses (designates) this concept".

He does not mention here anything about the specialised nature of terms or concepts, although he defines the essential characteristic of a term to be the "name (designation, expression form, reference etc.) of a special subject/scientific/technical concept".

From a philosophical point of view, it is generally agreed that what distinguishes a term from a word, or from other linguistic "signs" is that its "semantic extension is defined in relation to the concept rather than the name" (Rondeau, 1987). In other words, whilst lexicography is interested in the meaning of a lexical unit, terminology is interested in the concept represented by the lexical unit. Rondeau also points out that a concept is always defined in relation to other concepts within a particular field, so that terms should not be considered in isolation but as part of a larger group. Whilst words can also be classified in hierarchies or ontologies, this is not crucial to their meaning in the same way as it is with terms.

It is unsurprising that the correct association of term and concept is important for machine translation (MT) (Dubuc and Lauriston, 1997), but this is also true for almost any application of computational linguistics involving specialised languages. Translation of technical texts involves finding a suitable equivalence between the source language (SL) term and the target language (TL) term. In order to establish this equivalence, Dubuc and Lauriston claim that equivalence between the SL and TL concept must first be found (Dubuc and Lauriston, 1997). For MT, the problem generally lies even deeper, since the SL and TL concepts may not have a direct correspondence. So, although translation appears to be simply a matter of establishing a direct term-term correspondence, it is not always quite so simple. Cognitive issues also have to be considered; a concept is not fixed in stone but can vary according to the viewpoint of an individual or a group of people. So establishing the correspondence between SL and TL concepts is not always straightforward.

Whilst we are not aiming to perform machine translation in this work, similar principles apply. We must have some comprehension of the relationship between term and concept in order to understand term variation and ambiguity, both of which, we believe, are crucial to the theoretical study of terminology and to its practical applications. In order to perform term disambiguation, we need to be able to relate different occurrences of a term to different concepts, and in order to be able to account for term variation we need to be able to relate different linguistic realisations of a term to one concept. This is virtually impossible unless we already have a clear idea of what these concepts are.

2.2.2 Terms and termforms

At this point, it is necessary to introduce some clarifications as to the terminology we use, in order to avoid confusion about terms and concepts. Lauriston (Lauriston, 1995) makes an important distinction between *terms* and *termforms*, which we have so far ignored. He describes a termform as the actual lexical form or linguistic entity encountered in a text, and a term as "the intersection between a conceptual realm … and a linguistic realm", i.e. something that lies between the lexical form and the concept to which it is related. Frantzi (Frantzi, 1998) prefers to consider a term as an ordered pair consisting of a concept and a termform, in other words a superset of the two sets rather than an intersection between them. So theoretically when we talk about *terms*, we are not talking about the linguistic expression itself but something wider. However, since *term* is commonly used to denote the linguistic realisation, we shall continue to use this designation for the sake of simplicity, except in cases where we make an explicit distinction between Lauriston's terms and termforms.

2.2.3 Contexts

The overall picture cannot be completed without mentioning the linguistic contexts we are investigating, and how they should be viewed. They can be seen as related primarily to either terms or concepts (Dubuc and Lauriston, 1997). The psychological viewpoint sees them as being related to concepts, in that the context serves to position the concept within a particular subject field. We view contexts as being the key to understanding a term, because a term can be seen in different ways according to each individual instance. We can only be sure that two instances of a term are being used in the same way if we consider the context in which each instance is situated. So while the linguistic realisation may remain the same, the meaning may change (albeit almost imperceptibly), and we therefore wish to consider the relationship between meaning and context, rather than considering simply the linguistic forms of the two. It is thereby the concept rather than the term which appears to be most relevant to the context. In this way, the concept can be seen as a kind of *bridge* between context and term, as shown in Figure 2.5. Since we aim to illuminate the relationships not only between context and concept, but also between concept and term, an understanding of concepts is crucial.

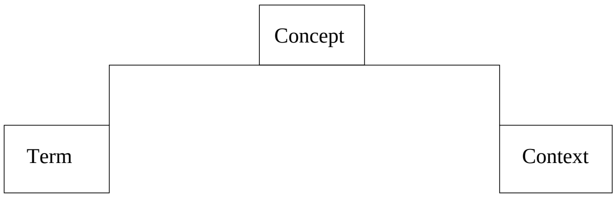

Figure 2.5: Relation between term, context and concept

The alternative approach views the context from a structural position, regarding the context as a source of collocational information and syntactic structures. In this case, the context is linked to the term itself rather than to its underlying meaning. Although it is undeniable that contexts do serve this purpose, we consider this to be an incidental rather than a primary function.

2.3 The Cognitive Dimension

In order to understand more fully the relationship between concepts and terms, we should not neglect the psycholinguistic aspect. Whilst we are not aiming for an automatic means of understanding terminology and its underlying concepts in the same way that people understand it, it may nevertheless be helpful to examine the way in which people, and in particular children acquiring language, develop an understanding of some of the underlying concepts of terminology, and build relationships between terms and concepts.

As children learning to speak, we do not have any real innate comprehension of concepts, but gradually acquire this knowledge as we acquire language, which in turn helps us to understand and use language. Aitchison (Aitchison, 1994) identifies three separate tasks in a child's acquisition of meaning:

- labelling
- packaging

- network building.

2.3.1 Labelling

Although it might seem as if children can correctly match names with concepts very early on, it is more likely that their concept systems are actually very d-ifferent from those of an adult. For example, they might say "horse" on seeing a horse on its own in a field, thereby appearing to form a correct association between the animal and its name. But they might not say it on seeing the same horse being ridden down the road. To them, this is a different concept which, as yet, they do not know how to name. Aitchison notes that early on, children tend to associate a name with not only one particular object, but also with that object in a particular situation. It is often only later that they become capable of associating the name with the object in different situations, and finally with different objects in different situations (e.g. recognising that a toy horse and a real horse are both horses).

Although NLP is not particularly concerned with the process of naming objects that are being seen, the principles are very similar. Instead we are faced with the problem of matching occurrences of lexical forms to terms. The problem for a machine is that once it has recognised a lexical form as being related to a particular term, it may form an association between these two which is hard to break. Similarly, we need to be able to recognise whether a word occurring in different contexts (or even the same context) refers to the same concept or not. Take, for example, the following sentences.

1. The dog growled at the cat.

2. The dog and the bitch are fighting.

3. The dog and the cat are playing.

The problem is to decide whether the word "dog" in each of these three sentences refers to the same concept in each case. In the first sentence the word "dog" refers to the species of animal, whereas in the second sentence it refers to the gender of the animal (which may or may not belong to the same species of animal as in the first sentence – for example both "dog" and "bitch" could be referring to foxes or wolves). Even assuming that we now know there are two meanings of the word "dog", in the third sentence we still have to decide which of the two meanings is being referred to. Given the above contexts, this is not an easy task. On syntactic grounds there is a closer relationship between sentences 2 and 3, because both 2 and 3 have the structure *the X and the Y are Z*, but on semantic grounds there is a closer relationship between sentences 1 and 3, because both 1 and 3 refer to a dog in connection with a cat.

In essence, this is the same problem that the child has. Given different situations, how do we tell whether the objects that appear in these situations are the same? As demonstrated by our examples, it is not always simply a case of comparing the similarity of the contexts (either the visual contexts for the child or the written contexts for the machine). For the child, it is easier since there are other available sources of information to be drawn on, such as pragmatic and world knowledge, which are hard to include in any automatic recognition system.

2.3.2 Packaging

The labelling problem is closely linked with the packaging problem (the ability to apply a particular name to the correct set of instances of an object). Clark describes the acquisition of semantic knowledge as

> "adding ... to the lexical entry of the word until the child's ... entry for that word corresponds to the adult's." (Clark, 1973)

Initially, a child's view of the world is essentially monoreferential and, as it acquires language, it begins to learn both that meanings can be extended and that objects can have more than one name. For the machine, the problem is also that of grouping together different words as meaning the same thing, or of correctly subdividing a hypernym into its relevant hyponyms.

Some of these forms may outwardly look similar but may actually be related to different concepts, in the same way that a tomato may appear to be a vegetable but is actually a fruit. By analogy and experience, a child assumes the former until it is specifically told the latter. In the same way, a machine may have to be specifically told about certain relationships which it cannot predict based on prior knowledge.

Both children and machines also tend to focus on superficial qualities of objects. For example, the fact that a moon is crescent-shaped is generally one of its defining characteristics, but not the fact that it changes shape. Since the moon is not always crescent-shaped, this is rather a superficial quality, but ask a child to draw a picture of a moon, and unless specifically asked to draw a full moon, the chances are he or she will draw a crescent shape. Similarly, to a child, a robber might be understood as something like "a bad man with a gun who steals things". Someone who breaks into a house to steal drawing pins would probably not be thought of as a robber, although the defining characteristic of a robber has more to do with the stealing element than the gun element.

In the same way, to "understand" a term, a machine must be able to distinguish superficial elements of its meaning from fundamental elements. The fact that a tomato is red is not an essential element of its meaning (since it can be other colours), but the fact that it is a fruit is essential, since this determines its relationship to other fruits and vegetables. However, there is a certain discrepancy between the defining characteristics of a term and those which are useful.

Knowing that it is a fruit rather than a vegetable is not particularly useful for us in real life, since tomatoes are eaten in the manner of vegetables rather than fruits. For example, making use of the defining characteristics of a word or term for structural disambiguation, e.g. (Vanderwende, 1995), is a technique which can run into problems if the definition does not correspond with the actual usage of the word or term.

2.3.3 Network building

The third stage in a child's acquisition of meaning is what Aitchison calls "network building", i.e. the ability to form networks linking objects together. Children initially find it hard to understand concepts such as hyponymy, for example they cannot grasp how an object can be both a horse and an animal at the same time. They have discovered that objects have names, and that these names can refer to different objects, but they have not grasped the concept of naming larger (or smaller) groups of objects. In computational applications, various types of hierarchy such as thesauri and ontologies may be used to provide this relational information about words. As will be discussed more fully in the following chapters, there are many problems inherent both in using pre-defined hierarchies (either with accessing or making use of the relevant information, or with the design of the hierarchy) and in creating or fine-tuning such a hierarchy automatically.

Although we have only highlighted some of the most important points, we can see that there is a clear connection between language acquisition for children and the understanding of language in some form by NLP systems. As humans we tend to be unaware of many of the cognitive processes that are involved in the processing of language, since for a large part they are subconscious. Recognising and understanding some of the problems faced by children acquiring language may help us to appreciate the difficulties faced by NLP systems, particularly in the area of terminology, where the relationship between terms and concepts, and the organisation of hierarchies is so crucial.

2.3.4 Recognising words and terms

The process of recognising terms is not a simple activity even for humans (which is partly the reason why automatic methods for performing this activity are so desirable). Clearly the mechanisms involved are not the same as those used to identify words. When we hear or see a lexical item, it is relatively easy for us (as adults) to decide if it is a word or not, within a fraction of a second and without any conscious thought. This involves decision criteria such as recognising "impossible" sequences of letters in a word (for a particular language). Word recognition is also largely about memory. If faced with an item we have seen before, we are more likely to assume it is a real word. This is proved by the fact that if faced with a non-word we have seen many times before (even if it is known to be a non-word) we begin to question whether in fact it is a real word.

While term recognition for humans naturally incorporates a certain amount of memory, it is not that simple for several reasons.

1. Termhood may vary amongst domains. An item may be a term in one domain but not in another. So though we might recognise it as having been a term before, if there is a change of domain it may not still be a term in the new domain.

2. The type of memory required is different for word and term recognition. It is not simply a case of recognising an item as "known" or "unknown" to us, for we may have come across the item without being aware as to whether it was a term or not. In the case of words, we are unlikely to have come across an item unless it was a real word, but we will naturally come across many items which are words but not terms.

3. Word recognition is almost always an artificial situation, whereas term recognition is a task which needs to be carried out in real life. Experiments with word recognition are simulated tasks, but although term recognition experiments may also be simulated, there is little difference between these experiments and the real life task.

4. Term recognition involves other factors apart from memory, to a greater extent than word recognition. When faced with an item we have seen dozens of times before, we are likely to interpret it as a word (based on the fact that we do not see many occurrences of non-words). So frequency is a good indicator of wordhood. Although frequency does play an important role in termhood (as demonstrated by the large number of automatic term recognition systems which take frequency of occurrence into account), it has less significance for terms because many words appear very frequently which are not terms, even in specialised texts. This is generally dealt with in automatic systems by the establishment of a stop list of frequent words which are excluded from consideration as candidate terms.

5. Term recognition requires conscious thought, whereas word recognition does not appear to do so. This is the main reason why we can recognise words more or less instantly, but we take longer to recognise terms (unless we have already identified them many times before).

6. Term recognition is a subjective process which is much less clear cut than word recognition. Word recognition is a binary distinction which can in most cases be verified by an appropriate source such as a dictionary. Term recognition is often scalar and is open to individual variation – as with translation there is not necessarily a single correct solution. This partly accounts for why some automatic term recognition systems rank candidate

terms rather than, or in addition to, awarding them a straight yes/no result. There is no real means of ranking words in a similar way.

Although many psycholinguists have studied the topic of word and non-word recognition in humans, e.g. (Coltheart et al., 1979), (Gardiner and Java, 1990), (McClelland and Rumelhart, 1981), there have not been similar studies for terms, largely because little research has been carried out intersecting the fields of terminology and psycholinguistics. This kind of study would, we believe, be beneficial to both fields, although the limitations of this thesis prevent us from carrying out such work here.

2.4 Term Formation

Understanding the theoretical issues involved in term formation is an important aspect of term recognition. Since it has been discussed in great detail elsewhere, e.g. (Sager, 1990; Rey, 1979; Lauriston, 1996; Sager, 1997), we shall restrict ourselves to summarising the most salient points.

Sager (Sager, 1997) differentiates between term and word formation in that term formation is a planned, conscious effort which takes into account pre-existing patterns, whereas word formation is more arbitrary. However, this does not mean that term formation is always predictable, because there are many different circumstances in which term formation can occur. For example, when something is discovered by chance, its naming procedure is often very ad hoc.

Sager describes three major patterns of primary term formation:

- use of existing resources;
- modification of existing resources;
- creation of new linguistic entities.

Within these patterns there can, of course, be many subdivisions, but we shall only give a brief overview here.

2.4.1 Use of existing resources

When a new concept arises, an existing term may have its meaning extended to cover it. For example, the term "computer" has changed its meaning over the years to cover a much broader concept than its original meaning. This pattern includes the frequent method of inter-domain borrowing, where a term from general language or a specific domain is used to denote a different concept in another domain. For example, the term *mouse* used in computing is borrowed from the animal meaning of the term in general language, because of the visual similarities between the two objects.

2.4.2 Modification of existing resources

Many terms are created through morphological or syntactic transformations such as derivation, affixation and compounding (e.g. *visual display unit*, *disambiguation*). There is much debate among terminologists as to whether lexical variants should be classified as individual terms, treated as variations of the same term, or excluded from being classed as terms at all. We shall not discuss this here as it is dealt with more fully in Sections 2.6 and 2.7. Suffice it to say that, in those cases, the lexical variants largely relate to the same concept but differ slightly in their surface realisation, whilst here we are more concerned with variations in surface realisation which relate to different concepts.

2.4.3 Creation of new linguistic entities

New linguistic entities, or neologisms, tend to be created in response to the demand for the naming of a new concept. Although there is much philosophical debate about the existence of linguistically unrealised concepts, it is generally accepted that concepts can exist without being realised by *terms* (Lauriston, 1996). Neologisms tend to be created either by borrowing from other languages, e.g. *ohm*; by using the name of the person who developed or discovered them (known as *eponyms*), e.g. *Bowman's membrane*; or by inventing totally new forms, e.g. *cyberspace*[1]. Rondeau (Rondeau, 1987) distinguishes between *original* and *auxiliary* neonymy (creation of new terms). Terms created through the naming of a new concept by its creator in his working language are said to be original, whilst terms arising from the renaming of concepts in a different language from its original one are said to be auxiliary.

2.4.4 Primary and secondary term formation

Sager also differentiates between primary and secondary term formation (Sager, 1997). Primary term formation is the creation of a new term for a new concept. Secondary term formation is the process of creating a new term for a prior concept. This can be either planned (change of designation) or accidental (transfer of knowledge). In the first case, revision of a terminology (within a domain) may result in a change of designation of some of its concepts. In the second case, when specialised knowledge is transferred across domains, new terms are necessary.

2.4.5 Summary of term formation

Although term formation tends to be a well-structured process, following one of a variety of set paths, it can nevertheless be hard to deconstruct. This is partly because the methods by which terms can be formed are so different, and partly because there are no rules about which method of formation should be used in any instance. Some terminological domains have more clear-cut approaches

[1] coined by William Gibson in his science fiction novels of the 1980s

than others. For example, in the field of immunology, Ananiadou (Ananiadou, 1994) found it possible to recognise and analyse neoclassical terms, because the formation process adhered to strict rules. In many other domains, however, and in particular non-medical ones, term formation is much less regular. Many domains are also more dynamic, and terminologists and linguists cannot rely on finding classical term formation processes, either because new terms are being formed (by borrowing, invention, eponymy etc.), or because old terms undergo changes in meaning. It is important, however, to understand the processes behind creating terms in order to further our knowledge of their structure and use.

2.5 Multi-word Terms

So far we have focused largely on the theoretical aspects of single-word, or *unitary* terms. Since compound, or multi-word, terms are often composed of at least one *nested* single-word term, these theoretical aspects are all relevant. However, the majority of terms in use are multi-word, and there are some relevant differences which must be taken into account for term extraction. For example, the tagging of multi-word terms is by no means a straightforward task, so part-of-speech taggers are usually forced to tag individual components of a term, which is not always ideal.

2.5.1 Formation of multi-word terms

Multi-word terms are not necessarily composed of just terms, but may include any number of terms and non-terms in any order. It is extremely rare for them to exceed eight words in length, and the vast majority are far shorter. Term length and composition tends to vary according to the domain (Boguraev and Kennedy, 1997). Also, terminologists do not always agree on what constitutes a

Composition	Example
non-term + non-term	pigment granules
non-term + term	background retinopathy
term + non-term	vitreous remnant
term + term	vitreous aspirate
modifier + modifier + non-term	central fibrous core
modifier + modifier + term	anterior peripheral synechiae
modifier + binary term	cellular fibrous_tissue
binary term + term	spindle_cell melanoma
binary term + binary term	spindle_cell malignant_melanoma
ternary term + term	central_retinal_vein occlusion
modifier + ternary term	ulcerated basal_cell_carcinoma

Table 2.1: Examples of the composition of multi-word medical terms

term. For example, some would argue that terms can only be composed of nouns and adjectives; others would permit the use of prepositions and even verbs. We principally take the view in this work that combinations of nouns, adjectives and prepositions can form terms, but not verbs, as expressed by the linguistic filter for the NC-Value described in Chapter 6. We are willing to admit that there may be occasions when verbs could be considered terms, but for the purposes of this work, we do not implement this idea. Table 2.1 identifies some examples of different types of composition of multi-word terms in the medical domain. Note that we do not specify whether a modifier is a term or not, though it will frequently be an adjective which is not generally considered to be a term in its own right.

2.5.2 Interpretation of multi-word terms

An important part of term formation, and of understanding terms, is the way in which the terms are put together. With binary terms (terms composed of two units), it is generally easy to see that one unit modifies the other. With longer terms, it is not always so evident how the units are associated. Not only is this a problem for understanding how the term is formed, but it also makes comprehension of the term harder, or even impossible, by anyone or anything except a domain expert. We shall discuss the problem (and necessity) of term interpretation in more detail in Chapter 3.

2.6 Ambiguity in Terms

As we have already discussed, terms are the embodiment of specialised concepts (Sager, 1990) and, ideally, each term should therefore refer to one specific concept, and each concept should be represented by one specific term. In practice, however, terms exhibit a high degree of both polysemy and synonymy.

The problem of term ambiguity is therefore strongly correlated with the problem of defining the relationship between terms and concepts. Essentially, a term is ambiguous if more than one concept can be represented by a single surface form, i.e. by a term. For example, the term "plant" can refer to various different concepts according to the domain. Webster's Revised Unabridged Dictionary (Merriam-Webster, 1999) differentiates between:

1. A vegetable; an organized living being, generally without feeling and voluntary motion, and having, when complete, a root, stem, and leaves, though consisting sometimes only of a single leafy expansion, or a series of cellules, or even a single cellule.
2. A bush, or young tree; a sapling; hence, a stick or staff. "A plant of stubborn oak." –Dryden.
3. The sole of the foot.
4. The whole machinery and apparatus employed in carrying on a trade or mechanical business; also, sometimes including real estate, and whatever represents investment of capital in the means

of carrying on a business, but not including material worked upon or finished products; as, the plant of a foundry, a mill, or a railroad.

5. A plan; an artifice; a swindle; a trick.
6. (a) An oyster which has been bedded, in distinction from one of natural growth. (b) A young oyster suitable for transplanting.

The first two definitions belong to general language or to the domain of horticulture. The third definition belongs to the domain of anatomy, the fourth to the industrial or commercial domain, the fifth to general language (slang), and the sixth to the domain of zoology. We could also add a further definition:

7. A particular move or tactic in team sports, such as ice hockey, designed to mislead the opposition.

WordNet (Miller et al., 1990) provides yet another definition specific to the domain of performing arts:

8. An actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience.

Conversely, terms are considered synonymous if they both relate to a single concept. For example, *peewit*, *lapwing*, *dastard plover* and *wype* all refer to the same species of bird in zoology, and are freely interchangeable.

The relationship between terms and concepts thus tends to be distinctly multivocal, i.e. a combination of these possibilities, as discussed earlier in Section 2.2.1.

Originally, guidelines for terminology creation and management stated that terms should never define more than one concept. The International Organisation for Standardisation (ISO-704, 1987) states that there should only be one term created for a concept. More recently, it has been acknowledged that this is no longer possible or even necessarily desirable:

> "Modern terminological theory accepts the occurrence of synonymic expressions and variants of terms and rejects the narrowly prescriptive attitude of the past which associated one concept with only one term. It is recognised that one concept can have as many linguistic representations as there are distinct communicative situations which require different linguistic forms." (Sager, 1990)

However, Sager does not admit this possibility arbitrarily, but only where there is a situational need for it, i.e. where a single term is not sufficient to express a concept in practical usage.

In their guidelines for terminology management, Bowman et al. propose the following advice:

> "Describe one concept, and only one, per definition. A term that refers to more than one concept ("has more than one meaning") shall have more than one entry, as long as the other concepts are relevant to the subject field in question." (Bowman et al., 1997)

So they not only admit the possibility of a term referring to concepts in different fields, but also to that of a term referring to more than one concept within a particular domain.

The level of ambiguity can be reduced substantially by restricting the domain to a sublanguage, but ambiguous terms are by no means eliminated, since ambiguity can still exist within a single subdomain. For example, the term *drug* can refer to either an illegal narcotic or to a prescribed medicine. Either or both meanings might equally well appear within a text in the field of medicine. Also, there is nothing to prevent general language meanings of terms from occurring within a specialised text, since the two domains are not mutually exclusive. Furthermore, it is rare that any single domain does not admit the possibility of terms from other domains occurring within it. This means that we cannot always discount all meanings of terms other than those restricted to the domain in question, though the possibility of them occurring is less likely on statistical grounds. According to Cimino (Cimino et al., 1989), we should be able to use the position of a concept within a network to characterise its meaning. However, this is not always of much practical use, since almost all hierarchies still suffer from the problem of multidimensionality, whereby the same concept can be classified in more than one way. We shall discuss this topic in more detail in Section 2.6.1.

2.6.1 Types of ambiguity

There are several different ways in which terms can be ambiguous. We divide them into four main types, which can be classified as follows:

- syntactic;
- semantic;
- componential;
- multidimensional.

Syntactic ambiguity

Syntactic ambiguity occurs in both words and unitary terms, but is generally considered less of a problem in terminology than in general language, due largely to the fact that only nouns are usually considered as terms. If a word has a possible nominal interpretation, that interpretation can be considered as a candidate term and any other interpretations can be ignored. For example, the noun *plant* could be considered as a candidate term and the verb *plant* could be ignored. Since part-of-speech taggers are now reasonably accurate (typically 95-99% accuracy) (Karlsson and Karttunen, 1996), the correct syntactic identification of candidate terms is on the whole not seriously problematic. The only real problem caused by syntactic ambiguity is the chance of over- or under-generation of candidate terms by an extraction system, and the possibility of skewed statistical information, but this is largely alleviated by accurate part-of-speech tagging.

Some terminologists also accept terms other than nominal expressions – most commonly these are verbs or verbal expressions. In this case, the syntactic ambiguity may be less important, since distinguishing between cases of noun/verb ambiguity may not be necessary. Any erroneous syntactic tagging will not alter the outcome of the recognition process. For example, it would not matter whether *plant* was identified as a noun or a verb since in either case it would still be recognised as a term. The fact that noun/verb ambiguity amongst terms is so common also adds fuel to the argument of those who believe terms do not have to be nominal and that syntactic variations of terms are permissible. This point is discussed further in Section 2.7.

Semantic ambiguity

The case of semantic ambiguity is one which is particularly problematic for computational terminologists. It is often disregarded, either because its importance is not recognised, or because it is simply considered too awkward to deal with. Since we adopt a meaning-based approach to term recognition, we must therefore take into account any discrepancies such as semantic ambiguity which may affect our method.

We further subdivide semantic ambiguity into three types:

- homophony;
- homonymy;

- polysemy.

Homophones are words which are pronounced the same but spelt differently, such as *ceiling* and *sealing*. Naturally, homophones are not an issue where term recognition is performed on written texts, but if speech recognition is part of the process then it must be taken into account. Although there are currently few term recognition systems which involve speech, this is likely to change in the foreseeable future with systems such as YPA (deRoeck et al., 1998). However, since this is a problem associated with speech processing rather than with terminology per se, and since we can disregard it for the purposes of our work, we shall not elaborate further here.

Homonyms are words unrelated in meaning or origin which are spelled the same, e.g. *mint* (the herb) and *mint* (the place where currency is coined) have no relation to each other apart from their identical surface forms. Included in this category are **homographs** which are spelled the same and unrelated in meaning but are pronounced differently, e.g. *tear* (the ocular fluid) and *tear* (a rip). Clearly homonymy does constitute a problem for term recognition from written language (although, in the case of homography, not for spoken language), since it could lead to erroneous recognition, either in the form of false positives or omissions. Because the words are unrelated, even if we ignore their meaning it is quite possible that one will be a term and the other will not. Using the restrictions of the domain to filter out alternative meanings may only be partially successful, and may be insufficient to weed out all ambiguities, since two meanings of a word may still exist within one domain.

The third type of semantic ambiguity, **polysemy**, is probably the most widespread and troublesome, partly because of its frequency of occurrence, but also because it is much more difficult both to recognise and to resolve. Polysemous terms are identical in surface form but have different meanings. Unlike homonyms, the meanings are related in some way. Many terms are polysemous because of the way in which they have been formed – through analogy or borrowing (as described in Section 2.4). An example of polysemy is the word *lens*. The Concise Oxford Dictionary (Fowler et al., 1999) gives four distinct meanings for this word:

1. a piece of a transparent substance with one or (usu.) both sides curved for concentrating or dispersing light-rays, esp. in optical instruments etc.;
2. a combination of lenses used in photography;
3. *Anat.* – crystalline lens;
4. *Physics* – a device for focusing or otherwise modifying the direction or movement of light, sound, electrons, etc.

The first of these definitions is a general purpose one; the following three are for the domains of photography, anatomy and physics respectively. However, these meanings are not restricted to appearing only in specific texts belonging to each domain. For example, radiophotography plays a large part in the world of medicine. Similarly, a text about ophthalmology would be likely to incorporate elements from both anatomy and physics, and the general language meaning of the word could appear in any domain-specific text. Whilst clearly all four meanings are related, we still need to be able to make distinctions between them, especially for applications such as machine translation and information retrieval. One of the reasons why it is often hard to make this kind of distinction is that it is a subjective one. This classic "lumping vs. splitting" problem[2] is a constant problem for lexicographers, and is one of the main characteristics which distinguishes their individual work. If this is hard to perform manually, it is likely to be even harder to perform automatically.

[2] This problem concerns the decision whether to combine different meanings (or shades of the same meaning) in one entry, or to separate them into individual ones

The distinction between polysemy and homonymy is also not always clear-cut, because it can be difficult to tell whether two meanings are related or not. However, this distinction is more important for lexicographers and theoretical linguists than for term recognition, since as long as we are able to recognise different meanings of a term, it matters less what exact form these differences in meaning take, unless we only wish to recognise one type of difference and not the other.

Componential ambiguity

The third type of ambiguity, which we call componential, occurs only in multi-word lexical units (be they terms or collocations). It exists where the individual components can be interpreted as being attached in different ways; in other words, the relationship between the components is ambiguous. There are two principal ways in which this can occur. The first one is essentially a constituent bracketing problem. Terms which are composed of more than two words frequently contain a nested term within them, but it is not always evident which words comprise the nested term. For example, the term *basal cell carcinoma* might be composed of *basal* + *cell carcinoma*, or it might be composed of *basal cell* + *carcinoma*. For a terminologist familiar with the domain, this is not so problematic, but for an automatic system, methods might be needed to resolve this. Worse still, when words are joined together to form a term, it may not be obvious where the split occurs. This problem is frequent in languages such as German, with terms such as *Wachtraum*, which could equally be composed of *Wacht* + *Raum* (guardroom), or *Wach* + *Traum* (daydream). Similarly in Dutch we find *bil* + *drulle* (a clumsy driver) or *bild* + *rulle* (a roll of film). In English, this type of word or term formation is less prevalent, so the potential for ambiguity is reduced, but it is nevertheless important to be aware of this point.

The second reason for this type of ambiguity lies in the interpretation of the compound. This frequently occurs with terms because they are composed of two (or more) nouns which have had connecting words such as prepositions eliminated. For example, a *cat thief* is someone who steals cats (a "thiever of cats"), but a *cat burglar* is not someone who burgles cats (a "burglar of cats") but someone who burgles places "in a cat-like manner". The interpretation of nominal compounds has been the focus of much research in semantics, often focusing on very specific topics, e.g. (Sparck Jones, 1985). The interpretation of multi-word terms has, on the other hand, been largely avoided because of its complexity:

> "Interpreting compounds can require arbitrarily extensive inference, and further, inference on pragmatic information, i.e. the most exigent form of analyser process." (Sparck Jones, 1985)

The need for interpretation of compounds as part of the term recognition process is widely debatable, but in either case, it is clear that the problem of componential ambiguity needs at least to be recognised. Whilst the ambiguity may only be at a basic rather than a deep level, in that it may be resolved by human terminologists without difficulty, it may still create a problem for automated methods.

Multidimensionality

A multidimensional network is a network which enables a concept to be classified in more than one way. Although this is ideally to be avoided in a concept representation system, for the sake of simplicity and clarity, it is more or less inevitable to some extent. The classification of concepts is a difficult task because it is subjective – there is no right or wrong solution. It depends not only on the view of the person designing the hierarchy, but also on the type of objects to be classified and on the purpose of the classification system.

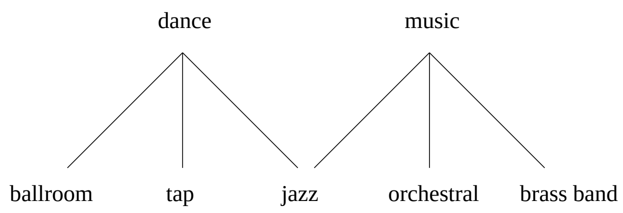

Figure 2.6: Example of multi-inheritance in a network

Various different terms have been used which express the idea of multidimensionality (Bowker, 1995), such as:

- multidimensional concept systems;
- polyvalent relationships;
- polyhierarchical relationships;
- facetted classifications;
- polydimensional series;
- polyhierarchic systems.

Because objects have a variety of defining characteristics, a decision has to be made as to which of these characteristics is used as the basis for the object's classification.

There are two main ways in which multidimensionality can occur. The most common one is *multiple inheritance* within a hierarchy. This is where a node has more than one parent (hypernym) in the tree. Figure 2.6 depicts an example of this, where *jazz* can be both a style of dance and a style of music simultaneously, but other forms of dance, such as tap, cannot be forms of music, and other forms of music, such as brass band, cannot be forms of dance. Another way to view this type of relationship would be with intersecting sets, as in mathematics (Figure 2.7).

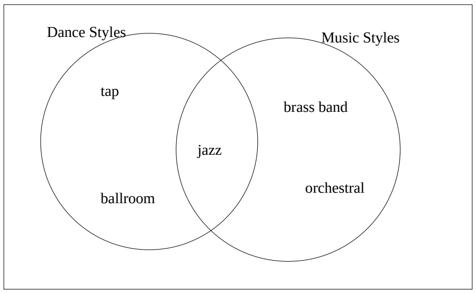

Figure 2.7: Multi-inheritance represented in set theory

Multiple inheritance is especially prevalent in medical terminologies (De Vries Robbé and Flier, 1994). For example, *pulmonary tuberculosis* is both a respiratory disorder and an infectious disorder. In the UMLS Medical Thesaurus (NLM, 1997), many terms are classified as both "signs/symptoms" and "diseases/abnormalities". For example, a *retinal haemorrhage* is a sign of *proliferative retinopathy*, but it is also an ocular abnormality itself. It could be argued that the classification system is at fault here for permitting such widespread, and in some cases, uniform, multiple inheritance to occur. This could be resolved by changing the parental nodes of the hierarchy to more suitable ones so that cross-classification cannot occur, but in practice this is virtually impossible to achieve, because of the very nature of terms in general, particularly medical terms.

The second way in which multiple inheritance can occur is where there is a kind of ambiguity in the term itself. The same linguistic entity can often be used to represent one or more of its components or constituents. For example, *coffee* can refer to the bean or to the whole drink. Similarly, *television* can refer to just the physical hardware, or to the contents of the display (hence the ambiguous question *"What's on the television?"*). In other languages, there may be more than one word to express these different ideas, e.g. in French we have *téléviseur* and *télévision* for the physical object and the contents respectively.

So in a hierarchy the term might be represented more than once, but unlike

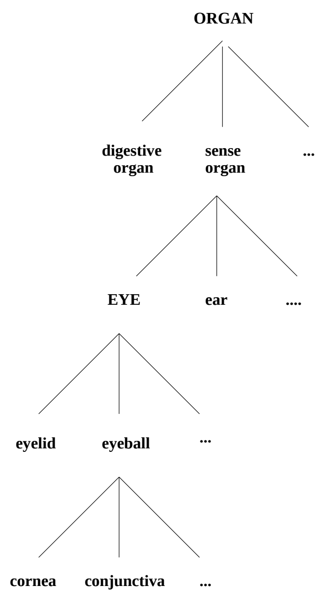

Figure 2.8: Section of a combinatory system (taken from (Felber, 1984))

in the previous case of a hierarchy with multiple inheritance, both instances of the term usually belong to the same section of the hierarchy. In this case, however, one would simply be higher up than the other, such that one is an ancestor of the other (usually with other nodes in between). This may also be because the hierarchy contains different types of relations, such as generic-specific and partitive (part-whole) relations (Bowker, 1995). This type of system is termed *combinatory concept system* (Felber, 1984) or simply *mixed system* (Picht and Draskau, 1985). For example, a concept contained in a partitive hierarchy might then be subdivided into generic-specific relationships, or vice versa. Felber (Felber, 1984) gives the following example of the concept *eye* being placed in a generic-specific hierarchy of organ types, but being further subdivided into its constituent parts in the lower part of the hierarchy (Figure 2.8). In this case, the term may only be represented once in the hierarchy, but have different meanings depending on the path one is following in the hierarchy, i.e. whether it is being viewed as a hyponym or a hypernym.

Although multidimensionality is an inherent problem of hierarchical networks, it is really the result of a more cognitive issue – the way in which we view the world. For example, the term *football boot* can be seen as either an item of footwear or as a piece of sports equipment. Under which category should it therefore be placed in a hierarchical representation? Most probably, it would be placed under both, or there would be a link from one position to the other. How the term is best categorised might depend on the individual circumstance. In a text about football, it might be more useful to know that a football boot is an item of footwear; in a text about footwear, it might be more useful to know that it is a piece of sports equipment, since in each case, the more specific information about the term is provided. Since it is virtually impossible to avoid a multidimensional representation, we need to find some means of accessing the term in its most relevant location according to the situation.

2.6.2 Term ambiguity in the field of medicine

Term ambiguity is a particular problem in the field of medicine because there can be very serious, even fatal, consequences if communication is not clear. The field is particularly susceptible to ambiguity because it is both multi-faceted and dynamic:

> "Medical nomenclature is more diversified than the other codes because besides actual objects it designates states, conditions, processes and operations, and it is not only subject to expansion, as the other codes are, but to a greater extent subject to revision as knowledge of human physiology, diseases and operations increases, and new methods of diagnosis and treatment are developed." (Sager, 1990).

Blois (Blois, 1988) claims that the problem is partly due to the "vertical" structure of the field of medicine, i.e. that "the ability to be precise is challenged as you go up the vertical scale" (Campbell et al., 1998). Blois views the domain as a hierarchical structure, ranging from the very precise subdomains, such as physics and chemistry, to the more general ones, such as sociology and psychology. Not only does this mean, as Blois claims, that the likelihood of ambiguity increases as soon as one moves away from the more precise areas at the bottom of the scale, but it also has more far-reaching implications. Because knowledge from a wide range of fields is necessary, problems associated with multidimensionality and with inter-domain meaning change are magnified.

Crookshank (Crookshank, 1923) goes so far as to criticise the medical profession for the communication problems caused by the lack of precision of the field of medicine. In particular, he claims that the field is not scientific because of the proliferative ambiguity arising largely from the multifaceted correspondence between terms and concepts:

> "Medicine ... [has] forfeited pretension to be deemed a Science, because her Professors and Doctors decline to define fundamentals or to state first principles, and refuse to consider, in express terms, the relations between Things, Thoughts and Words involved in their communication to others."(Crookshank, 1923)

Whilst Crookshank's ideal of greater precision is a laudable one, and indeed, one which terminologists strive for, the inherent problems not only in the field of medicine, but in the nature of language and terminology itself, make it an unrealistic option.

A further problem is caused not by imprecision, but by terminological changes evoked by time, increased knowledge and new discoveries. It is well known that general language changes over time, according to a variety of social, cultural and historical factors. It would be surprising if, given the dynamic nature of scientific fields such as medicine, terminology did not also change over time. The question of the correct naming of concepts has been argued for centuries, dating back to philosophers such as Plato and Aristotle. The problem of increased knowledge is typified by Frege's classic example of the morning and evening star (Frege, 1892). Because people did not originally know that the morning and evening star both referred to the same thing (the planet Venus), they gave them separate names. At that time, the names were therefore not ambiguous, but since we now know that both refer to the same concept, they have become ambiguous. Similar examples of these can be found in medicine, where increased knowledge about diseases has led to the discovery that seemingly different terms actually refer to the same thing. Conversely, one term may refer to several different diseases, because originally no differentiation was made between the diseases. For instance, the term *diabetes* was originally used to refer to two completely different diseases[3] with similar symptoms but very different causes and treatments. Even in the last decade, it has been established that even type II diabetes is actually composed of a variety of different diseases, again with different causes and sometimes different treatments. Naming conventions have not been fully established yet, but the terms *type II 1/2* and *type III* diabetes are becoming more widespread. Similar differentiations have been made between the different types of other diseases such as *hepatititis A, B* and *C*, and *viral* and *bacterial meningitis.* Failure to be precise about diseases such as these can easily lead to serious problems and even death, so it is crucial that communication is explicit. Even in the 1920s, Crookshank was well aware of the effect that this terminological imprecision could have:

> "In modern Medicine this tyranny of names is no less pernicious than is the modern form of scholastic realism. Diagnosis, which ... should mean the finding out of all there is wrong with a particular patient

[3]They are now known by a variety of names: most commonly *type I* and *type II* diabetes.

> and why, too often means in practice the formal and unctuous pronunciation of a Name that is deemed appropriate and absolves from the necessity of further investigation. And, in the long run, an accurate appreciation of a patient's "present state" is often treated as ignorant *because* it is incompatible with the sincere use of one of the few verbal symbols available to us as Proper Names for Special Diseases." (Crookshank, 1923)

Crookshank also criticised the medical profession for "...symbolizing various clinical happenings, in different way at sundry times, and in divers places ...", and for claiming that the "biological properties" of diseases were changing over time, instead of realising that it was the simply the naming procedure which was at fault, and that the designation of concepts was changing. He also claimed that there were no attempts made to actually define a disease:

> "In a word, no attempt is made to distinguish between what we observe in persons who are ill, on the one hand, and the general notions we form in respect of like illnesses in different persons, together with the 'linguistic accessories' made use of by us for purposes of communication concerning the same, on the other." (Crookshank, 1923)

Unfortunately, although people such as Crookshank were aware of the terminological problems in the field of medicine over 60 years ago, little has been done to resolve such issues.

2.7 Term Variation

As discussed in Section 2.2.1, there should ideally be a one-to-one correspondence between terms and the concepts they represent. In practice, however, there is often a many-to-many correspondence. Not only can a term represent more than one concept, but a concept can be represented by more than one term. Such terms are generally referred to as *synonyms* or *term variants*.

The presence of term variation within a subject field tends to be much more widely acceptable than term ambiguity. The International Organisation for Standardisation (ISO-704, 1987) permits the use of term variants (synonyms) within a subject field, but only permits the use of ambiguous terms (homonyms) where the ambiguity lies across, rather than inside, a domain. It recommends that preference is selected for one term amongst the variations, which should be the most commonly used and one which is not ambiguous and conforms to the guidelines, where possible. If such a term does not exist, it suggests creating a new term rather than using a pre-existing one which does not conform.

Term variation is a serious problem for information retrieval tasks such as internet searching. A search using a particular keyword or term may not match different variations of this term, which nevertheless refer to the same concept and are thus valid. For example, if we searched for references to *cell wall*, we might not retrieve documents containing phrases such as *cellular wall*, which clearly would be relevant. Similarly, searching for *lapwing* would not necessarily recognise instances of *peewit*.

It is also problematic for many areas of terminology such as automatic abstracting, information extraction and thesaurus generation. Information about terms may be left unidentified because it is unclear that the term is being referred to. For example, a system which identifies relations between terms and their context (which might be used for any of these terminological applications) could fail to recognise either that the same term is being referred to in each case, and thus fail to correlate the two. Worse still, it might totally disregard any variations of the term if it is not recognised that they are related, and thereby fail to identify any useful contextual information relevant to it.

Type	Subtype	Example
Morphological	Inflection	enzyme activities
	Derivation	enzymatic activity
Syntactic	Insertion	enzyme amidolytic activity
	Permutation	activity of enzyme
	Coordination	enzyme and bactericidal activity
Semantic		fermentation

Table 2.2: Types of variants of the term *enzyme activity*

Terms can vary according to morphological, syntactic or semantic criteria. These categories can also be further subdivided. Morphological variants can be either derivational or inflectional (or a combination of the two). We also adopt Jacquemin and Royauté's distinction between three types of syntactic variants: coordination, permutation and insertion[4] (Jacquemin and Royaute, 1994). Table 2.2 gives examples of each of these types and subtypes. Thes types were orginally defined for French, and it should therefore be pointed out that the incidence of the permutaion variant is less likely to render a term in English than in French. For example, the example given in the table, *activity of enzyme*, is unlikely to be considered a term in English, although its French equivalent, *activité d'enzyme* would. Jacquemin and Royauté define a separate category of morphosyntactic variants: a combination of (derivational) morphological and syntactic variations. However, since all three types (morphological, syntactic and semantic) can be combined together in any way, the creation of this distinct category seems perhaps unnecessary.

A further type of variant, which we might call *orthographic variant* should also be mentioned. For example, an unweaned piglet is termed a *shoat* or *shote*. Both terms are identical in meaning. Such variants may also be linked to pragmatic criteria such as American versus British English. This may have

[4]insertion was later renamed "modification" (Jacquemin, 1996)

2.7.1 Formation of variants

Lauriston also discusses term variation under the guise of term formation (Lauriston, 1996; Lauriston, 1995). He discusses three types of term formation process which do not actually result in new terms being created, but rather in variants of existing terms. These he calls:

- conversion;
- compression;
- semantic drift.

Conversion

Conversion is similar, but not identical, to Jacquemin's syntactic variation, where a new term is created by changing the syntactic category of an existing term. Lauriston's conversion is a narrower concept in that it only exists in the case where the term does not outwardly change its form. This means that there is syntactic but not morphological change. Alternatively this could be described as a syntactic change and a case of zero-affix derivational morphology. In English, this can take place between a variety of syntactic categories, although the most common are verb-noun and noun-verb transformations, e.g. *plant*, *estimate*; and adjective-noun and noun-adjective transformations, e.g. *resident*.

Compression

Compression is the process of shortening a term to form an acronym or abbreviation. For example, the term *central processing unit* is also known as *CPU*. Again, this is not so much a new term that has been created but a variation on an existing term. Jacquemin (Jacquemin, 1996) does not really account for this type of variation, presumably categorising it as a semantic variation (with which he does not deal as yet). The problem with this type of variant is that it is difficult to establish when use of such an acronym becomes sufficiently common to warrant it becoming a valid term.

In addition, the identification of acronyms and other abbreviations is not always correctly handled by automatic term recognition systems. Systems based on statistical methods which do not associate compressed forms with their extended forms may produce misleading frequency information about the occurrence of such terms, whilst linguistic methods may not handle these compressed forms correctly. Identifying and extracting abbreviations is a far from trivial task that normally requires individual processing in a separate module. Methods for identifying acronyms may be slightly different from those used for other abbreviations, since acronyms are composed of the initial letter or letters of the original term or phrase, whereas abbreviations are simply some kind of shortened form of the original. The KEP Program (Bowden et al., 1998), which extracts conceptual knowledge from texts, incorporates an automatic acronym acquisition module. This works by recognising expanded versions of the acronyms in the text and conflating the two. Where no expanded version exists, however, the system fails, so this method is only useful where the two variants co-exist.

Acronyms and abbreviations can also be ambiguous, since they may refer to different complex forms. This is not a widespread problem, in view of the fact that generally they will not be ambiguous within a specific domain, but only across domains. Even within a domain, however, ambiguity can still occur. Mountain Data Systems' Acronym Finder (Mountain Data Systems, 1988) lists the following expansions for the acronym ATR:

```
Above Threshold Reporting
Acceptance of TEMPEST Risk
Acceptance Test Report
```

`Advanced Tactical Radar`

`Aerial Target Review`

`Aided Target Recognition`

`Air Traffic Regulations`

`Air Transport Rack`

`Air Turbo Rocket`

`Alternate Trunk Routing`

`Ammunition Transaction Report`

`Analog Tape Recorder`

`Antitransmit Receive`

`Automatic Target Recognition`

`Automatic Terminal Recognition`

`Automatic Trouble Reporting`

No mention is even made of its usage in the field of terminology (Automatic Term Recognition) and doubtless other possibilities. It is clear from this list, however, that there is still potential within a specific domain for confusion to arise.

Not only can abbreviations be ambiguous with respect to the complex terms they refer to, but also with respect to ordinary words or terms which are not related (i.e. they are homographs). Although acronyms are generally written in capital letters – sometimes with full stops between each letter (although this is becoming increasingly rare) – this is not always the case. Furthermore, many recognition systems do not distinguish between upper and lower case letters. Lauriston (Lauriston, 1996) lists the following examples for English of homographic acronyms:

> ACE, ALP, ARM, ART, ASP, BIB, BITE, CAM, CASE, CELTIC, CHAR, CONS, COT, EARS, ED, FAW, FEE, FIB, IT, LAP, MAD,

MAP, MART, MASE, ME, MUM, MUT, NEXT, PAD, PI, PIN, RAN, REV, ROSE, ROW, SAME, SAP, SO, SPADE, SPITE, STAB, SUB, TAP, TIC, TIE, TON, TUG, VAT, VIA.

It is also common not only for the compressed term to become valid as a variant, but to replace the original complex term in usage, thereby reversing the roles of term and variant, so that the abbreviation becomes the main term and the original term the less common variant. Naturally, this is a possibility with all variants, but is far more likely to occur in this case as the abbreviation becomes more widely known and used (because in general people prefer to use simpler forms of complex terms). For example, DNA and TNT are much more commonly used than their respective complex forms Deoxyribonucleic Acid and Tri-Nitro-Toluene. Once an acronym or abbreviation becomes well-established, it can even taken on different morphological forms. For example, in the medical field the acronym OD is used for the noun *overdose*, but can also be used as a verb – "he collapsed after OD-ing on heroin". In cases like this, it is clear that the compressed form behaves in the same way as the extended form and is really nothing more than a spelling variant (although some would class it as a semantic variant, since the two lexical forms share little in common and do not have a root form or stem in common). There are also, however, cases where syntactic categories of the abbreviation exist that are not a possibility for the extended form. In the field of education, the abbreviation ARD is used in place of the more complex term *Attention, Review and Dismissal* (used for special needs students). Although the complex term is a nominal form, the abbreviation can be used as a verb – "We ARDed the student into special education". The complex term cannot be used in this way, i.e. it is impossible to replace the above expression with "we attention, review and dismissalled the student" or "we attended, reviewed and dismissed the student".

Semantic drift

Lauriston's third term formation process, semantic drift, is more about ambiguity than term variation. He defines semantic drift as the case where an existing term takes on a new meaning, without any change to the morphology or syntax. This is not so much a question of variation but ambiguity. Lauriston calls these terms

> "... the wolves in sheep's clothing stealing through terminological pastures." (Lauriston, 1996)

By this he means that they are hard to recognise, not only for ATR systems but even for humans, and he claims that no ATR system has ever attempted to detect them. This is because often the semantic drift can be quite subtle, although ambiguity in general is also not necessarily about clearly defined sets of meanings. Where semantic drift entails a total replacement of meaning rather than just an alternative meaning (as often happens over time), this ceases to become a case of ambiguity and instead becomes a case of meaning change. Semantic drift can occur between domains, where a term is borrowed from a different subdomain, but it can also occur within a domain.

2.7.2 The importance of variants

In spite of the difficulties associated with recognising what Lauriston calls "deviant usage" of terms, it is nevertheless important for term recognition. In line with the distinction he makes between terms and termforms, which we have discussed earlier, he claims that some of these distinctions in meaning can only be detected with the help of discourse structure and possibly even pragmatic knowledge, and until then, term recognition will only be termform recognition. By this he means that there will be no understanding of what is recognised, because systems will fail to "detect new terms in old termforms".

Lauriston performed a set of experiments to test the productivity of different methods of word formation in the telecommunications field, and discovered that the frequency of appearance of new terms over a period of four years was astonishingly high. Of these, the most common methods were conversion, compression and semantic drift, which suggests that variation and ambiguity account for a high proportion of the appearance of new terms. It is often claimed that cases of ambiguity are greatly reduced when dealing with specialised languages, because ambiguities frequently occur across domains or between general and specialised languages, but in Lauriston's study of semantic drift, over three-quarters of the cases were identified as occurring within the same domain.

2.8 Summary

In this chapter we have examined and discussed some of the major theoretical issues underlying the study of terms, drawing on research in the fields of terminology, psycholinguistics, cognitive science and theoretical linguistics. These issues all play an important role in our understanding of the theory of terminology, which is necessary for developments in its practical applications such as term recognition and disambiguation. We first investigated the relationship between terms and concepts from a variety of viewpoints, and then highlighted two important issues stemming from this relationship, namely ambiguity and variation. There still remains much to be understood about these issues from a theoretical perspective, which in turn poses a significant problem for many practical applications, as will be discussed in the next chapter.

Chapter 3

Practical Approaches to Terminology

3.1 Introduction

In the previous chapter, we examined various theoretical issues in terminology, in particular the relationship between terms and concepts. In this chapter we aim to provide an overview of some of the main practical approaches which have been used to deal with these problems. The chapter is divided into three main sections: approaches which deal with term recognition, disambiguation, and variation.

3.2 Term Recognition

Approaches to term recognition generally fall into one of three categories: linguistic, statistical, or a combination of the two. Multi-word term extraction consists almost entirely of the latter, since single-track approaches are largely insufficient if they are not to be restricted to very limited applications. Combinatory approaches tend to be biased heavily in one or other direction – mostly towards the statistical side. Independently from this, the knowledge used to identify terms can be either intrinsic or extrinsic to the term. Intrinsic information might involve morphological, syntactic, semantic or statistical knowledge about the individual term or its components (lexical items or even morphemes). Extrinsic knowledge makes use of either contextual information (either linguistic or statistical) or information from exterior resources such as dictionaries, ontologies or other corpora.

Approaches to term recognition really require an understanding of the nature of terms, and thereby "not only contribute to the applications of computational linguistics but also to the theoretical foundations of terminology." (Kageura and Umino, 1996). Some comprehension of the behaviour of terms is thus crucial to any successful approach to automatic methods of term recognition and disambiguation.

3.2.1 Term recognition and term interpretation

A distinction is commonly made between term recognition and term interpretation (Lauriston, 1996). Lauriston claims that the recognition performed by current ATR systems is really termform recognition, precisely because it does not involve any kind of interpretation or understanding of terms. In this we agree with him, but we do not admit that term interpretation is necessarily beyond current capabilities. Often, however, term interpretation is taken to mean an understanding of the relationship between the components of a compound term. However, many terms are non-componential units, such that their meaning cannot be determined from the meanings of their individual components, e.g. *angle_closure_glaucoma, ghost_epithelium, descemet's_membrane, epithelial_basement_membrane.* Therefore, understanding the relationship between the units may not help in understanding the meaning of the term.

Ananiadou, on the other hand, claims that

> "recognition of a compound structure cannot be carried out in isolation from interpreting it" (Ananiadou, 1988).

Although interpretation of compounds in general language is intrinsically hard:

> "Interpreting compounds can require arbitrarily extensive inference and further inference on pragmatic information, i.e. the most exigent form of analyser process" (Sparck Jones, 1985),

Ananiadou claims that this is only true for general language, and that compound interpretation in sublanguages does not require such extensive information. This is mainly because there are more restrictions placed on compound terms than on compound words, particularly relating to the number of elements contained in the compound.

The main method of interpreting the relationship between the elements of a multi-word term is to look at the strength of association between components of a lexical unit. This approach is used both as a means of collocation extraction, and also as a means of determining how the term can be broken down into its various constituents.

For binary terms, it can generally be seen that one component is a modifier of the other. Since, in English, modifiers precede their objects, the first item is almost always the modifier of the second. Or to look at it another way, the second item is generally the head of the phrase. For example, *granulation tissue* is a type of tissue; *descemet's membrane* is a type of membrane, *proliferative retinopathy* is a type of retinopathy. For terms of more than two words, identifying the relationship can be more problematic. We know, by the same process, that *striated muscle fibre* is a type of fibre, but is *striated* a type of *muscle fibre* or is the *fibre* part of a *striated muscle*? In formal terms, should the term be analysed as *striated (muscle fibre)* or *(striated muscle) fibre*? Even where adjectives are present it is not so obvious. Take the term *perforated corneal ulcer*. Since *corneal* is an adjective, we know that there cannot be such a term as *perforated corneal*, since the last element should be a noun. However, the noun *cornea* could have been transformed into an adjective during the formation process of the longer term.

Determining how the elements of a multi-word term are attached to each other is very similar to the process of determining *nested terms*[1]. The only difference is that if two elements are attached as part of a larger term, it does not necessarily indicate that these two elements actually comprise a term rather than a lexical unit. For example, the term *cellular fibrous tissue* is analysed as *cellular (fibrous tissue)*, *fibrous tissue* being a nested term. Although it is not important at this stage whether the attached elements actually comprise a nested term or are simply attached as a lexical unit, the principle is the same. For simplicity's sake, we shall describe the method as if for nested terms.

There are two main methods for distinguishing which elements of a longer term form a nested term (or lexical unit), both of which are connected with the statistical measure *mutual association* – a measure of how strongly attached various elements of a lexical unit are. The first method is to look for examples of the candidate nested term on their own in the corpus. The second is to look for them as part of other larger terms, i.e. to see if they can modify or be modified by other elements.

Method 1: Nested terms occurring singly

If we take the term *striated muscle fibres*, we see that there are two interpretations possible:

1. striated (muscle fibres)

2. (striated muscle) fibres.

[1] terms in their own right incorporated as part of larger terms

We examine the rest of the corpus for examples of the two possible nested terms: *muscle fibres* and *striated muscle*.

We find examples of *striated muscle* but not of *muscle fibres*. This would seem to indicate that (1) is the correct interpretation. One of the problems with this method, though, is that we do not find negative examples in a corpus, i.e. just because something does not occur does not mean that it cannot exist.

Looking at another example, *basal cell carcinoma*, we find two similar interpretations possible:

1. basal (cell carcinoma)

2. (basal cell) carcinoma.

Here, we find examples of both potential nested terms in the corpus, so we turn to frequency of occurrence. If the difference in number of occurrences is above a certain threshold, we might assume preference of one interpretation over another. In our corpus, we find 906 occurrences of the former, and 33 occurrences of the latter, which suggests that the former is more likely to be a nested term. Of course, the possibility also exists that both are actually nested terms, which means that we cannot easily distinguish which of the two is in fact being used as part of the longer term in this case.

Method 2: nested terms as part of other terms

If we apply the second method to the above two examples, the results are somewhat different. In this approach we look for examples of the candidate nested term being modified by a different word or term to form a new term. The theory behind this is that the more different terms the candidate nested term can appear in, the more likely it is to be a valid interpretation. However, this can be misleading, as the example shows.

Looking for examples of *X muscle fibres*[2], we find two terms: *ciliary muscle fibres* and *skeletal muscle fibres*. We do not, however, find any examples of *striated muscle X*. This would seem to indicate that (2) is the correct interpretation (contrary to the indications from Method 1). However, we can take this a step further. We now have the following examples of candidate nested terms:

1. striated muscle
2. ciliary muscle
3. skeletal muscle
4. *(muscle fibres)

and the following examples of longer terms:

1. striated muscle fibres
2. ciliary muscle fibres
3. skeletal muscle fibres

All the alternative longer terms found containing *muscle fibres* also contain candidate nested terms of the form *X muscle* found in the text on their own. The alternative nested term *muscle fibres* never occurs in the text on its own. From this we might assume that the best interpretation of our original term is *(striated muscle) fibres*.

If we take a look at the second example, *basal cell carcinoma*, we find both *basal cell X* and *X basal cell* occurring in the text. We also find *X cell carcinoma*:

1. basal cell papilloma
2. basal cell papillomata

[2] where X stands for a random word

3. cystic basal cell

4. squamous cell carcinoma

But, as in the previous example, we also find examples of the nested term *squamous cell* occurring on its own. In the same way, this suggests that the correct interpretation is therefore *(basal cell) carcinoma*, particularly since *basal cell* is found in more different terms than *cell carcinoma*.

Clearly these approaches to term interpretation are rather simplistic, and are far from foolproof, but aside from validating against dictionary entries, they are the only real automatic methods available. With these small examples, we do not attempt to prove or disprove either Lauriston's or Ananiadou's claims, but rather to illustrate the importance and problems associated with term interpretation.

The ability to determine nested terms, however, is not only important for term interpretation, should we choose to perform this at any level, but also for term extraction. Although we wish to extract the longest terms possible, we may also wish to extract nested terms if these are valid in their own right, even if they never appear on their own. Also, if a statistical method of term extraction is used, it is important that nested terms are taken into consideration when considering frequency values (Frantzi, 1998).

3.2.2 Approaches using linguistic information

Methods of ATR involving purely linguistic information have been fairly limited. They fall into two main categories: those which use extrinsic information and those which use intrinsic information (relative to the term). The types of information used tend to be syntactic for the former, and syntactic or morphological for the latter. Little use has been made of semantic information for the purpose of term recognition, largely because, in contrast to morphological and syntactic information, it is hard both to identify and to manipulate such information easily.

Morphological approaches

Ananiadou (Ananiadou, 1988) provides a method using domain-specific morphological rules to recognise single-word terms. Ananiadou's model was developed for the field of immunology, but it is unlikely that such a method could be successfully applied to subdomains outside the medical field which rely less heavily on neoclassical forms.

Barrier methods

The LEXTER tool (Bourigault, 1992) extracts candidate terms in the form of maximal noun phrases from French texts, using certain syntactic categories as delimiters, and then applies further linguistic analysis to reduce these to shorter candidate terms where appropriate. For French, there are two categories used as delimiters:

1. verbs

2. any preposition except "de" and "à" followed by a determiner.

Bourigault makes two main claims to support his method: firstly, that syntactic information alone is sufficient for an extraction system, and secondly, that an automatic extraction system cannot be expected to provide perfect precision (i.e. it will necessarily overgenerate). So LEXTER is designed only to select initial candidate terms, which must then be verified by a human terminologist. Although it was developed for the field of information documentation, using a large corpus as the basis for the creation of rules, the method is one which could theoretically be applied to different domains, since it does not rely on any prior knowledge of the domain or external lexical resources. If ported to a new domain, the analysis rule base would have to be refined, but the same corpus-based methods could be reused for this.

Early experiments with similar barrier methods were undertaken by Tersmette et al. in the medical domain (Tersmette et al., 1988), and later implemented more fully by Nelson et al. (Nelson et al., 1995). In this domain they have proved fast and effective at selecting initial candidate terms. Nelson et al. created a barrier list of all words which were not nouns or adjectives, which was then manually reviewed and adjusted. This list was then used to find "chunks" of text which were suitable as candidate terms. Their barrier list was much longer than those used by other similar approaches, which resulted in better recall, but much manual work was necessary to obtain satisfactory precision.

Mixed approaches

A few methods use a mixture of intrinsic and extrinsic information for term extraction, though the emphasis is usually on one or the other. Oakes and Paice (Oakes and Paice, 1998) developed a system for term extraction for the purpose of automatic abstracting in the domain of crop husbandry. This initially uses a pattern matching system to identify candidate terms (i.e. strings which occur within fixed lexical environments). Then the candidate terms are "lexically validated" by verifying that the internal structure is acceptable, both syntactically and semantically. This lexical validation is achieved by ensuring that the components of the candidate string belong to a fixed vocabulary and are combined according to a pre-defined set of grammar rules.

3.2.3 Approaches using statistical information

Approaches to ATR have traditionally relied heavily on statistical information. The methods used tend to be borrowed from other fields such as information retrieval and collocation extraction.

Frequency of occurrence

Frequency of occurrence is one of the simplest and most popular methods of statistical term extraction. It is useful in that it is domain-independent and requires no external resources. In its simplest form, candidate terms are awarded a ranking based on their frequency of occurrence in the corpus. The more frequently a candidate term appears, the more likely it is to be a term. Generally some kind of initial selection also has to take place. This tends to be linguistic, making use of some kind of syntactic filter that only permits certain categories, or certain syntactic patterns, to be selected as candidate terms.

Justeson and Katz (Justeson and Katz, 1995), Dagan et al. (Dagan et al., 1994), Enguehard and Pantera (Enguehard and Pantera, 1994), and Lauriston (Lauriston, 1996) all use frequency of occurrence to rank candidate terms.

Variations on frequency of occurrence

Frequency of occurrence is not really sufficient on its own, however, so it is often used as a starting point for further statistical or linguistic procedures, as in the C-Value and NC-Value approach (Frantzi and Ananiadou, 1996; Frantzi, 1998; Frantzi and Ananiadou, 1999). It may also be used as a base for connectionist approaches (Marshall and Bandar, 1999).

Smadja (Smadja, 1993) uses information about the position of the component words of the term. The more frequently the component words appear in the same position with respect to each other, the more likely it is that they constitute a term.

Mutual information

There are several variations on the mutual information approach, but this method essentially measures the unithood of a candidate term, i.e. the strength of association of its components to each other. Church and Hanks adopted this measure for collocation extraction (Church and Hanks, 1990), but it is also used by Damerau (Damerau, 1993) and Daille et al. (Daille et al., 1994) for the extraction of binary terms. According to Frantzi and Ananiadou (Frantzi and Ananiadou, 1995), it is difficult to implement for longer terms.

Damerau's method (Damerau, 1993) compares the relative frequency of the term in the document with its relative frequency in a reference corpus. It is used to extract both single-word and bi-word terms, in a simple but effective manner, with no recourse to any other information. This also has the advantage that, provided a suitable reference corpus exists, it is domain-independent.

Daille et al. experimented with various other statistical methods such as the *log-like coefficient* proposed by Dunning (Dunning, 1993) and the Φ^2 coefficient proposed by Gale and Church (Gale et al., 1991). They found that for binary terms, frequency of occurrence was the best approach and that combinations of measures did not bring significant improvements.

Statistical barrier methods

We have already discussed barrier methods as a linguistic approach to term extraction. A similar type of approach is taken by Nakagawa and Mori (Nakagawa and Mori, 1998), but they make use of *statistical* information about boundaries to delimit candidate terms. They consider information about the type of words which immediately precede and follow a candidate term, i.e. "boundary words". Since only a small number of single-word nouns contribute towards most multi-word terms, they allocate a score to each boundary noun based on the number of compound nouns that incorporate that boundary noun. Interestingly, the method compares well with other methods such as the NC-Value quantitatively (i.e in terms of recall and precision), but they find a qualitative difference in the type of terms extracted.

3.2.4 Hybrid techniques

Recently, approaches to ATR have been inclined towards using both statistical and linguistic information. The main part of the algorithm is generally the statistical part, with much less emphasis being placed on the linguistic part. Linguistic information is incorporated into the algorithm in the form of a syntactic filter which only permits certain combinations of syntactic categories to be considered as candidate terms. There is some debate as to the ideal filter which will allow maximum recall without lowering the precision rate.

Dagan et al. (Dagan et al., 1994), Daille (Daille et al., 1994), van der Eijk (van der Eijk, 1993), Justeson and Katz (Justeson and Katz, 1995) and Frantzi (Frantzi, 1998) all use hybrid techniques, combining a linguistic filter with the statistical measures described above. With the exception of Frantzi's approach, they differ mainly in the type of filter they adopt. Frantzi uses the same filter as Justeson and Katz, but her method differs more in the statistical part of the measure. Ranging from the very strict to the more open, the filters are as follows:
Dagan: $N^*\ N$;
Daille: $(N|Adj)N$;
Van der Eijk: Adj^*N^+;
Frantzi; Justeson & Katz: $((Adj|N)^+|((Adj|N)^*NPrep(Adj|N^*)N$.

The NC-Value uses linguistic information to delimit lexical strings on which a series of statistical measures are performed. The linguistic information takes the form of a syntactic filter on the strings to be selected as candidate terms. Statistical information about frequency of occurrence and string length is compiled, to generate an initial term ranking. A second pass is then performed using contextual information. Words from the context are passed through another syntactic filter, to select those which are most important, and statistical information is then gathered about the co-occurrence of context words with the candidate terms. The weights generated by this process are then combined with the original ranking of the terms to form a new ranking.

3.2.5 Lexical atoms and collocations

Term extraction has been influenced by measures used for other applications such as the identification of collocations and lexical atoms[3]. For example, Lin (Lin, 1998) extracts collocations from corpora by using syntactic information to extract dependency triples, and then using mutual information to identify those which are more than coincidental relationships. This is a very similar approach to the hybrid methods used for term extraction. As we have mentioned, Gale et al. previously used the mutual information measure not for term extraction but for collocation extraction.

Collocation extraction has also provided ATR with another statistical measure - that of *unithood*: "the degree of strength or stability of syntagmatic combinations of collocations" (Nakagawa and Mori, 1998). These are adopted by many of the statistical and hybrid approaches to ATR such as the NC-Value.

Evans and Zhai (Evans and Zhai, 1996) adopt a hybrid approach by combining corpus statistics with linguistic heuristics (in the form of noun phrase analysis) to create "indexing phrases" for information retrieval. By identifying linguistic

[3]lexical atoms differ from collocations in that their meaning is non-componential, e.g. *hot dog*, whereas collocations may be componential or non-componential in meaning. Many terms are therefore also lexical atoms.

relations between the elements of the noun phrases extracted, they attempt to capture some kind of conceptual content rather than the traditional flat structures more usually obtained in information retrieval.

Evans and Zhai claim that a shallow understanding of the content of the document is necessary (and achievable) for information retrieval, although for other tasks, such as machine translation, it may not be sufficient. Although most ATR systems do not attempt to capture this kind of knowledge at all, we would argue that at least a level of knowledge similar to that of Evans and Zhai's system is useful for term extraction as well as for information retrieval.

3.3 Recognition of Term Variants

There has been relatively little work focusing on the development of tools for widespread variant recognition. Some early work was performed at Cambridge University in the 1980s to generate variants of terms for information retrieval (Sparck Jones and Tait, 1984). This produced both syntactic and semantic variants, although the distinction between them was blurred. For example, variation according to number was regarded as semantic, whilst variation according to verb aspect was regarded as syntactic. On the other hand, the production of synonyms (i.e. terms sharing the same meaning) was managed by the syntactic component. In essence, variant generation was motivated by practical rather than linguistic considerations. Unlike the approach taken by Jacquemin et al. with the FASTR tool (Royauté and Jacquemin, 1993; Jacquemin and Royaute, 1994; Jacquemin, 1996), Sparck Jones et al. were essentially dealing with producing, rather than recognising, term variants. So their approach consisted of systematically deriving all the possible linguistic realisations of a concept, and then matching these with the text. Although this has the potential to generate a lot of noise in the form of "nonsense strings", e.g. *retrieval by information* as a variant of *information* *retrieval*, Sparck Jones et al. claim that this does not matter since such nonsense strings will never be found in the text anyway. However, there is still the potential for overgeneration with the method, since it may pick up "false segments", i.e. a string of words which does not constitute a meaningful lexical unit.

Inflectional morphology is generally dealt with by stemming algorithms, e.g. (Porter, 1980), or some form of normalisation procedure reducing the different morphological forms of a word to its root. These techniques are commonly used in information retrieval, e.g.(Krovetz and Croft, 1992). Derivational morphology is more problematic to handle because there is a much greater risk of incorrect conflations. Generally an electronic dictionary is used to match whole words, rather than treating the stem and suffix separately.

FASTR deals with term variation by recognising morphological, syntactic and morphosyntactic variants separately, although it cannot, as yet, deal with semantic variation. Inflectional and derivational morphology are handled using a grammar which links sets of potential suffixes for the stemmed terms. Conflation between term and variant is performed by matching with a base form, e.g. an infinitive verb or singular noun. If the two forms can be conflated, their suffixes are then compared with the list of possible suffixes to ensure they are compatible (i.e. that both forms can be related by derivational or inflectional morphology). Again, overgeneration is a possibility because of the potential that stemming techniques present for erroneous conflations, particularly with short words (Krovetz and Croft, 1992), but the restrictions imposed on the suffixes help to minimise this risk. Using this technique, morphologically related nouns, adjectives and verbs can all be handled.

3.4 Disambiguation Techniques

The problems, pitfalls, and even necessity of *word* sense disambiguation have all been widely debated over the years, and methods for disambiguation vary from the purely linguistic to the purely statistical, with many varieties in between. But though approaches may vary, the task is essentially the same: to relate occurrences of a word found in a text to different senses, usually defined by a dictionary. This is not quite the same as relating occurrences of a term to different concepts, since although senses may differ according to the particular dictionary or other lexical resource, the senses are reasonably established and are concrete (represented in a dictionary by definitions). Concepts related to terms are not so easily definable and are not generally described in any formal way. Although some specialised dictionaries and other lexical resources are available, they are almost always both incomplete with respect to their coverage of terms in the field, and limited in terms of definitions. The whole idea of a concept is far more fluid than that of a dictionary sense. If a concept is not something concrete but abstract (a "mental picture"), then multidimensionality becomes even trickier to explain.

3.4.1 Levels of ambiguity

There has been some debate about just how useful it is to be able separate words into their various senses, i.e. to perform word sense disambiguation (Kilgarriff, 1997b). Kilgarriff argues that the whole idea of word senses is a controversial issue and that word senses only exist relative to a particular problem. He claims that "there is no reason to expect a single set of word senses to be appropriate for different NLP applications" (Kilgarriff, 1997c). We believe that this line of reasoning can equally be applied to terms and their senses.

In previous work (Maynard, 1997), we have proposed that syntactic ambiguity can be viewed at three different levels: at the most basic level whereby a naive parser will, for example, identify any structure of the type V + NP + Prep + NP as being ambiguous; at the next level whereby a more comprehensive parser will employ some more sophisticated means of disambiguation such as the encoding of semantic information or frequency-based techniques; and finally at the highest level, which views ambiguity in human terms. The fact that humans find relatively few sentences genuinely ambiguous in terms of structure leads us to suggest that much can be learnt from studying human techniques for disambiguation.

In the same way, it may also be helpful to break semantic ambiguity down into three levels. At the most basic level, any word which has more than one possible meaning is ambiguous. At the next level, any simple WSD algorithm can use basic syntactic information to distinguish between some senses of words, or use frequency information to rule out the most unlikely senses. The highest level is again analogous with human understanding, whereby higher levels of processing such as discourse cues, semantic knowledge about the context and world knowledge are used. If situated in a context, even highly ambiguous words generally pose no problem for human understanding.

Most disambiguation techniques are based on the assumption that, in line with dictionary and thesaural definitions, words have a fixed number of senses and that it is possible to distinguish between them for any given occurrence of a word. In reality, this is not always the case. It is not always easy, or even necessarily appropriate, to do this, because distinctions between word senses are commonly fuzzy rather than absolute.

Along the lines of Kilgarriff, we believe that different NLP applications may require different levels of granularity when distinguishing between senses. For machine translation, it may not be necessary to operate at too fine-grained a level of distinction, since we can often take advantage of a "free ride" in this case. By this we mean that where a source language and a target language word are both ambiguous in the same way, there may be little point in bothering to disambiguate the word in the first place, since for both meanings the translation into the TL will be the same. On the other hand, for an application such as hypertext linking, it may not be desirable to forge a link between two occurrences of a word which do not have the same exact meaning. For this purpose, a much finer-grained level of distinction may be required, possibly at a level greater than would be provided by dictionary senses. This would seem to be particularly true for technical terminology, where terms with slightly different meanings may be referring to completely different things.

3.4.2 Approaches to disambiguation

There has been an abundance of literature on different approaches to WSD, which we shall therefore not discuss in great detail here. For further pointers, see (EAGLES Computational Lexicons Working Group, 1996; Ng and Zelle, 1997; Chandrasekar and Bangalore, 1998; Kilgarriff, 1997a; Resnik, 1997). A variety of methods has been used, such as supervised learning from large corpora, e.g.(Yarowsky, 1995), multilingual approaches using the difference between languages, e.g.(Dagan et al., 1991), statistical models, e.g. (Bruce and Wiebe, 1994), and using linguistic and contextual information, e.g.(Ng and Lee, 1996).

Traditional methods of sense disambiguation involve comparing the distribution of the words in a context window with typical distributions of the word's senses to find the closest match (Gale et al., 1992). The use of lexical semantic information has largely been ignored for term sense disambiguation, although it is increasingly being used for other NLP applications.

3.4.3 Using semantic information

The idea of using semantic information from the context for disambiguation purposes seems an intuitive one. Psychological studies have shown that we make use of this kind of information when processing text, e.g. (Connine et al., 1984). Connine suggests that we use information about selectional restrictions and thematic roles as a form of case frame, enabling us to make preferences about the likelihood of attachment. For example, knowing the semantics of the verb allows us to make decisions about which constituent is attached to it. In the same way, we might use this idea in reverse, i.e. knowing that a particular constituent is attached to the ambiguous word might enable us to select one sense of the word over another.

A more probable explanation for the human sentence processing mechanism is the use of contextual clues, in the form of pragmatic or real-world knowledge. For example, if we see the phrase *the table with a broken leg*, we would assume that it is the inanimate sense of leg that is being described, rather than the animate sense. The reason we make this assumption is because we know that tables are inanimate. It seems evident that this kind of information would be useful for automatic disambiguation, but the problem lies in how to impart it (Maynard, 1997). In order to be able to summon up the relevant piece of world knowledge, overwhelmingly large amounts of information would need to be stored, and there would then still be the question of establishing which of these pieces of information would be necessary to disambiguate a given word. A better solution would appear to lie in replicating the results of these methods using different techniques.

3.4.4 Acquiring semantic knowledge

Although semantic information has been largely ignored for term disambiguation, it has been extracted for a variety of other purposes, such as information extraction, dictionary construction and knowledge acquisition. Semantic knowledge can be acquired in two main ways: either by extracting the information from dictionaries or thesauri (Rayson and Wilson, 1996), (Riloff and Lehnert, 1992), (Montemagni, 1995), or by using knowledge-based or statistical techniques to extract the information directly from corpora (Grefenstette, 1994), (Reinert, 1986), (Jouis, 1994), (Wilms, 1996). Such techniques will be discussed in more detail in Chapter 5.

Most word sense disambiguation techniques rely on dictionaries or other lexical resources to provide the necessary information about word senses, thereby establishing their semantic interpretation. This can then be used as a reference point for establishing the correct meaning of a particular wordform, by establishing (by whatever means) the sense which seems the most likely or relevant to the wordform in question. Although we do make use of lexical resources (the UMLS Metathesaurus and Semantic Network), we do not use them in quite the same way. Instead, we view the context, rather than the definition, as being the key to the meaning of the term. That is to say, we do not necessarily seek to associate a termform with an established definition or meaning, but rather to group different termforms according to their senses as used in their particular context. So, although two instances of a term may have the same basic meaning, they may be used in different ways by exhibiting different facets of meaning according to the context in which they are found. We consider that it is often appropriate to distinguish between these different facets.

3.4.5 Word sense and term sense disambiguation

In many ways, term sense disambiguation (TSD) is similar to the problem of word sense disambiguation (WSD). However, it differs in two main respects. Firstly, terms are domain-specific, which means that general language resources and techniques may not be appropriate. Some work has been carried out on tuning lexicons to specific domains, e.g. (Basili et al., 1995). Secondly, the majority of technical terms consist of more than one word. The ambiguity of multiword terms is generally not caused by different senses of the individual components of the term, but by different senses of the term as a whole. The different meanings of the term may be linked to different domains, but they may equally be present within a specific domain. Although domain-specific text is likely to contain the meanings of a term related to that domain, it does not rule out the possibility of the general meaning of that term being used as well. For example, the term *drug* found in a medical text could refer to either an illegal substance or simply to a kind of medicine (as discussed in Chapter 2). Similarly, the term *desktop* in a business text could refer to part of the furniture or to the layout of windows on a computer. Because of the differences between word sense ambiguity and term sense ambiguity, techniques applied to WSD are not always appropriate for TSD.

3.5 Summary

The majority of methods used for term extraction focus largely on statistical information. This is unsurprising given the enormous volumes of text which are often involved in practical applications, and which necessitate semi- or fully automatic methods of term extraction. Where specialist corpora (which often contain smaller amounts of text) are involved, statistical methods alone may not be the best solution. Recently, hybrid approaches involving some linguistic analysis have been proposed, but the information used is largely syntactic and is fairly shallow. What most current methods lack is any kind of understanding of terminology ("term comprehension") and its use in the surrounding text. Occasionally term interpretation has been carried out, though it is acknowledged that this task is extremely difficult except in very specialised corpora.

The problem of term ambiguity is strongly correlated to the task of term comprehension. Little work has been carried out on term sense ambiguity, for the very reason that term comprehension is also largely ignored, and therefore it is deemed unnecessary to perform term sense disambiguation. We therefore turn to methods for word sense disambiguation which, in contrast, has been widely researched.

Methods for word sense disambiguation are also far more diverse than those for term recognition, ranging from the purely statistical to the purely semantic, together with all kinds of mixed approaches and methods borrowed from other fields such as machine translation, information retrieval and information extraction.

In the next three chapters, we shall investigate methods of incorporating contextual information into a statistical method for term extraction, and how this same information can be used for term sense disambiguation. We introduce a set of context weights which convey information about different parts of the context, and explain exactly how this is used for term extraction.

Chapter 4

Extracting Contextual Information

The NC-Value method has shown that statistical methods of multi-word term extraction, such as the C-Value, can be enhanced by the addition of contextual information (Frantzi, 1998). In this approach, contextual information is used in a primarily statistical way. The basis for this stems from the idea that terms are less easy to modify than non-terms because of the inherent restrictions on the modifiers allowed by terms, or "extended lexical units" (Sager, 1990). Since the type of modification is different for terms and non-terms, it should be possible to make use of this information to help distinguish one from the other. From a statistical point of view, the more often a modifier appears with a term, the more likely it is that any other lexical unit it modifies is also a term.

In this chapter, we discuss ways in which the NC-Value method can be enhanced through the addition of further contextual information and, in particular, using deeper forms of linguistic information. We first examine how contextual information has been identified and used in other work, and then investigate ways in which these methods could be adapted to suit our needs, and ways in which they could be extended to include other types of information. Finally, we outline our approach for extracting and combining the different types of contextual information we require, in order to produce a weighting score for each context word based on its usefulness.

4.1 Use of Contextual Information

As we have discussed, term extraction could benefit greatly from the addition of contextual information. The real problem lies in establishing exactly what kind of information should be incorporated. To answer this question, we shall first take a look at some of the ways in which contextual information has been used before, both in term extraction and in other NLP applications.

4.1.1 Terminography

The identification of knowledge-rich contexts is an essential task for terminographers, since it enables them to analyse the concepts underlying the terms in a particular domain (Davidson et al., 1998). Traditionally this is achieved by using a KWIC index to produce context windows surrounding the term, and then manually selecting those which are most relevant to the task. For terminography, linguistically and conceptually rich contexts are therefore required (Davidson et al., 1998). Extracting these is a time-consuming and laborious task when carried out manually, so it is of great benefit to terminographers if this can be achieved automatically, or at least semi-automatically.

4.1.2 Term extraction

In term extraction, little use has been made of contextual information other than from a syntactic point of view. This generally takes one of two forms – barrier methods and co-occurrence methods.

Barrier methods are based on the principle of identifying boundaries demarcating candidate terms, as discussed in Chapter 3. Though successful, such methods tend to require substantial manual or semi-automatic techniques. They may also have been tested only on a very specific language or domain.

Co-occurrence methods such as the NC-Value (Frantzi, 1998) are based on the principle that terms are restricted in the ways in which they can be modified. If some particular words in the context occur frequently with terms and infrequently with non-terms, it seems more probable that any candidate term occurring with one such word will be a real term.

4.1.3 Semantic information

Deeper forms of contextual information, such as semantics, have been largely ignored for term extraction, but are more commonly used for other purposes, such as the identification of synonyms (Grefenstette, 1994), and the definition of terms through the identification of other terminological relations (Davidson et al., 1998; Meyer et al., 1999; Oakes and Paice, 1998).

Identification of synonyms

This approach is based on the premise that two terms which share the same contextual environment are likely to have the same meaning. Grefenstette's approach focuses on the semantic properties of the term and its context. These properties, called *attributes*, are used to determine the similarity between term and context words, using a form of weighted Jaccard measure. In this way, context words are ranked in order of similarity to a term, and thereby determine synonyms, which can then be used for automatic thesaurus construction and other NLP applications.

From this idea we can surmise that the reverse philosophy is also true, and that we may be able to detect cases of ambiguity where two terms (with the same linguistic form) do not share similar contextual environments. Unlike Grefenstette, we do not rely solely on semantic features, but combine a variety of types of linguistic and statistical information, although emphasis is placed on semantic aspect.

Identifying relationships between term and context

Contextual knowledge in terminology is frequently used for obtaining semantic information about terms, either in order to define the meaning of the term or to define relationships for thesaurus construction or similar purposes. This is usually achieved by some kind of pattern matching: either lexical, syntactic, semantic, or a combination of these. Good examples are found in the domains of composting (Davidson et al., 1998) and crop husbandry (Paice and Jones, 1993). Similar pattern matching approaches are found in general language for applications such as message understanding in the MUC projects (Sundheim, 1995), information extraction in the CRYSTAL project (Riloff and Lehnert, 1994; Soderland et al., 1995), thesaurus generation (Grefenstette, 1994), and abstracting (Oakes and Paice, 1998).

However, for these purposes, at least within the terminological domains, the type of text used tends to be descriptive, such that the contexts serve to define the terms, and can be used for establishing relations such as hyponymy (generic-specific), meronymy (part-whole) and attribute information.

For example, the COGNITERM system (Davidson et al., 1998) identifies relations in the domain of composting. The example below shows some of the contexts Davidson et al. found with the term *compost*. Many of these contexts express clear definitions of the term or relations between term and context. The method they use relies heavily on the fact that the contextual information is so explicit.

```
... organic materials such as compost ...
Compost is a practical method of recycling organic
matter and reducing solid waste.
Compost is perhaps best defined as a pile of organic
materials deliberately assembled for fast
decomposition.
Compost is classified as a soil conditioner rather
than a fertilizer.
Compost is a wonderful soil amendment for the
vegetable garden or landscape, improving soil
structure and texture while adding valuable plant
nutrients.
Compost is the partially decomposed remains of plants.
... a humus-rich soil amendment known as compost.
```

The field of medicine is more unusual (and problematic) in that the type of text used for analysis tends to be something like patient records, which do not carry the same kind of information. For example, the following contexts occurring for the term *diabetic retinopathy* are typical of those found in our eye pathology corpus.

```
Chorio-retinal atrophy (presumed to be effect of
laser retinopathy).
No histopathological evidence of diabetic retinopathy
in plane of section examined.
Diabetic retinopathy. Recurrent vitreous haemorrhage
```

```
in left eye.
Preretinal glaucoma secondary to crvo or diabetic
retinopathy.
Diabetic retinopathy. Has had vitrectomy, but
developed retinal detachment.
Lens induced endophthalmitis. (History of diabetic
retinopathy).
Painful blind left eye. History of previous retinal
detachment operation and vitrectomy for diabetic
retinopathy.
```

The contexts differ in that they are less explicit about providing clues to the term's meaning. This problem is one of text type rather than domain, since there are types of medical texts which do convey this kind of contextual information. Patient information leaflets (PILs), such as those accompanying drugs and other medicines, are one such example. Here the text type differs in that it is *definitive* rather than *reportive*. By this distinction we mean that definitive text types are designed to provide information, and therefore the terms tend to be explicitly defined and explained. The contexts surrounding these terms will thus be useful in providing defining information about the terms. Example of such texts would be instruction manuals, information guides, textbooks and rulebooks. Reportive texts, on the other hand, are designed more to convey details of events or descriptions. They assume that the reader already has knowledge of the terms they contain, and do not therefore attempt to define them, but only to convey an information report. Examples of this type of text might be patient medical records, reviews, reports, journal articles, weather bulletins, etc. The distinction between the text types is in some ways analogous to Pearson's "formal" and "semi-formal" defining expositives (Pearson, 1998).

The crucial factor regarding text type is the presumed difference in expertise between writer and reader. The larger the gap in knowledge (such that the writer has more knowledge than the reader), the more likely that the text will be an informative one. The smaller the gap, the more likely that the text will be descriptive or explanatory. One exception to this might be the case where the writer has *less* knowledge than the reader. This is a rare scenario, for the purpose of communication is to convey information, and if the writer has less knowledge than the reader, he is unlikely to be able to convey much useful information about that topic. However, it can occur in the case where the reader is in some way judging either the writer's knowledge, or his ability to convey information, as in the case of a student's essay. In such a case, the type of text may vary considerably, depending on the situation, but it is unlikely that such text types will form part of a large-scale corpus.

The domain of medicine also imposes further constraints on semantic knowledge acquisition, in that the type of knowledge required is extremely diverse. As has been discussed in the literature on both terminology and medical informatics, e.g. (Sager, 1990), (Blois, 1988), and as we mentioned in Chapter 2, understanding of the medical field requires a wide range of knowledge from the general to the very specific.

4.2 Identifying Relevant Contextual Information

Our aim in this section is to investigate ways of identifying and extracting knowledge-rich contexts, i.e. those contexts or parts of contexts which contain domain-specific information about terms or concepts. This information can then be used to help extract terms, disambiguate them, or extract relations between terms or concepts. As we have seen in the previous section, there is a wide variety of types of information which can be used for such purposes, and a number of ways in which this can be achieved. The choice of method depends on factors such as the type of information available, the purpose for which it is being used, and the depth of knowledge required. For term extraction purposes, some types of contextual information are clearly more useful than others.

The NC-Value method uses linguistic information from the context to help identify terms, but it only make use of shallow information, in the form of a simple syntactic filter. This permits certain syntactic categories of words to be considered as context words. The categories accepted are those which are considered to be most “meaningful” as context words. As always, a distinction has to be made between precision and recall. A more open filter (i.e. one which permits more syntactic categories) would increase the chance of recovering useful context words, but would also increase the noise, since many of these context words might not be useful. So a more open filter would be expected to increase the recall but decrease the precision. Conversely, a more closed filter would be expected to increase the precision but lower the recall.

It is not only adjustments to the syntactic filter which would make a difference to the NC-Value. No distinction is made between the syntactic categories (Frantzi, 1998), nor is the relative position of the context word or its semantic properties taken into account.

In order to improve the chance of identifying the most useful context words, we therefore need to turn to other methods. The task of identifying which parts of the context are most useful and what type of information is relevant is one that is far from trivial (Maynard and Ananiadou, 1999a; Maynard and Ananiadou, 1999d). The context is generally considered simply as a “bag of words” and its structure, word order and relationship to the term in question are ignored in most systems. One exception to this is the dynamic matching algorithm developed by Radford et al. (Radford et al., 1997). This takes into account the linear order of the words within the context window. However, it still does not make any distinctions according to the type of information in the context. While dynamic matching does improve the efficiency of the algorithm, it is computationally expensive, even though semantic information is still ignored.

The question of relevance can be viewed from various different angles. Distinctions about the context could be made from the point of view of any of the following:

- Syntactic category;
- Linear position with respect to the term;
- Syntactic relationship between term and context word;
- Semantic properties of the context word;
- Semantic relationship between term and context word;
- Some other kind of relationship.

It is also necessary not only to decide which type of information, or which combination of types, is the most useful, but also to take into account whether this information is easy both to identify and to manipulate. Many kinds of semantic information would undoubtedly be useful for NLP applications, but they are largely ignored because of the difficulties of implementing methods which make use of them.

We therefore restrict ourselves to three main properties of the contextual information.

1. **Syntactic category**: the weight of the context boundary word should vary accordingly.

2. **Terminological status**: a context word which is also a term has greater significance.

3. **Semantic information**: context terms having similar semantic properties to the candidate term are accorded higher weights.

4.2.1 Syntactic information

There are essentially two ways in which the selection of appropriate syntactic information can be undertaken. Either the relevant information is identified and extracted, or the irrelevant information is identified and filtered out. Both methods have the same final goal, but one may be easier to achieve than the other. In fact, we combine them in a two-tiered approach. It is relatively simple to spot some initial trivial syntactic information which does not play any essential role, so this can be filtered out immediately. Rather than simply retaining a few syntactic categories as in Frantzi's approach with the NC-Value (nouns, verbs and adjectives), we retain a less dismissive approach and only filter out categories which are almost certainly unimportant for our purposes, such as determiners, conjunctions and adverbs. Once the context has been reduced in this way, it becomes easier to discriminate amongst the remaining information and to make more informed judgements about relevancy based on other criteria such as semantic "informativeness".

The initial elimination of empty syntactic categories is intuitive. Clearly the presence, absence or choice of a determiner is not going to add much information about a term. Although terms are restricted in their modification, they can be more or less freely modified by determiners, so this is of little help for term extraction purposes. The same can be said of adverbs (since they do not modify noun phrases, but verbs or verb phrases), conjunctions and most other linguistically closed categories. One exception to this is prepositions, which may have some importance in view of the fact that they can form part of a complex term[1].

[1]depending on the view taken by the terminologist on this point

Empty categories such as these should not be neglected totally, however. As we have seen, they can be useful as barriers which demarcate the outer edges of complex terms, i.e. the boundaries beyond which a term is unlikely to go. Since terms do not normally include categories such as conjunctions, the presence of such a category can be used to indicate the beginning or end of a complex term.

Although barrier methods of term recognition have been quite successful when combined with further linguistic processing (Bourigault, 1992), we do not make use of this method as such, since we prefer to use a more restrictive selection process. Instead of using a filter which is external to the term, we use one which is *internal*. This means that rather than setting boundaries and allowing anything within the boundary to be accepted as a candidate term, we force deeper restrictions on the categories that can occur inside the boundaries, i.e. the categories which constitute the candidate term.

On the other hand, we do make use of the barrier method in a slightly different way. Instead of only permitting certain syntactic categories to be acceptable as boundary limits to candidate terms, we discriminate between boundary words, such that some are more heavily weighted than others. The larger the weight, the more likely that the boundary word signifies a term. We shall explain in detail in Section 4.3.1 how these weights are determined.

4.2.2 Terminological information

Another source of information which is largely ignored in term extraction methods is that of pre-existing terminological information. Evidently we do not have much information of this type that can be applied directly to the candidate term *a priori*, but what is potentially available is information which can be applied to the context. A bootstrapping technique is useful here, in that once a first pass has been made regarding the identification of terms, this information can be used to fine-tune the categorisation procedure in a second pass. Furthermore, the terminological status of the context words may have some relevance to its importance (Maynard and Ananiadou, 1998a). We shall describe in Chapter 6 how this kind of information is incorporated in our algorithm.

4.2.3 Semantic information

The NC-Value method is limited in that the only kind of linguistic information it makes use of is syntactic category. It is clear that semantic information has a role to play with respect to the identification of terms (Maynard and Ananiadou, 1999b), since syntactic information is insufficient in some ways. For example, whilst many noun phrases found in specialised texts and consisting of two nouns are indeed terms, clearly there are many exceptions to this. Even adding statistical information about frequency of occurrence and "unithood" (the strength of association between the elements of a term) is still not completely adequate. It seems natural that the meaning of the candidate terms and that of their surrounding contexts should also be taken into consideration. It is not always easy to incorporate semantic information into such a process, however. This is partly because of the sheer enormity of information available, and partly because of the inaccessibility of some of the information which might be useful. For these reasons, the information we choose to make use of is domain-specific and pre-defined, and therefore readily available. The information is also generic rather than word-specific. Using this has the advantage that generalisations can be made about groups of words. This point will be discussed in more detail in Chapter 7, but the idea is that the more generalisations that can be made, the easier the acquisition and manipulation of semantic knowledge, since it avoids the potential for redundancy and speeds up the processing. Furthermore, the generalisations themselves are particularly useful, since we rely on them to help with the disambiguation process. This will be discussed in more detail in the next chapter.

4.3 Obtaining Contextual Information

Our aim is a knowledge-rich approach which makes use of as much readily available information as possible without extending the computational processing excessively. We therefore combine the three types of information – syntactic, semantic and terminological - described above, using a set of weights which are assigned to each context word to determine its relevance. In this section we shall explain how these weights are obtained and combined. We shall leave full details of the semantic weight till the next chapter, since this is a primary part of our work, and restrict ourselves in this chapter to a brief outline. Similarly, we shall leave it until Chapter 6 to explain fully the SNC-Value, which combines the context word weights together to give the candidate terms a ranking.

4.3.1 Syntactic weight

The syntactic category of the context word is clearly important, since we know that some categories provide more useful information than others. For example, it would be expected that a context word which is a noun would provide more information than one which is a preposition. The NC-Value does distinguish between syntactic categories, but only in a very basic way, by filtering out all syntactic categories except for nouns, adjectives and verbs. This syntactic filter can be made more open or closed according to the type of restriction we wish to place on the terms extracted. If we make the filter more open by allowing more categories to be accepted, this will allow a greater number of candidate terms to be extracted (increasing the recall) but also increases the chance of extracting false terms (decreasing the precision). Conversely, by making the filter more closed and allowing fewer categories to pass through the filter, the precision may be increased but the recall will be decreased. However, adjusting the filter in this way does not allow for very precise changes to the extraction results. Simply allowing or restricting a whole category to be used as context words is rather vague, since the importance of a context word relies on more than just its syntactic category. It also implies that all categories are equally important, which seems to be too much of a generalisation.

Calculating the weight

Each of the four syntactic categories permitted by the filter is therefore allocated a syntactic weight. This is calculated in the following way.

1. For each candidate term detected previously by the C-Value method, any boundary context words are found (the context word immediately preceding and following the candidate term). Boundary context words cannot cross sentence barriers, so if a candidate term commences or ends a sentence, there will be no context word before or after it respectively. Note that sentence internal punctuation is not considered, since it is harder to establish which types of punctuation might be relevant. Clearly some types of internal punctuation, such as bracketing, could be taken into account, although due to the style of text used in the medical reports, it is less frequently used than in many other forms of text.

2. Syntactic categories are allocated to boundary words.

3. For each candidate term, the total number of boundary words of each category is calculated.

4. Each candidate term is then checked against a list of manually evaluated terms to identify whether it is a valid term or not.

	Non-Terms		Terms		
Category	Total	Percent	Total	Percent	Weight
Verb	13147	28	16170	32	1.2
Prep	12050	29	15551	31	1.1
Noun	11500	26	12425	24	0.9
Adj	7682	17	6604	13	0.7
Total	44379	100	50750	100	

Table 4.1: Distribution of boundary words by category for terms and non-terms

5. The list of candidate terms (containing the totals of boundary words) is then divided into terms and non-terms.

6. For each list, the total number of boundary words in each category is calculated as a percentage of the total number of boundary words, to find the distribution of each category for terms and non-terms.

7. For each category, the percentage value for terms is divided by the percentage value for non-terms to give an overall weight for each category.

The results of this procedure are described below. Table 4.1 shows the frequency and percentage of boundary words of each category occurring with terms and non-terms.

The figures show two main findings. Firstly, the proportion of boundary words which are verbs and prepositions is higher for terms than for non-terms. This means that boundary words which are verbs or prepositions are more likely to occur with terms than with non-terms. The proportion of boundary words which are nouns and adjectives is lower for terms than for non-terms, indicating that boundary words which are nouns or adjectives are less likely to occur with terms than with non-terms. Secondly, the distribution of boundary words is such that for terms, the boundary word is most likely to be a verb or preposition (the former being marginally ahead of the latter), less likely to be a noun, and least of all likely to be an adjective. Note that the figures themselves are not strictly relevant: rather, it is the difference between them that is important. For example, although there is not much difference between the percentage of terms and non-terms occurring with verbs, it is the distinction between these differences for each category which is important (e.g. the fact that there is an inverse relationship for terms and non-terms occurring with verbs and nouns).

The weights we have allocated to each category using the above method reflect these findings accordingly, with verbs receiving the highest weight and adjectives the lowest. The difference between the weights matches the difference in likelihood of each category appearing as a boundary word. So the weights are designed to distinguish not just the likelihood of a boundary word of that category occurring with a term (as opposed to a non-term), but the *relative* likelihood of this, as opposed to the likelihood of a different category of boundary word occurring with a term. This means that rather than just measuring the importance of a particular category of context word as a term indicator, it measures the relative importance of that category as a term indicator.

Ideally, we should also distinguish between boundary words which occur before a candidate term and those which occur after one. It might be that some categories are better at distinguishing terms when they occur before a candidate term, but not when they appear after it, or vice versa. However, we have ignored this distinction in order to maintain consistency with the rest of our approach, which does not currently consider the position of the context word with respect to the term, other than to make the distinction between boundary word and non-boundary word.

4.3.2 Context term weight

The first parameter we consider here is the status of each context word. If a context word is also a term, we predict that it is more likely to be significant than a non-term, in other words, that terms are better indicators of other terms. This stems from the fact that terms do not tend to occur singly or randomly, but in groups, particularly in domain-specific texts. From this point, we shall refer to this type of context word as a "context term".

Since we do not know in advance which context words are terms, this step can only be undertaken in a second pass, i.e. once we have a preliminary list of candidate terms. For this we use the top third section of the list of terms extracted by the NC-value approach (Frantzi and Ananiadou, 1997), since this should contain the "best" terms (or, at least, those which behave in the most term-like fashion). A context term (CT) weight is assigned to each candidate term based on how frequently it appears with a context term. This is calculated by summing the frequency of occurrence of each context term appearing together with the candidate term. Note that this is not the same as calculating the total frequency of occurrence of each context term, for we only consider the number of times the context term appears *in conjunction with* the candidate term in question. So a context term will not have a fixed weight – the weight will vary according to the candidate term which is being considered. While this means that the weight has to be calculated for each candidate term, it does not significantly increase the processing time since the calculation is so simple. We recognise that the total frequency is also important, but it has already been measured in the NC-Value approach (for all context words, not just for context terms), and is retained as part of the overall weight. This means that for the purposes of the context term weight, we can ignore it.

The CT weight is formally described as follows:

$$CT(a) = \sum_{d \epsilon T_a} f_a(d) \tag{4.1}$$

where

a is the candidate term,

T_a is the set of context terms of a,

d is a word from T_a,

$f_a(d)$ is the frequency of d as a context term of a.

4.3.3 Similarity weight

We also predict that context words which are not only terms, but also have a high degree of similarity to the candidate term in question, are more likely to be relevant. This is linked to the way in which sentences are constructed. Semantics indicates that words in the surrounding context tend to be related, so the more similar a word in the context is to a term, the more informative it should be.

Our claim is essentially that if a context word has some contribution towards the identification of a term, then there should be some significant correspondence between the meaning of that context word and the meaning of the term. This should be realised as some identifiable semantic relation between the two. Such a relation can be exploited to contribute towards the correct identification and comprehension of a candidate term.

Since it is such an integral part of the extraction process, we shall not dwell on the similarity weight further here, as the next chapter is devoted solely to this purpose.

4.4 Summary

In this chapter we have examined some of the ways in which contextual information can be used to help identify terms. In particular, we focused on ways of distinguishing those parts of the context which are most relevant. We have outlined a measure for incorporating this information into the overall term extraction procedure, by combining information from different elements of the context to produce a weight for each candidate term. In the next chapter we shall take a closer look at the main part of this, the Similarity Weight, which aims to ascertain the informativeness of different context words using semantic information.

Chapter 5

Measuring Similarity

5.1 Introduction

In this chapter, we first examine some of the resources available to us for extracting semantic information. We look both at theoretical notions of ontologies, and at existing lexical resources. We then describe the approach we take to extracting semantic information from the UMLS and measuring similarity between term and context, in order to produce the semantic weights for the SNC-Value Measure.

5.2 Methods for Extracting Semantic Information

Although semantic information has been rarely used within term extraction, it has been extensively used for other purposes such as knowledge acquisition and word sense disambiguation. There are two common approaches to the extraction of semantic information: corpus-based, statistical approaches; and knowledge-rich, linguistic approaches involving external resources such as dictionaries or thesauri.

5.2.1 Corpus-based approaches

The first type of approach involves the use of a disambiguated corpus for training. Methods used vary considerably, but tend to be statistical in nature. Riloff and Lehnert (Riloff and Lehnert, 1992) developed an algorithm to derive relevancy cues from training texts, which they used for information extraction. CRYSTAL (Soderland et al., 1995) is a system which identifies concepts in a text, by means of linguistic features which reliably identify the conceptual content of a phrase. SEXTANT (Grefenstette, 1994) is a hybrid, knowledge-poor approach to the automatic extraction of semantic information from large corpora, using only syntactic information.

Corpus-based approaches have the advantage that they are self-contained, since they do not require any external resources. This means that they can be used not only for general language, but also for terminology, since they can be tailored specifically to the domain. As with any statistical method, they work best with large-scale corpora, which are not always readily obtainable in specific domains. They also ensure that the information will be appropriate to the domain, although there is no guarantee that this information will be accurate or unambiguous, so it may also be misleading. While the possibility of being able to use statistical information from the corpus is appealing, it may again cause skewing problems. Whereas a pre-defined lexical resource is likely to be more or less complete, there may be gaps in the coverage of a corpus. Lack of information in a dictionary tends to be less problematic, because it is easier to identify and may be manually (or otherwise) rectifiable. Table 5.1 outlines some of the main advantages and disadvantages of using a corpus-based approach to semantic acquisition.

Advantages	Disadvantages
– real examples of language – information tailored to domain – statistical information available	– techniques may be unreliable – coverage may be inadequate – large corpus may be needed – appropriate information may be unavailable – information may be ambiguous

Table 5.1: Advantages and disadvantages of corpus-based approaches for acquiring semantic information

5.2.2 Dictionary-based approaches

Because they are not used specifically for terminology, these methods tend to use general language dictionaries or thesauri. Usually, semantic information is extracted from the dictionary definition of a word. This may then be compared with the words in the surrounding context in the corpus. For example, Lesk (Lesk, 1986) used an online dictionary to calculate the overlap between dictionary definitions of word senses; Smeaton and Quigley (Smeaton and Quigley, 1996) calculated word-word similarity between related words in a thesaurus; while Yarowsky (Yarowsky, 1992) captured the meanings of word senses from semantic categories in Roget's Thesaurus.

Dictionary methods have proved to work well in applications such as word sense disambiguation. However, they are not so appropriate for use in terminology, because they are designed to deal with general language. This has two main implications. Firstly, such techniques are not really suitable for terminology because of the different nature of terms, in particular multi-word terms, as discussed in Chapter 2. Secondly, the lack of suitable resources creates a problem, since technical terms are not covered sufficiently in any general language dictionary or thesaurus. Terminological applications almost always require that their lexical resources be tailored to the domain. On the other hand, there is a lack

Advantages	Disadvantages
– accuracy of information	– information may be too general
– even and neutral coverage	– few specialised dictionaries available
– unambiguous information	– type of information unsuitable
– information readily available	– bound by subjectivity of lexicographer

Table 5.2: Advantages and disadvantages of dictionary-based approaches to semantic knowledge acquisition

of suitable domain-specific resources available. The field of medicine is unusual in that there are various existing ontologies, although these may still not be specialised enough to suit any individual corpus. Finally, using a lexical resource means that the application is constrained by the subjectivity of the compiler or lexicographer. This is one of the main reasons for the current trend towards the design of automatically-generated dictionaries and thesauri. Table 5.2 illustrates some of the main advantages and disadvantages of a dictionary-based approach to semantic knowledge acquisition.

5.3 Ontologies, Dictionaries and Thesauri

5.3.1 Theoretical notions of classification systems

Any proposed system of concept classification or ontology must be based on some solid theoretical foundations. From a cognitive science perspective, a variety of views exists concerning the formation of concepts and the relationships held between them. Keil (Keil, 1979) proposes a rigid hyponymic structure similar to those used to store terminological knowledge. Rosch (Rosch, 1978) proposes a more flexible structure, in which there are varying degrees of relationship between an item and a concept, depending on the similarity between that item and a prototype. Dahlgren (Dahlgren, 1988) also prefers this flexible, prototypical approach over a hyponymic hierarchy, but allows cross-classification in her ontology, using a feature-based classification system.

Lauriston discusses these theoretical approaches in more detail in his PhD thesis (Lauriston, 1996). He proposes the following set of criteria for a classification system used to attribute meaning to lexical units:

- it should apply equally to nouns, verbs, adjectives and adverbs;
- it should focus on the meaning of the individual units rather than any relations between them;
- there should be 3 major classes – Entity, Activity and Property – which should not be restricted to any specific syntactic category;
- it should be universally applicable to all domains;
- categorisation of concepts should take the form of general groupings rather than fine-grained distinctions;
- it should be based on psycholinguistic theory;
- it should allow for imprecision and systematic ambiguities;
- it should use groupings from an established source of reference;
- it should be validated on a real corpus.

Lauriston's proposed concept system uses a feature system based on "quaternary semantic features" rather than the traditional binary or multi-valued features. He thus has a set of four possible feature values: true, false, indeterminate and inapplicable. The "indeterminate" value captures generality, i.e. the concept is not specific enough to be allocated a value for the feature (e.g. it cannot be determined whether "person" is male or female). The "inapplicable" feature captures ontological knowledge in the sense that the feature does not apply to that concept (e.g. a book cannot be male or female[1] – it is not just unknown, but impossible).

For the field of telecommunications, Lauriston defines a basic set of 14 semantic features:

- material / abstract
- temporal / state or attribute
- collective / individual
- natural / social or man-made
- goal / non-functional
- living / non-living
- animate / inanimate
- intellectual / irrational
- symbolic / non-symbolic
- positional / non-positional
- relational / non-relational
- quantity
- quality
- colour

[1] in English – in other languages, inanimate objects such as books may be allocated genders

It appears, however, that not all of the features listed really have the capacity to be quadrivalent, especially the latter ones. The difference between something not having a colour and a colour feature being inapplicable to the concept, for example, seems negligible.

5.3.2 General vs. domain-specific ontologies

It is clear that any theoretical basis on which ontologies are founded is inclined to be general. However, bearing in mind the need for domain-specific ontologies, it is important to specify the theoretical foundations on which they should rely. In the same way that special languages are not simply a subset of general language, domain-specific ontologies cannot simply be a subsection of general language ontologies (although they may contain elements from them).

Although Lauriston's criteria for a classification system recommends that it should be applicable to all domains, this is not really practical. Different subdomains really require individual ontologies. For example, Lawson (Lawson, 1997) specifies the following "classes of vocabulary" for the subdomain of patenting:

- Construction
- Function
- Substances
- Geometry, Motion and Position
- Degree
- Properties of Invention or Prior Art
- Names etc.

While all classification systems such as these are subjective, it is clear that although this may be perfectly suitable for the field of patenting, it would be of little use within the field of medicine. Similarly, the following top-level structure of the concept system presented in the UMLS Semantic Network is suited to the medical field, but would not be applicable to patenting:

- Entity
  - Physical Object
  - Conceptual Entity
- Event
  - Activity
  - Phenomenon or Process

This system is separated into two main concepts: entity and event. In some domains, such as patenting, "event" would not be classed as a primary concept, although for medicine it is clearly relevant.

It is not only the categories of concepts, but also the entire structure of the ontology, that differs between these two approaches. It is hard to make a branching structure out of Lawson's primary concepts, since they are all rather different, but the field of medicine, as presented in UMLS, can more easily be divided into two main sections – Entity and Event - or "objects" and "actions". Beneath this lie the various secondary concepts: physical and conceptual types of entity, and planned and unplanned types of event (activities and phenomena). In this way, the main concepts of Lawson's patenting ontology will form a shallow, flat structure, whereas the main concepts of the UMLS ontology will form a deeper, branching structure, as depicted in Figures 5.1 and 5.2.

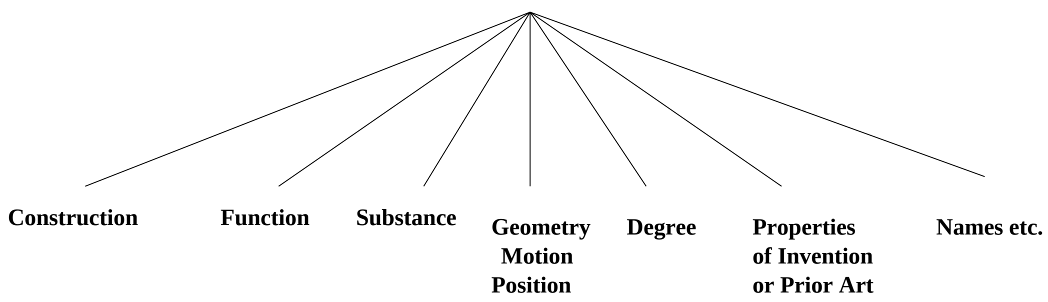

Figure 5.1: Lawson's primary concepts for the field of patenting

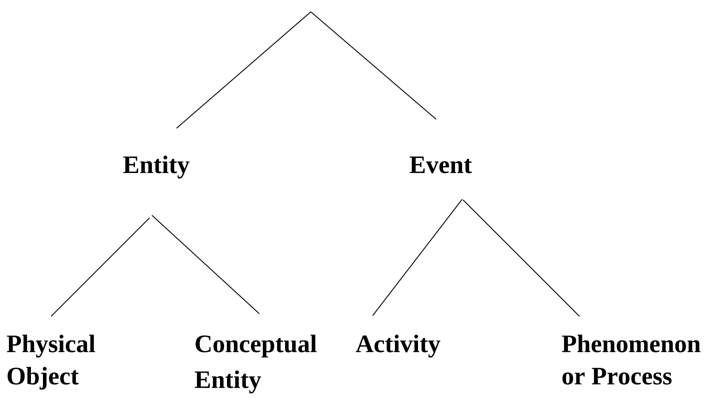

Figure 5.2: UMLS Semantic Network's primary concepts for the field of medicine

5.3.3 Lexicalisation

Works in lexical semantics generally agree that there exists an underlying set of universal meaning components which can be realised in different ways (or left unrealised) in different languages. Conceptual systems may appear to vary according to the language (or even sublanguage), but this could be explained by the fact that these are only surface conceptual systems. In other words, there is still a universal deep set of components which is realised in each language or sublanguage by a different set of concepts, each of which is then realised by different lexical forms. This allows for the universality of meaning, whilst accepting the individuality of both surface linguistic forms and conceptual systems. To what extent conceptual systems differ between languages is largely debatable, but it is clear (as seen in fields such as machine translation) that there is not always a one-to-one correspondence between concepts in different languages.

The problem of realising concepts in different languages is typified by the design of the EuroWordNet project (Vossen et al., 1997), (Vossen, 1997), a multilingual thesaurus which links the individual wordnets of the 8 languages involved (English, French, Dutch, Spanish, Italian, German, Estonian and Czech). The main aim of the EuroWordNet project is to allow non-expert users easy access to the constantly expanding areas of electronic information, in a variety of European languages, by means of concepts rather than specific words. Whilst the English wordnet already exists (Miller et al., 1990), the wordnets for the other languages have had to be created from scratch. Each is composed of synsets (conceptual units) in the same format, linked by semantic relations, and each meaning is also linked via equivalence relations to a WordNet1.5 synset. There are several other possibilities for the structure of such a system. Theoretically, a universal conceptual system to which the lexicalised wordnets of each language are linked would seem to be ideal. Practical considerations must, however, be taken into account:

for example, when new words are added to one or more of the languages, the whole conceptual structure may have to be altered.

5.3.4 General language resources

We now turn our attention to some of the most popular lexical resources which might be considered for semantic knowledge acquisition. As we have discussed, general language resources are not really sufficient for terminological applications, but we may gain some insight into the properties a suitable specialised resource should contain. More detailed comparisons of lexical resources can be found in (Litkowski, 1999; Wilks, 1997; EAGLES Computational Lexicons Working Group, 1996).

Lexical resources can take the form of dictionaries, thesauri, or other ontologies. The difference between them is not always evident, but is important nevertheless. An ontology (in the field of AI) can be defined as:

> "An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them." (Howe, 1998)

In general, thesauri contain implicit definitions and explicit synonyms; monolingual dictionaries contain implicit synonyms and explicit definitions. Dictionaries are also ordered alphabetically, while thesauri have a hierarchical structure (although some modern American thesauri are now ordered alphabetically). The choice between dictionary and thesaurus is largely governed by the application. As early as the 1950s, Masterman argued in favour of a thesaurus, such as Roget's Thesaurus, for NLP, because of its strong underlying structure which she believed was more compatible with the way in which humans process language (Masterman, 1956).

LDOCE

The Longman Dictionary of Contemporary English (Proctor, 1979) is a learner's dictionary, i.e. one designed for non-native English speakers. It is probably the most commonly used dictionary in NLP tasks, and is particularly favoured for work on word sense disambiguation, e.g. (Lesk, 1986; McRoy, 1992; Cowie et al., 1992; Bruce and Wiebe, 1994; Wilks and Stevenson, 1997). This is largely because it has the special feature of subject codes which indicate particular subject areas. In this way, different senses of a word can be tagged and thereby easily distinguished. For example, the word *wind* has the following subject codes marking its different meanings:

```
DZP    physiology
HFZH   hunting
ML     meteorology
MU     music
NA     nautical
```

LDOCE has also been used for tasks such as text retrieval by classification (Walker and Amsler, 1986), sense linking (Krovetz, 1992), and automated lexicon construction (Wilks et al., 1996).

WordNet

The Princeton WordNet 1.5 (Miller et al., 1990) differs from traditional thesauri in that it stores semantic information about *synsets* (conceptual units) rather than about individual lexical items. This semantic information is also entirely based on relations, rather than features or attributes. Many of the conceptual classes are artificial and thus non-lexicalised. WordNet has been used in a number of projects for word sense disambiguation, e.g. (Resnik, 1995a; Dorr and Jones, 1996), and for various other forms of knowledge acquisition such as conceptual indexing

(Stairmand and Black, 1996) and information extraction (Chai and Biermann, 1997).

Roget's Thesaurus

Roget's Thesaurus is also organised in the form of a semantic hierarchy, with words arranged in groups of categories. Links may be established between neighbouring categories. There is also a cross-referencing link, where a word in one category is also found in another category, thereby supporting multidimensionality. Roget's Thesaurus has been used for word sense disambiguation, e.g. (Yarowsky, 1992), and also for lexical chaining[2] (Morris and Hirst, 1991). Ellman discusses some of the difference between Roget's Thesaurus and WordNet as a resource for lexical chaining (Ellman, 1997), preferring the simplicity of the former over the wider coverage of the latter. Wilks also argues in favour of Roget's Thesaurus (Wilks, 1997), claiming that, although WordNet has become one of the most popular lexical resources for NLP, the abstract classification system of Roget's Thesaurus is superior (for this purpose) to the lists of partial synonyms found in WordNet.

5.3.5 Specialised resources

The field of medicine is one of the few subdomains which is relatively rich in lexical resources, largely because there is a huge demand for terminological applications and for standardisation in this field. According to Lynch:

> "A medical terminology entry should be organised for optimum reference utility, readability, interchangeability, and flexibility, using a terminology database management software package that offers a range

[2] This can be used for a number of tasks such as word sense disambiguation, text summarisation and information retrieval.

> of record layout options and the appropriate manipulation capabilities for the task at hand." (Lynch, 1997)

SNOMED

The Systematised Nomenclature of human and veterinary Medicine (Côté, 1993) is an international classification system comprising 11 different modules (e.g. Diseases/Diagnoses, Living Organisms). It contains over 144,000 terms, each of which is assigned a code, allowing relations to be formed between terms. A basic hierarchy is thereby defined, enabling the identification of synonyms, hyponyms and hypernyms. The terms are not restricted to a single conceptual representation, and may be cross-referenced so that they appear in more than one section of the hierarchy.

UMLS

The Unified Medical Language System (NLM, 1997) also contains information about medical terminology, organised in a hierarchical structure. The UMLS is unusual in that it consists of a *set* of resources, which comprise four main components:

- Metathesaurus
- Semantic Network
- Source Vocabularies
- Specialist Lexicon

Not only does it provide a classification system for the terms, but it also contains morphological, syntactic and semantic information.

The **Metathesaurus** contains semantic information about the terms that appear in various controlled vocabularies and classifications, such as SNOMED

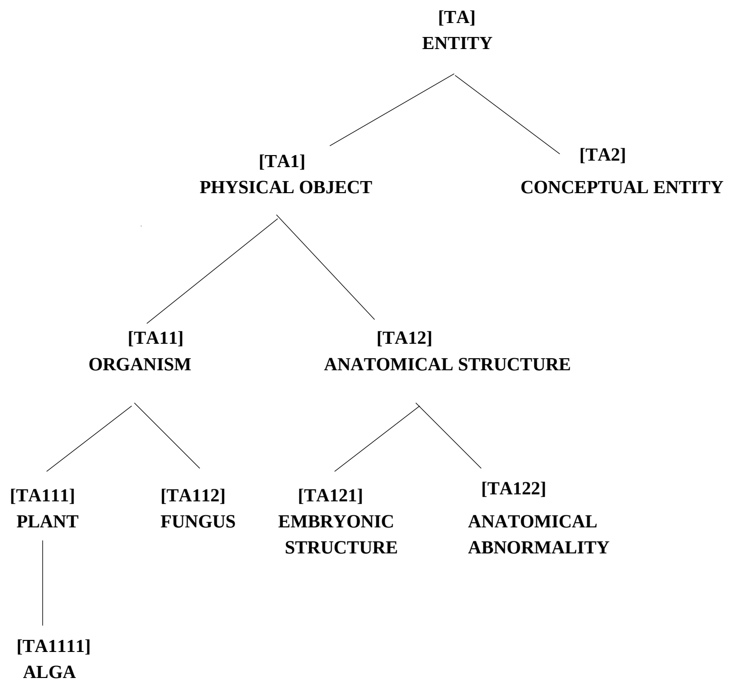

Figure 5.3: Fragment of the UMLS Semantic Network

and MESH. It lists over 330,000 concepts and over 735,000 terms, from over 30 sources. The meanings and relationships are preserved from the source vocabularies, but some additional information is provided, and new relationships between concepts and terms from different sources are established.

UMLS makes a distinction between *concepts* (meanings) and *concept names* (terms). The Metathesaurus is organised by concept, establishing relationships between alternate names and views of the same concept, and between concepts themselves. The representation takes the form of three levels: a set of general concepts, represented by codes known as CUIs (Concept Unique Identifiers); a set of concept names, represented by codes known as LUIs (Lexical Unique Identifiers); and a set of strings, represented by codes known as SUIs (String Unique Identifiers). A section of the concept hierarchy is depicted in Figure 5.3. The Metathesaurus identifies various types of relationships between concepts, including synonyms, hyponyms, hypernyms, part-whole relations, and various "undefined relations".

The **Semantic Network** contains information about the set of basic semantic categories that may be assigned to the concepts, and about the relationships between these categories. In Version 4, there are 135 semantic types (represented as nodes) and 51 relations (represented as links). The main relation is that of hyponymy, but there are also 5 main types of non-hierarchical relations: physical, spatial, temporal, functional and conceptual. Relational information is generic rather than specific, i.e. details are provided about semantic types rather than about individual concepts, and relations do not necessarily hold for every concept belonging to that type. For example, a relationship exists between the semantic classes of *disease or syndrome* and *acquired abnormality*, such that the former is a *result_of* the latter. This does not imply that any disease is the result of any acquired abnormality, but simply that there is a general relationship of this kind between the two classes, such that diseases or syndromes can be the result of acquired abnormalities.

The **Specialist Lexicon** provides detailed syntactic information about both biomedical terms and common English words. An individual lexical entry is created for each spelling variant and syntactic category of a word. These are then grouped together to form a unit record for each word, defined in a frame structure consisting of slots and fillers. Full morphological and syntactic information is provided, such as syntactic category, number, gender, tense, adjectival type, noun type, etc. as appropriate. Version 4 of the Knowledge Sources only contains an experimental version of the Specialist Lexicon, though it has since been more fully completed.

The **Information Sources Map** contains details of the original sources of the terms. It consists of a database of records describing the information sources, with details such as scope, utility and access conditions. The sources themselves are varied and include bibliographic databases, factual databases and expert systems.

Other ontologies

The READ codes, or "Clinical Terms" version 3 (NHS Information Authority, 1999a), were developed in the 1980s by the NHS to enable electronic patient records to hold clinical concepts in a concise and unambiguous way. They later became the UK General Practice Standard, and are used in many primary care patient record systems. Each term represents a clinical concept, which can range from the very general to the very specific.

The READ system has a concept-based file structure and is represented as a directed acyclic graph (DAG). It is similar in structure to SNOMED, in that it has explicit relational tables, multiple parentage, and many levels of granularity. It largely differs from other ontologies in content – for example, the same terms are used, but can mean different things and can be related to different children or parents. A project is being undertaken to combine the READ codes with S-NOMED, to create the "SNOMED Clinical Terms" (NHS Information Authority, 1999b), but this is still currently under development.

5.4 Using Semantic Information from UMLS

The reasons we choose to use UMLS as our source of semantic information are multiple. Firstly, the relationship it expresses between terms and concepts allows us to exploit the multidimensional nature of terms. We shall discuss this in more detail below. Secondly, it is structured in such a way that it is very easy to manipulate. Thirdly, it is the most comprehensive of the medical resources considered, since it not only contains all the SNOMED terms, but also information from other sources. Finally, it is easy to obtain for research purposes, and updates and backup support are available. At the time of commencing our research, Version 4 was the latest available, and so was used for our experiments, although updated versions have since been released. We did not consider that the differences between versions were significant enough to necessitate upgrading to these later editions. We initially make use of two of the four components of UMLS: the Metathesaurus and the Semantic Network. The Metathesaurus is used to associate candidate terms found in the text with their semantic type, and the Semantic Network is then used to measure similarity between semantic types.

The UMLS Semantic Network differs from traditional thesauri in two main ways. Firstly it is non-lexicalised, providing a hierarchical structure of concepts (or semantic types), rather than terms. Files are provided in the Metathesaurus which link these concepts to the terms they represent. It is also designed as an information resource rather than a lexical resource in this respect, since the nodes can be viewed as representations of language, rather than lexicalised natural language expressions. This issue is discussed further in the concluding chapters of this thesis. The second difference is partly linked to this, in that the entries in the UMLS are derived from a number of different source vocabularies. The resulting ontology is thus a conglomeration of meanings, or concepts, from a variety of viewpoints. The implication is that the concepts represented can only really be interpreted in their *extensional* meaning (whereas the source vocabularies from which they are derived provide the *intensional* meaning of the concepts (Campbell et al., 1998)).

Campbell et al. propose that the concept in UMLS therefore holds a special relationship with the meaning it represents, in that it

> "... takes on the emergent meaning through linkage to terms in different terminology systems" (Campbell et al., 1998).

This emergent meaning may not be the same as its original meaning in those

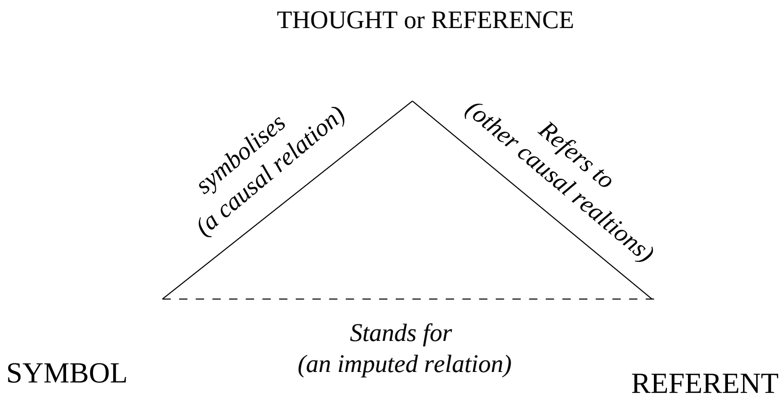

Figure 5.4: The Ogden-Richards meaning triangle (taken from the original 1923 illustration)

systems. In this way, the UMLS Semantic Network can be seen as a *possible world*:

> "... a mathematical formalism that represents propositions about some perspective or interpretation of the physical world" (Campbell et al., 1998).

They continue this idea by showing how the UMLS can be viewed in terms of Ogden and Richards' *meaning triangle*, illustrated in Figure 5.4. They explain the semantics of the thesaurus by means of the two different types of meaning labelled by Frege as *Sinn* and *Bedeutung* (Frege, 1892), and usually translated as *sense* and *referent*. The Ogden-Richards triangle emphasises the point that there is not always a straightforward correspondence between a concept and its linguistic realisation, as we have discussed in Chapter 2. It indicates that the preferred route from one to the other is via a thought or reference, and thus that the transition from one to the other is a subjective one. The dotted line from symbol to referent in the diagram suggests that some kind of direct correspondence between the two is possible, but that it is far from absolute, and entails much hidden complexity.

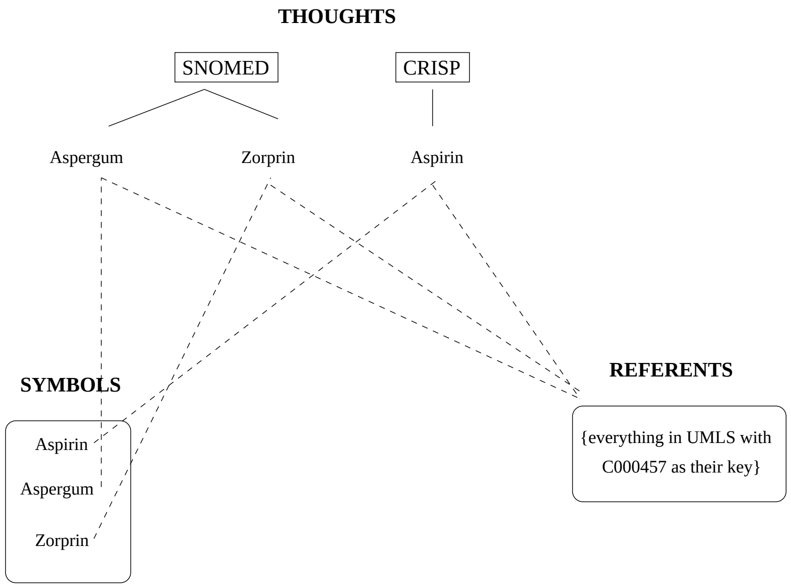

Figure 5.5: Understanding the semantics of "aspirin" using the UMLS and the Ogden-Richards framework (adapted from (Campbell et al., 1998))

Figure 5.5 is a simplified version of Campbell et al.'s diagram which demonstrates how the UMLS represents semantic ambiguity. The concept C0004057 is labelled *aspirin*, but a variety of terms is used to describe the same concept. This is partly because in some of the source vocabularies, there is more than one term for that concept (as in SNOMED), and partly because more than one source vocabulary is used, and the term relating to that concept may differ according to the source (e.g. CRISP and SNOMED label it differently). The diagram shows how the intensional meaning of the symbol depends on its originating source, but also demonstrates that what each symbol designates is the extensional meaning of all the symbols which are labelled in UMLS as referring to the concept C0004057 (aspirin). Because there is no direct path from symbol to referent, the argument is that we cannot link one symbol to one representation in the UMLS, such that each one can be distinguished from the others. Instead, we must follow the path via the "thought" or intensional meaning, taking into account the intended

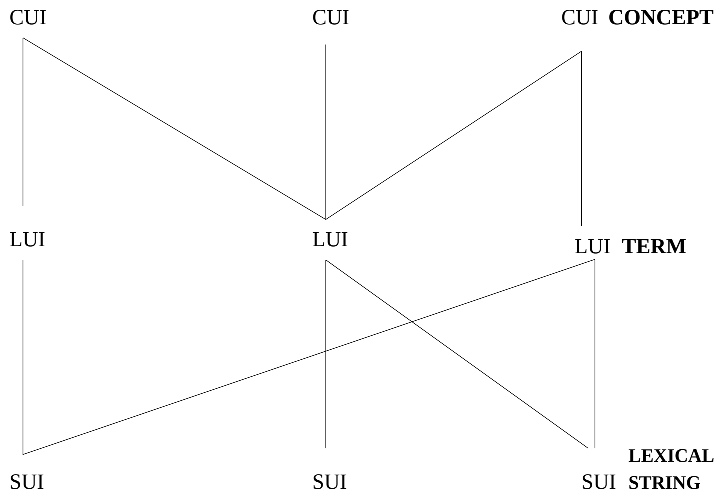

Figure 5.6: Relations between strings, terms and concepts in the UMLS

meaning of each symbol in the source vocabularies.

What practical implications does the UMLS hold for us as a "possible world thesaurus"? Because it treats the relationship between terms and concepts in a unique way, this influences our use of it to help understand the meaning of terms. It might be argued that the UMLS conceptual system is far too complicated to exploit fully, but we reason that rather than seeing this as an impediment, we can make use of its complexity. Although we do not make specific use of the source vocabularies from which the thesaurus is compiled, by recognising that the "path to meaning" lies through this route (in line with the Ogden-Richard framework), rather than directly from term to concept, we refrain from forging univocal correspondences between term and concept, as other thesauri and ontologies force us to do. In this way, we are able to more fully exploit the multidimensional nature of terms.

Figure 5.6 depicts the structure of the UMLS as a whole, showing how the lexical string is linked to the terms, which are then linked to the concepts, and how ambiguity and variation can occur. Figure 5.7 depicts these same relationships, but from a different perspective, highlighting the difference between lexical and

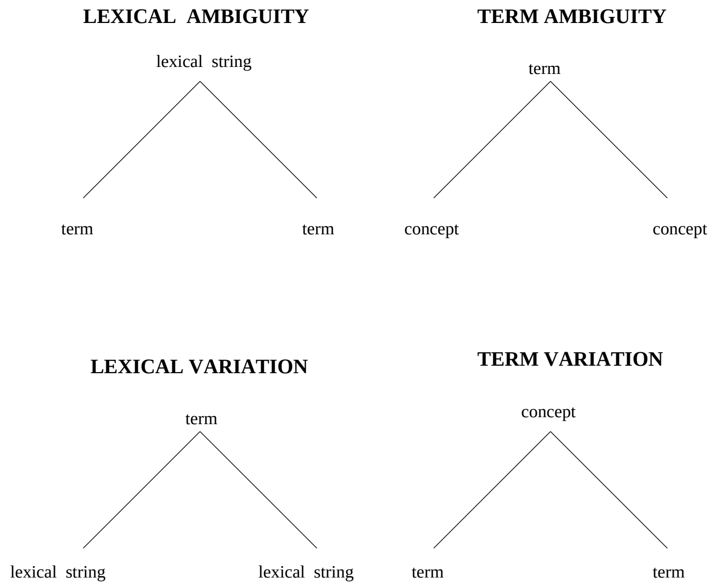

Figure 5.7: Ambiguity and variation in the UMLS

term ambiguity and variation. Table 5.3, taken from the UMLS documentation (NLM, 1998) gives a more concrete example of this. The string *atrial fibrillation* and its plural *atrial fibrillations* have different SUIs (string identifiers) but are both linked to the same LUI (lexical or term identifier). The string *auricular fibrillation* and its plural *auricular fibrillations* are linked to a different LUI, but both LUIs are linked to the same CUI (concept identifier) since they are considered to have the same meaning.

In Appendix F, we also give some examples of the different ways in which the term *epilepsy* is represented in the UMLS. We find 228 variants, matching 67 different concepts, thereby demonstrating both lexical variation and ambiguity.

Concept (CUI)	Term (LUI)	String (SUI)
C0004238 (preferred) atrial fibrillation atrial fibrillations auricular fibrillation auricular fibrillations	L0004238 (preferred) atrial fibrillation atrial fibrillations	S001668 (preferred) atrial fibrillation S0016669 (plural variant) atrial fibrillations
	L0004327 (synonym) auricular fibrillation auricular fibrillations	S0016899 (preferred) auricular fibrillation
		S0016900 (plural variant) auricular fibrillations

Table 5.3: Relationship between *auricular fibrillation* and *atrial fibrillation* in the UMLS

5.5 Methods for Measuring Similarity

As we have discussed, there are two major criteria to consider when deciding on a method of measuring semantic distance in this way. Firstly, we must decide whether to use a corpus-based approach or a dictionary-based approach[3] for extracting semantic information. Secondly, we must decide on a method of measuring the similarity. These two criteria are co-dependent. Making the choice between corpus and dictionary restricts our choice of similarity measure, and choosing the similarity measure restricts our choice of semantic extraction method. Additionally, if we choose, for example, a method of measuring similarity which uses a pre-defined hierarchy, we must ensure that the dictionary we choose for extracting semantic information has a suitable hierarchy which can be used for measuring semantic distance.

Word similarity measures have been used in a range of applications such as example-based machine translation (EBMT), information retrieval, information

[3]We use *dictionary* here in its broad sense to cover dictionaries, thesauri and other ontologies.

extraction and word sense disambiguation. Although the purposes for which they are used differ, the main idea behind them is to find words or phrases which are closest in similarity to an input word or phrase. Similarity may be measured according to different features: for example, it might take into account syntactic category, subcategorisation information, semantic features or even pragmatic features. So when choosing a measure, we have to consider not only which type of method should be used, but which type of features should be considered. It is also important to establish other factors such as how fine-grained a measure is required (i.e. how similar words need to be to constitute a match), and how the results need to be classified (e.g. whether they need to be ranked in order of preference, or whether a Boolean classification is sufficient).

Methods can be broadly divided into three main types: EBMT-based methods, IR-based methods, and thesaurus-based methods, although there is a certain amount of intersection between them. For example, some of the EBMT-based methods involve a thesaurus hierarchy.

5.5.1 EBMT-based methods

EBMT aims to select the most similar experience for a given problem, by defining a set of relevant experiences and comparing them with the problem in terms of their context and configuration. These methods usually combine hierarchical structures with a database of examples taken from a corpus.

Weighted feature graphs

Zhao's Analogical Translator (Zhao, 1996) computes similarity using *weighted feature graphs*. Each feature graph describes the properties of a linguistic object. Linguistic objects are then compared in terms of their feature sets. The similarity of the feature graphs is computed by combining the values of the similarity of each feature pair in the graph.

Feature similarity is calculated by determining the "least generic common denominator" of two features in the general feature graph, and calculating its combined weight. This combined weight consists of two factors: position and importance. The *positional* weight is the distance between the root and the vertex in question. The greater the distance, the more concrete the feature and the heavier the weight. The *importance* weight depends on the type of feature in question. Different features have a different effect on similarity; for instance, predicate verb similarity is considered more important than modifier similarity. The importance weight is determined from the average weight of the dominating vertices, but can be fine tuned or overridden manually if necessary. The main features used are syntactic, but semantic and pragmatic features can be accessed if necessary, for finer classification. The feature indexes can also be used in parallel rather than hierarchically.

Example-based word sense disambiguation

Tsutsumi (Tsutsumi, 1993) performs word sense disambiguation using example sentences in which the words have been disambiguated, and a set of taxonym and synonym hierarchies. No special semantic categorization is needed, and even closely related meanings can be disambiguated. The plausibility of each interpretation of a sentence is calculated by taking the sum of all the closeness values between the input sentence and the example sentences. A measure of *word closeness* is established for each pair of words, based on the taxonym and synonym hierarchies. For example, two words which are synonyms will have a score of 70, whilst a one-level superordinate relation between two words will have a score of 65. Sentence closeness is then calculated, based on a weighted score of closeness between subjects, objects, verbs and modifiers, obtained from the word closeness scores.

Example-based weighted semantic distance

Sumita and Iida's EBMT system (Sumita and Iida, 1991) uses a manually created thesaurus to calculate the semantic distance between input and example words. Distances are calculated according to the word's attributes, which (for nouns) consist of the lexical subcategory, the existence of a prefix or suffix, and its semantic code in the thesaurus. Two measures are used: the attribute distance and the weight of the attribute. The attribute distance is determined by the MSCA obtained from the thesaurus hierarchy. The MSCA, or Most Specific Common Abstraction (Kolodner, 1993), is the lowest common denominator of two nodes, i.e. the lowest node which is an ancestor of both nodes. If the thesaurus has n layers, *(k/n)* is assigned to the concepts in the kth layer from the bottom. (i.e. the unique top node would have a value of 1 and the bottom nodes a value of 0.). The weight of the attribute is the degree to which the attribute influences the selection of the translation pattern. For example, if a particular attribute is always translated by another unique attribute in the corpus, this attribute will have a weighting of 1, since it greatly influences the selection of the translation pattern. The total distance between the input and example is calculated by multiplying the total attribute weight by the total attribute distance for all attributes.

5.5.2 IR-based methods

IR matching techniques differ slightly from EBMT ones in that they were originally designed for matching an input word or phrase (or a set of these) with relevant documents, rather than with other words or phrases. However, some of the matching techniques used have been successfully adapted to the task of matching two words or phrases. They tend to be frequency-based, which may, however, be more appropriate for IR, where very large data sets tend to be used.

Weighted Jaccard measure

This approach (Grefenstette, 1994) is similar to Zhao's in that it compares words based on their linguistic features. He uses the syntactic features of the context to judge word similarity. The *binary* Jaccard measure between two words is the number of shared attributes divided by the total number of attributes. The *weighted* measure uses attributes which have been assigned global weights between 0 and 1, depending on how many different objects the attributes associate with and how often they appear. It is unclear, however, exactly how these weights are assigned.

Co-occurrence techniques

Rigau et al. (Rigau et al., 1997) performed a set of experiments to test various heuristics for word sense disambiguation, combining different types of knowledge. They used two main information retrieval-based techniques for word matching, which both depend on frequency:

- **Simple co-occurrence**: the co-occurrence weight between two words can be given by Co-occurrence Frequency, Mutual Information or Association Ratio. Given a hyponym definition and a set of candidate hypernym definitions, this method selects the candidate hypernym definition which returns the maximum score given by the co-occurrence weight of the two words.

- **Co-occurrence vectors**: The similarity between two definitions can be measured by the dot product, the cosine function or the Euclidean distance between two vectors which represent the contexts of the words presented in their definitions. The vector for a context is calculated by adding the co-occurrence information vectors of the words in the definition (found using simple co-occurrence as described above).

Dictionary-based inverse sense frequency

Braden-Hader uses multiple sources of information from LDOCE and other dictionaries for word sense disambiguation, borrowing weighting strategies from information retrieval (Braden-Hader, 1993). She uses vector similarity measures to determine the closeness of the sense entry and target word. The advantage of this is that senses can be ranked according to relevance, and that sense distinctions can be based on the level of granularity required for a particular application.

The features used are based on linguistic structure (from explicit LDOCE codes) and semantic relationships (synonymy and hypernymy). The weighting measure is called *inverse sense frequency* and is similar to the IDF (inverse document frequency) measure in information retrieval, based on the number of entries in the dictionary which contain a particular cue (e.g. nouns classed as abstract in LDOCE). The similarity of 2 vectors is calculated using a basic cosine correlation (the angle between 2 vectors in a space).

5.5.3 Thesaurus-based methods

These methods rely heavily on a hierarchically-structured thesaurus. The nodes are generally (though not always) annotated with weights based on frequency or probability. The notion of the MSCA is also commonly used.

Conceptual distance

Another heuristic tested by Rigau et al. involved the simple measure of conceptual distance (Rigau et al., 1997). This uses a structured hierarchical net as the basis for determining closeness in meaning among words. The conceptual distance is the length of the shortest path that connects the concepts in the hierarchy. Surprisingly perhaps, this method produced the highest precision out of all 8 tested, although the recall figure was low.

Semantic distance

Smeaton and Quigley (Smeaton and Quigley, 1996) approximate the similarity between two words from the **information content** of the first class in the hierarchy that subsumes both classes, i.e. the MSCA. The information content is calculated by estimating the probability of occurrence of the class in a corpus. In this case, it was computed from an analysis of the frequency of occurrence of WordNet terms in a large corpus.

Sense proximity

In the CRISTAL project (Magadur and Tabuteau, 1996), disambiguation is carried out based on the proximity of sense using the Dicologique[4] conceptual dictionary. The similarity measure used is a calculation of the semantic distance between word senses. The basic semantic distance between two senses is the total number of concepts of the same type which are ancestors of one of the two words but not both. If the semantic distance is below a certain threshold for an interpretation, the two words are mutually disambiguated by this interpretation.

Syntactic dependence

This algorithm developed by Lin relies on the intuition that two different words are likely to have similar meanings if they occur in identical local contexts (Lin, 1997). Similarity between two words is measured by the ratio between the amount

[4]Dicologique is a trademark of the French company Memodata

of information needed to state the commonality of A and B and the information needed to describe fully what A and B are. Generally speaking, the commonality of A and B is that each belong to the maximally specific superclass of both A and B. Each node in the hierarchy is given a probability that a randomly selected word belongs to that node. So, the higher up the tree, the greater the probability that a word belongs to that node.

Information content

The measure of *semantic distance* (the length of the shortest distance between two words), does not take into account the fact that most networks are non-uniform and that some parts of the hierarchy are more detailed than others. Resnik deals with this problem by extending the algorithm and associating probabilities with the concepts in the hierarchy (Resnik, 1995b). Probabilities are simply assigned with respect to frequency in a large corpus. The basic notion is that the further up the hierarchy a word is, the greater the probability of encountering an instance of that word. So if there is a unique top node in the hierarchy, it will have a probability of 1, and an information content of 0. The more information two concepts share, the more similar they are, and thus the greater the information content of the concept that subsumes them.

Best match algorithm

Uramoto uses a database of example sentences as a means for word sense disambiguation (Uramoto, 1994). Word similarity is measured in terms of conjunctive relationships, since the elements of coordinated structures are supposed to be good indicators of the interchangeability of words. If no exact matches are found amongst the examples, a best-match algorithm, strongly based on the thesaurus, is used. A candidate set of words which are interchangeable with the input word is searched for in the example base, and a preference value is attached to each example found. This is calculated by multiplying the certainty value of each candidate by its frequency. The certainty value is calculated by 1/(the number of candidates for resolving the ambiguity).

5.5.4 Summary of methods

One of the problems with measuring word similarity is that it can be difficult to compare the different algorithms. As mentioned before, the choice of measure depends largely on the type of information available and the intended use of the algorithm.

Thesaurus-based methods seem a natural choice for our task, because to some extent they already define relations between words. However, a distinction needs to be drawn between the problems of word similarity and concept similarity. One word can represent different concepts which will thus appear more than once in the hierarchy. In measuring word similarity, Resnik notes that it is really the relationship between word senses that matters, and a similarity measure should be able to take this into account (Resnik, 1995b). He does adapt his algorithm to deal with this, although it is not entirely clear how he achieves this.

Simple thesaurus-based methods also fail to take into account the non-uniformity of hierarchical structures, as noted again by Resnik. It may be useful to augment these with some kind of weighting scheme to take this into account, although it is not conclusive that schemes based solely on frequency of occurrence are the most effective.

Methods such as Resnik's information content have the drawback that the assessment of similarity in networks only involves taxonomic (is-a) links. This means that they may exclude some potentially useful information. Resnik notes this, but chooses to ignore it. Tsutsumi's method examines different types of relation by using synonym as well as taxonym hierarchies, although again it is not entirely clear which types of relationships are considered.

A final point on the use of thesauri concerns their coverage. General language thesauri such as WordNet and Roget's Thesaurus are only really suitable for general-language domains, and even then have been found to contain serious omissions. If an algorithm is dependent on resources such as this, it can only be as good as is dictated by the resource. Domain-specific resources are still few and far between, although they are better suited to technical corpora, and tend to be highly individual, so that a method based on a domain-specific thesaurus may not be easily portable to different domains or to other applications.

5.6 Similarity Measurement in the TRUCKS System

Our approach is modelled mainly on the EBMT-based techniques of Zhao (Zhao, 1996) and Sumita and Iida (Sumita and Iida, 1991). This is based on the premise that the position of the MSCA within the hierarchy is important for similarity. The lower down in the hierarchy the MSCA, the more specific it is, and therefore the more information is shared by the two concepts, thus making them more similar. We combine this idea with that of semantic distance (Rigau et al., 1997; Smeaton and Quigley, 1996; Magadur and Tabuteau, 1996). In its simplest form, similarity is measured by edge-counting – the shorter the distance between the words, the greater their similarity. The MSCA is commonly used to measure this. It is determined by tracing the respective paths of the two words back up the hierarchy until a common ancestor is found. The average distance from node to MSCA is then measured: the shorter the distance to the MSCA, the more similar the two words. We combine these two ideas in our measure by calculating two

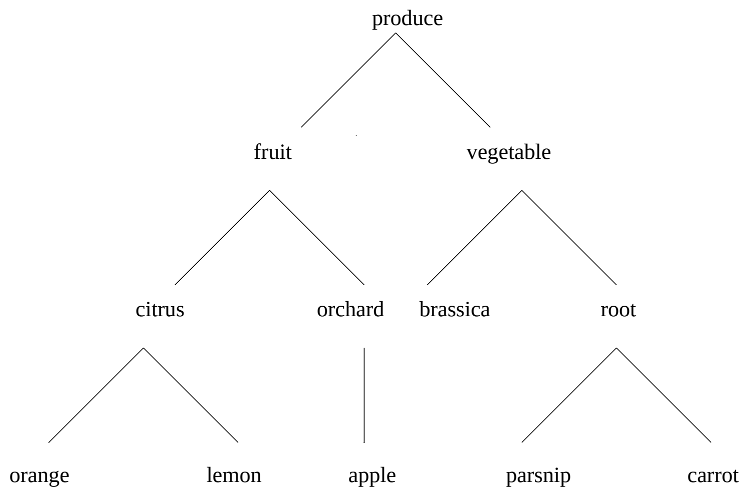

Figure 5.8: Fragment of a food network

weights: one which measures the distance from node to MSCA, and one which measures the vertical position of the MSCA.

To explain the relationship between network position and similarity, we use the example of a partial network of fruit and vegetables, illustrated in Figure 5.8, since this is easier to understand than the UMLS network. Note that this diagram depicts only a simplistic relationship between terms, and does not take into account factors such as diagonal relationships and multidimensionity. We claim that the height of the MSCA is significant. The lower in the hierarchy the two items are, the greater their similarity. In the example, there would be higher similarity between *lemon* and *orange* than between *fruit* and *vegetable*. Although the average distance from *lemon* and *orange* to its MSCA (*citrus*) is the same as that from *fruit* and *vegetable* to its MSCA (*produce*), the former is lower in the hierarchy than the latter. This is also intuitive, because not only do *lemon* and *orange* have the *produce* feature in common, as *fruit* and *vegetable* do, but they also share the features *fruit* and *citrus*.

Our second claim is that the greater the horizontal distance between words in the network, the lower the similarity. By horizontal distance, we mean the

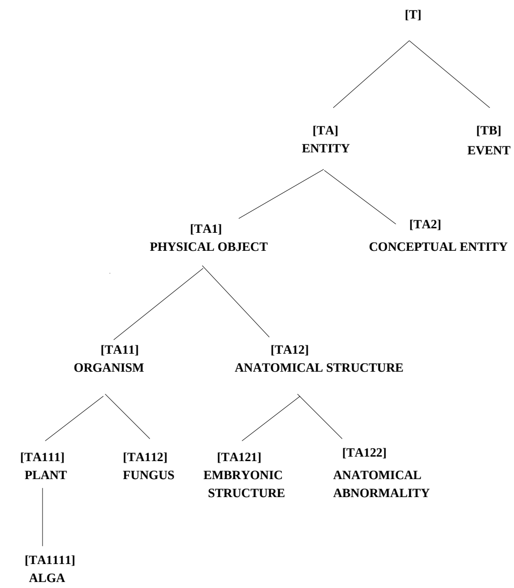

Figure 5.9: Fragment of the Semantic Network

distance between two nodes via the MSCA. This is related to the average distance from the MSCA, since the greater the horizontal distance, the further away the MSCA must be in order to be common to both. In the food example, *orange* and *carrot* have a greater horizontal distance than *lemon* and *orange*, because their MSCA (*produce*) is further away from them than the MSCA of *lemon* and *orange* (*citrus*). Again, it is intuitive that the former are less similar than the latter, because they have less in common.

Taking these criteria into account, we define the following two weights to measure the vertical position of the MSCA and the horizontal distance between the nodes:

- *positional*: measured by the combined distance from root to each node
- *commonality*: measured by the number of shared common ancestors multiplied by the number of words (usually two).

The nodes in the Semantic Network are coded such that the number of digits in the code represents the number of leaves descended from the root to that node, as shown in Figure 5.9. Similarity between two nodes is calculated by dividing the commonality weight by the positional weight to produce a figure between 0 and 1, 1 being the case where the two nodes are identical, and 0 being the case where there is no common ancestor (which would only occur if there were no unique root node in the hierarchy). This can formally be defined as follows:

$$\mathrm{sim}(w_1...w_n) = \frac{\mathrm{com}(w_1...w_n)}{\mathrm{pos}(w_1...w_n)} \tag{5.1}$$

where

com($w_1...w_n$) is the commonality weight of words 1...n

pos($w_1...w_n$) is the positional weight of words 1...n.

It should be noted that the definition permits any number of nodes to be compared, although usually only two nodes would be compared at once. Also, it should be made clear that similarity is not being measured between terms themselves, but between the semantic types to which the terms belong. So a similarity of 1 indicates not that two terms are synonymous, but that they both belong to the same semantic type.

5.6.1 The semantic hierarchy

For our experiments, the similarity between context term and term is measured using the pre-defined conceptual hierarchy of the UMLS Semantic Network, and applied to a corpus of eye pathology reports. The hierarchy is modified slightly in order to make it more general, since some of the distinctions are extremely specific, and we consider that in the first instance it is perhaps too fine-grained. Naturally, the granularity of the hierarchy is a variable which could be altered according to the application or to any specific requirements the user might have.

We modified the structure of the UMLS Semantic Network by assimilating the bottom layers of the network, such that no concept lies deeper than 6 nodes from the root. Any concept lying below this point is automatically allocated to the node of its immediate predecessor in the hierarchy, starting from the lowest set of nodes, until no concepts exist below the cut-off point. The advantage of this incremental approach is that the same process can be adopted should we wish to move the cut-off point either up or down the hierarchy, simply by altering that variable. The hierarchy of concepts is given in Appendix B.

5.6.2 An example

Let us take an example from the UMLS. The similarity between a term belonging to the semantic category *Plant* and one belonging to the category *Fungus* would be calculated as follows:-

- *Plant* has the semantic code TA111 and *Fungus* has the semantic code TA112.

- The commonality weight is the number of nodes in common, multiplied by the number of terms we are considering. TA111 and TA112 have 4 nodes in common (T, TA, TA1 and TA11). So the weight will be $4 * 2 = 8$.

- The positional weight is the total height of each of the terms (where the root node has a height of 1). TA111 has a height of 5 (T, TA, TA1, TA11 and TA111), and TA112 also has a height of 5 (T, TA, TA1, TA11 and TA112). The weight will therefore be $5 + 5 = 10$.

- The similarity weight is the commonality weight divided by the positional weight, i.e. $8/10 = 0.8$.

The similarity measure we have described is used to allocate weights to context terms, depending on their relevance as a term indicator. It can also, however, be used as a means of disambiguating both candidate and context terms, as explained in the next section.

5.7 Using Similarity for Term Disambiguation

Although ambiguity is not as prolific in specialised texts as in general language, it can nevertheless cause problems for terminological applications. Usually this takes the form of multidimensionality, whereby two meanings of a term are related to the same concept but are classified in different ways in the hierarchy. For example, the term *conjunctivitis* is classified in UMLS both as a *sign or symptom* and as a *disease or syndrome*. We shall refer to such a term as having two (or more) *meanings*, although the meanings may not strictly be different. (We could argue that the term has the same meaning in each case, but that there are different viewpoints from which it may be regarded).

There are essentially two issues at stake concerning term sense ambiguity. Firstly, it is beneficial to term extraction if we can classify the terms according to their meaning, so that if a term has two different meanings, we preserve this distinction in the extraction process. The importance of this depends on the ultimate use of the extracted terms. For applications such as information extraction, hypertext linking and machine translation, the granularity of distinction between term senses can vary considerably. For ontology creation or lexical tuning, it may be useful to make very fine-grained distinctions about meaning. Secondly, because of the way in which similarity is measured, if a term is positioned in two or more different places in the hierarchy, this creates problems, because a decision must be made as to which position should be used for the calculation.

The way in which we measure similarity means that it is only really possible to deal with the second issue. This is because we do not examine each occurrence of a candidate term separately, but we look at all its occurrences in the text together. The similarity measure calculates similarity between the candidate term and every context term it occurs with, rather than just similarity between those context terms occurring with a particular instance of it. This means that we cannot use the similarity measure in this way to identify different meanings of a candidate term which occur in the same corpus. We can, however, use it to identify one particular meaning for that term within the corpus. This is more positive than it may appear, because experience shows that within a given corpus, different occurrences of a term are not so likely to have different meanings. Future work, however, could certainly attempt to deal with this issue and extract terms according to their meaning in each instance.

5.7.1 Disambiguation of candidate terms

To disambiguate candidate terms using the similarity measure in this way, there are two intuitive possibilities. One solution would be to measure the similarity between the candidate term and context term for each meaning of the candidate term, and take an average of the results. However, this is liable to give a false impression of how relevant the context term is. For example, a context term might be highly relevant but might happen to have another meaning, situated in a distant part of the hierarchy, which is completely irrelevant. If the semantic distance is averaged out between these two, we would end up with a weighting far removed from either of the possibilities.

An alternative approach is to pick one of the two options, based on some kind of default criteria. This might be frequency-based or score-based. Since we use a relatively small corpus (approximately 1 million words), the frequency-based option is not necessarily a wise choice, because each candidate term may only occur relatively infrequently. The score-based option, on the other hand, seems intuitive. This involves measuring the similarity between each meaning of the candidate term and the context terms it occurs with. The result with the highest score wins, because we assume that the more similar the context terms are to the candidate term, the more likely that that meaning is correct.

5.7.2 Some examples

To illustrate the disambiguation procedure, we shall give some examples from the corpus of eye pathology records.

Example 1

The candidate term *cholesterol_granuloma* has two possible semantic types associated with it:
1. Finding/Result (TA22)
2. Disease/Syndrome (TB222).

In the corpus, it occurs in the following contexts:

```
THE NODULE CONSISTS OF NUMEROUS FOREIGN BODY GIANT CELLS AND
HISTIOCYTES ARRANGED AROUND CHOLESTEROL CLEFTS, THE WHOLE
BEING EMBEDDED IN A FIBROUS STROMA. DIAGNOSIS: CHOLESTEROL GRANULOMA.
```

```
THE SUBRETINAL SPACE CONTAINS A MASS OF SEROUS EXUDATE WITH
SUBSTANTIAL HAEMORRHAGE. MANY PIGMENT LADEN MACROPHAGES ARE
PRESENT.  ONE SMALL CHOLESTEROL GRANULOMA  CAN BE SEEN POSTERIORLY.
```

From the first context, we can surmise that since the cholesterol granuloma is a *diagnosis*, the meaning *Finding/Result* is more relevant than the meaning *Disease/Syndrome*. Similarly, for the second context, the fact that the cholesterol granuloma can be *seen* also indicates that it is a *Finding*. The total similarity weight for meaning 1 is higher than for meaning 2 (2.0 compared with 0.6), which is supported by our intuition from the corpus. So in this case, the meaning *Finding/Result* would be correctly retained.

Example 2

The candidate term *degenerative_change* has the same two possible semantic types associated with it: *Finding/Result* and *Disease/Syndrome*. In the corpus, it occurs in contexts such as the following:

```
THE TISSUE CONSISTS OF A SMALL PIECE OF CONJUNCTIVAL TYPE
EPITHELIUM SHOWING DEGENERATIVE CHANGE IN THE COLLAGEN OF
THE SUBSTANTIA PROPRIA CONSISTENT WITH A PINGUECULA.
```

The total similarity score for the meaning *Finding/Result* is 20.727, while the score for the meaning *Disease/Syndrome* is 14.723. From the contexts, we consider that the first meaning is the best, and this is also the one with the highest score.

These examples by no means prove that the method is accurate. However, it is often very difficult to form an opinion about which meaning is most suited to a term in these kind of circumstances. For example, it is virtually impossible to say whether the meaning *Organ or Tissue Function* or the meaning *Disease or Syndrome* is more suitable for the term *graft_rejection*,in contexts such as:

```
THE PATHOLOGICAL FEATURES ARE CONSISTENT WITH LATE GRAFT
 REJECTION ALTHOUGH THERE IS NO EVIDENCE OF VASCULARISATION.
```

This is one of the reasons why an automatic method of disambiguation is so useful.

Candidate Term	Position	Context Term	Position	Similarity
allergic conjunctivitis	TB222	conjunctivitis	TA22	0.4444
allergic conjunctivitis	TB222	conjunctivitis	TB222	1
connective tissue	TA123	lying	TB11	0.1111
connective tissue	TA123	lying	TA125	0.8
epithelial cyst	TA12	lesions	TB222	0.1111
epithelial cyst	TA12	lesions	TA214	0.2222
epithelial cyst	TA12	lesions	TA22	0.4

Table 5.4: Similarity between candidate terms and ambiguous context terms

5.7.3 Disambiguation of context terms

The same method is also used when context terms are found to be ambiguous (Maynard and Ananiadou, 1998b; Maynard and Ananiadou, 1998c). In this case, similarity between candidate term and each meaning of that context term is measured, and again, the highest score wins. The meaning of the context term which has the lower score is simply discarded (for that particular candidate term). Since the meaning of the context term may be different in another situation in the text, the process is repeated when it is found occurring with another candidate term.

For example, in the corpus we find the candidate term *allergic conjunctivitis* occurring with the context term *conjunctivitis*. This context term has 2 possible positions in the hierarchy, as shown in Table 5.4. One of these belongs to the same node as the candidate term, producing a similarity of 1, whilst the other is in a completely different section of the hierarchy and only has a similarity of 0.4444.

A manual evaluation of a set of ambiguous context terms was carried out, to establish how well this approach worked. For each possible meaning of an ambiguous context term, we identified its semantic category (i.e. the label of its hierarchical position) and attempted to establish whether the meaning with the highest similarity score was in fact the one being used in that context. In some cases, it was not possible to differentiate between the two, since the meanings were very similar, but we estimated that in about 80% of cases, the correct meaning was indeed the one with the highest similarity score. For example, the term *lying* can be classified either as a spatial concept or as a form of social behaviour. It occurs with the candidate term *connective tissue* in the context "*Lying among loose connective tissue is a collapsed cyst...*". Clearly the first meaning is the correct one in this case. The similarity between this meaning of *lying* and *connective tissue* is 0.8, whereas for the other meaning it is 0.1111, which corresponds with our intuitive interpretation.

For the ambiguous context terms, therefore, the scores are calculated for each meaning, and only the highest score (and associated meaning) is retained. If both meanings have an equal score, one meaning is picked at random, since for the time being we have no alternative method of selection.

5.8 Conclusion

In this chapter, we have examined approaches to semantic knowledge acquisition, in particular with regard to lexical resources, and methods of measuring similarity. We have explained in detail the method for measuring similarity which we adopt in order to produce the semantic weight for the SNC-Value. We do not claim superiority of any particular approach over another, but propose simply that different criteria may be important, depending on the application. We have also discussed a method for disambiguation of both candidate and context terms. Although this method appears to be effective, it would undoubtedly benefit from further expansion and more rigorous testing. We leave further evaluation of the similarity measure to Chapter 8, where we discuss its effectiveness and compare it with different statistical approaches.

In the next chapter, we explain how the individual weights for the context are combined to form the Importance Weight, and give details of the complete procedure for applying the SNC-Value to a corpus.

Chapter 6

The SNC-Value

6.1 Introduction

In this chapter, we describe the SNC-Value method in detail. We first explain the C-Value and NC-Value approaches on which it is based, and then discuss the way in which the individual weights are constructed and combined in order to extract and rank candidate terms from a plain corpus of text. For full details of the C-Value/NC-Value, we refer the reader to (Frantzi, 1998; Frantzi and Ananiadou, 1999). Each stage of our approach is illustrated with examples, but full listings of the final results are left to the Appendix.

6.2 The Statistical Approach: C-Value

The C-Value approach uses both statistical and linguistic information to extract terms, but the emphasis is strongly placed on the statistical side. The linguistic aspect consists of 3 parts: a part-of-speech tagger, a linguistic filter, and a stop-list. The statistical part applies a measure called C-Value to the candidate strings produced by the linguistic processor.

6.2.1 The linguistic module

The purpose of the linguistic processing is to break down the corpus into manageable linguistic units and to eliminate unlikely candidates for terms. The linguistic module consists of three elements:

1. Syntactic tagging

2. Linguistic filter

3. Lexical filter

Once the corpus has been pre-processed, it is tagged using Brill's rule-based part-of-speech tagger (Brill, 1992), such that each word is assigned a grammatical category. This process is necessary for the following stage, the linguistic filter. The corpus is passed though the linguistic filter so that only certain combinations of syntactic category are eligible as candidate terms. The more open the filter, the greater the number of terms that will be recognised, but this also encourages more noise. Conversely, the more closed or restrictive the filter, the better the precision, but the greater the chance that terms will be missed. The C-Value does not restrict the filter specifically, because the choice of filter may depend on the application. The two filters possible are:

1. $(Adj|N)^{+}N$

2. $((Adj|N)^{+}|((Adj|N)^{*}(NPrep)^{?})(Adj|N)^{*})N$

In the SNC-Value method, and for the purposes of our experiments, we have used the second filter, which is more open and therefore allows greater recall. The reason for this is that once candidate terms are eliminated at this stage, they can never be regained. If noise is obtained, however, this can still be eliminated later on. In this way, by choosing a more open filter, we can maximise recall without harming precision, since the following stages of the procedure aim to fine-tune the precision. In practice, this means that while more candidate terms will be extracted using this filter, the increase in error will be minimised in the re-ranking procedure whereby such terms can be pushed towards the bottom of the list.

The final part of the linguistic module is the stop list. This is in effect a lexical filter (as opposed to a syntactic filter), i.e. it prevents certain lexical items being accepted. The aim is, as before, to avoid unnecessary noise caused by extracting strings which are unlikely to be terms. Once again, there is a trade-off between precision and recall when deciding which words should form part of the stop list. In Frantzi's experiments (Frantzi, 1998), the stop list consists of 225 function and content words picked from a random sample of the corpus (10%), and is detailed in the Appendix. We see no reason to change the stop list for our experiments, since we are using the same corpus of data, though if a different corpus were to be used, the list would need reformulating acordingly.

Finally, a frequency threshold is also applied, which limits candidate strings to those which occur 3 or more times in the corpus. The reason for this is that strings which occur only once or twice are unlikely to be terms. The threshold is therefore useful in eliminating unnecessary noise. While this also could be altered, since it was tested by Frantzi on the same corpus as we use, we see no reason to alter it.

6.2.2 The statistical module

The statistical part of the process uses a measure called C-Value, which is based on 4 main characteristics of the candidate string produced by the linguistic module:

1. frequency of occurrence of candidate string;
2. length of candidate string;
3. frequency as a nested string of longer candidate terms;

4. number of these longer candidate terms.

We shall examine briefly each of these parameters. For more detailed information, see (Frantzi, 1998; Frantzi and Ananiadou, 1999).

Frequency of occurrence

This is the most common method of establishing termhood in ATR systems, and can be used in itself as a measure, as we have seen in Chapter 3. But although it provides a good starting-point, not all terms have high frequency, and not all high-frequency strings are terms. Clearly there are other factors which need to be considered.

Nested terms

Nested terms, i.e. terms which are contained within longer terms, also pose a problem. This is because some nested terms occur by themselves, while others do not. Simple frequency of occurrence does not provide a way of counting nested terms which do appear on their own, whilst avoiding those which do not. To achieve this, we also need to consider the number of times a nested term appears within longer terms, and the number of longer terms containing the nested term in question. These two parameters are said to indicate the degree of independence of the nested term.

String length

The length of string (i.e. the number of words it contains) is also important when considering frequency of occurrence in a corpus. Because of the possibility of nested terms occurring, a short string is more likely to occur than a longer string. This means that the longer the string, the greater the importance it should be accorded.

The C-Value measure incorporates all the above parameters. Since some terms are never found as nested, and also maximum length terms also cannot be nested, in these cases the two parameters relating to this aspect are ignored. So two different cases apply: one for terms that are found as nested, and one for terms that are not. The C-Value measure is described formally as follows:

$$\text{C-Value}(a) = \left\{ \begin{array}{ll} log_2|a| \cdot f(a) & \text{a is not nested} \\ log_2|a| - \frac{1}{P(T_a)} \sum_{b \epsilon T_a} f(b)) & \text{is nested} \end{array} \right\}$$

Essentially, this means that if a candidate string is not found as nested, its termhood is calculated from its total frequency and length. If it is found as nested, termhood is calculated from its total frequency, length, frequency as a nested string, and the number of longer candidate terms it appears in.

6.3 Adding Contextual Information: the NC-Value

The NC-Value method builds on the C-Value method by incorporating contextual information into the algorithm. The set of context words is composed of all the nouns, adjectives and verbs in a fixed-size context window, either preceding or following a potential term.

Each context word is assigned a weight based on how frequently it appears with a term. The higher the proportion of terms to non-terms appearing with a context word, the more likely that anything appearing with it will be a term. So the weight is derived from the proportion of occurences with a term (out of all its occurrences).

Since at this stage we do not know which words are terms and which are not, a bootstrapping approach is used by taking the top-ranked terms extracted by the C-Value list. The list of candidate terms is sorted in descending order of weight, such that the highest weighted (and thus most likely) terms are at the top of the list. A term is considered top-ranked if it belongs to the top third of this list. To calculate their weight, context words are therefore considered in conjunction with top-ranked terms. This weight is described formally as follows.

$$Weight(w) = \frac{f_w(t)}{n(t)} \tag{6.1}$$

where

w is the context word to be assigned a weight,

$Weight(w)$ is the weight assigned to word w,

t is a top-ranked term,

$f_w(t)$ is the number of terms occurring with word w,

$n(t)$ is the number of terms considered.

Weights are first allocated to all context words in the corpus according to the above algorithm. Context words are identified by taking a window of a fixed number of words (which can be set by the user) either side of all candidate terms (which are the terms extracted by the C-Value measure), and filtering out all those which are not nouns, adjectives or verbs. The weight allocated to each context word is global rather than local. This means that it is only dependent on the corpus, and will remain the same for each candidate term it appears with.

Then, for each candidate term, the context words which occur with it are extracted, along with their weights. This is repeated for each occurrence of that term in the corpus. For each term, a total weight, known as the *context factor*, is then calculated by summing the total weights for the context words, multiplied by their frequency, This is described formally as:

$$CF(t) = \sum_{w \epsilon C_t} f_t(w) \cdot weight(w) \tag{6.2}$$

where

t is the candidate term,

$C_t$is the set of context words of t,

w is a word from C_t,

$f_t(w)$ is the frequency of w as a context word of t,

$weight(w)$ is the weight of w as a context word,

The NC-Value combines the C-Value with this context factor to produce a new ranking of the list of terms. A set of experiments, detailed in (Frantzi, 1998), was used to determine the way in which the C-Value and the Context Factor should be combined to produce the best formula for maximising precision and recall. The final formula combines the two in the ratio 4:1, as described formally below:

$$NCvalue(a) = 0.8 * Cvalue(a) + 0.2 * CF(a) \tag{6.3}$$

where

a is the candidate term,

$Cvalue(a)$ is the Cvalue for the candidate term,

$CF(a)$ is the context factor for the candidate term.

6.4 Adding Deeper Information: the SNC-Value

As discussed in Chapters 4 and 5, we propose that by incorporating deeper forms of linguistic information into our measure, we can better differentiate between context words which are significant and those which are not. Although the s-tatistical method used in the NC-Value does improve results over the C-Value, relying largely on frequency can be misleading, especially in relatively small cor-pora. The other issue that is not addressed by the NC-Value method is that some context words may be relevant to particular terms. Because the weights are calculated globally rather than locally, this problem is ignored in the NC-Value.

6.4.1 The Importance Weight

The three context weights described in Chapters 4 and 5 are combined together to make up the Importance Weight (IW). For each part of the weight, a total score is found by summing the individual scores for each relevant context word occurring with the candidate term. The Importance Weight is then calculated as follows:

$$IW(a) = \sum_{b \epsilon C_a} syn_a(b) + \sum_{d \epsilon T_a} f_a(d) \cdot sim_a(d) \tag{6.4}$$

where

a is the candidate term,

$C_a$is the set of context words of a,

b is a word from C_a,

$f_a(b)$ is the frequency of b as a context word of a,

$weight(b)$ is the weight of b as a context word,

$syn_a(b)$ is the syntactic (category) weight of b as a context word of a,

T_a is the set of context terms of a,

d is a word from T_a,

$f_a(d)$ is the frequency of d as a context term of a,

$sim_a(d)$ is the similarity weight of d as a context term of a.

Note that while the individual parts of the weight are calculated for the context words or context terms, the importance weight is calculated with respect to the candidate term. The individual parts of the weight are defined as follows. The syntactic weight is composed simply of the category weight allocated to the boundary word, defined earlier in Table 4.1. The semantic weight is defined in the previous chapter by Formula 5.1.

6.4.2 Calculating the SNC-Value

The Importance Weight is therefore calculated for each candidate term, and is then combined with the NC-Value to produce a re-ranking of the list of candidate terms. It is combined by simply adding the two scores together for each candidate term, i.e:

$$SNCValue(a) = NCValue(a) + IW(a) \tag{6.5}$$

where

a is the candidate term

$NCValue(a)$ is the NC-Value of a

IW is the Importance Weight of a

6.5 Applying the SNC-Value to Term Extraction

The main purpose of the SNC-Value measure is to identify and extract multi-word terms by ranking them in order of termhood. The term extraction procedure is carried out in stages. There are two main reasons for this. Firstly, at each stage, modifications can be incorporated in order to fine-tune the results. Secondly, the whole procedure takes the form of a bootstrapping approach, as we have mentioned previously. This means that the results from each stage of the process are used to fine-tune the results for the following stage. Thirdly, the results of individual stages also may have other applications. For example, the semantic component is also used later for the clustering of contexts, and the disambiguation procedure can be applied independently from the rest of the extraction procedure.

6.5.1 The data

The algorithm has been tested on a corpus of eye pathology records, although it could be applied equally to any form of medical text. The corpus is first pre-processed to convert it entirely into lower-case characters (for ease of processing, since the UMLS terms are in lower-case) and to eliminate any anomalous information such as extra spaces and symbols. It is then tagged so that each word is annotated with its syntactic category.

6.5.2 Stage 1: the C-Value

The first part of the procedure is the C-Value, which uses a combination of statistical and linguistic information about candidate terms to give them an initial ranking. This is carried out in two stages: the linguistic module and the statistical module.

The linguistic module

The linguistic module consists of 3 parts: part-of-speech tagging, the linguistic filter and the lexical filter. The corpus is tagged using Brill's part-of-speech tagger. For a detailed description of this, see (Brill, 1992; Brill, 1994). Before using this tagger, the corpus first has to be tokenised following the Penn Treebank Tokenisation (Marcus et al., 1993), such that punctuation is separated from words, and each sentence is on one line.

The tagging process consists of two stages. First, each word is assigned its most likely tag in isolation, using a lexicon file. Any word not found in the lexicon is assumed to be a noun, although other cues can be used, such as prefixes and suffixes.

In the second stage, contextual information is used to fine-tune the tagging, using various sets of morphological and syntactic rules, and a set of bigrams. The files containing rules and lexical information can be trained on the corpus and fine-tuned by the user. However, we did not make any alterations to the default files used by Frantzi for her NC-Value experiments, since we wished to be able to make comparisons between the SNC-Value and the NC-Value without considering other criteria such as the tagging accuracy. We do, however, recognise that the results of the tagging are not perfect and could have an influence on the results of our method.

Aside from errors and inconsistencies in the original data (for example, spelling mistakes, typographical errors and inconsistent use of hyphens and quotation marks), many words are wrongly tagged as nouns because they are not recognised. This appears to be a particular problem of using a general tagger on a domain-specific text. Although the tagger can be trained on a particular corpus, this is a time-consuming process and, for the reasons mentioned above, we did not make use of this facility. It is likely that the majority of unknown words in general language will be nouns, since both new technical terms and general language neologisms are mainly nominal. However, in the eye pathology corpus we found many examples of adjectives relating particularly to the medical field, which were unrecognised by the tagger and therefore tagged as nouns, e.g. *acantholytic, acanthotic, acetic, aqueous, adipose, anterior* and *bronchopulmonary.* There were also some unusual adjectives formed from general language words, such as *bean-shaped, cheese-like, cord-like* and *cottage-shaped.* Whilst the individual components would be recognised in the lexicon, the tagger does not really cater for compound words such as these, and treats them as unknown (and therefore nouns).

Once the tagging process is complete, the corpus is ready for the preliminary extraction of candidate strings using the linguistic filter, stop list and frequency threshold. As discussed earlier, there are two possible filters which can be adopted, of which we use the more open one:

$((Adj|N)^+|((Adj|N)^*(NPrep)^?)(Adj|N)^*)N$

Any strings matching this pattern are extracted, and submitted to the lexical filter, which eliminates any strings containing words which match an entry in the stop list. Finally, the frequency threshold is applied, so that only words which occur 3 or more times in the corpus are accepted.

The statistical module

The statistical module consists of applying the C-Value measure to the candidate strings extracted by the linguistic module. For details of the exact algorithm used to apply the measure, see (Frantzi, 1998). The C-Value threshold is set to 0, so that all strings with a value of greater than 0 will be included in the list. The resulting candidate terms and their values are ranked in order of termhood, with the highest appearing at the top of the list. The results of these rankings are shown in the Appendix.

6.5.3 Stage 2: the NC-Value

The second stage builds on the C-Value by adding contextual information to the candidate terms and re-evaluating their status. The entire list of candidate terms extracted by the C-Value method is used as a starting point. First, the context words are extracted from the corpus. Any noun, adjective or verb falling within 10 words of the candidate term is considered as a context word. A weight is calculated for each context word according to the formula 6.1, by dividing the number of top-ranked terms occurring with the context word by the number of times the context word appears in total. Then for each candidate term, a context factor is calculated. According to the formula 6.2, this is calculated by

Occurrence	Context Words	Weight
1	show	0.5
	cornea	0.2
	example	0.25
2	show	0.5
	keratitis	0.1
3	cornea	0.2
	example	0.25
	tissue	0.4

Table 6.1: Example of context words and weights occurring with the term *proliferative_retinopathy*

multiplying the weight of the context word by its frequency, and summing the total for all context words appearing with the candidate term. In practice, it is easier to simply collect a list of context words and their weights. Each time a context word appears, it is added to the list. Once the list is complete, the weights are added together to produce the same result.

Let us demonstrate this with an example. Suppose we have the candidate term *proliferative_retinopathy*, and we find the following occurrences of it in the text, as shown in Table 6.1. According to the formula, we need to take each context word and multiply its weight by its frequency. So we would get the following results.

```
show       0.5 * 2   = 1
cornea     0.2 * 2   = 0.4
example    0.25 * 2  = 0.5
keratitis  0.1 * 1   = 0.1
tissue     0.4 * 1   = 0.4
Total                = 2.4
```

Alternatively, we can just ignore the individual words and sum the weights, i.e.

`0.5 + 0.2 + 0.25 + 0.5 + 0.1 + 0.2 + 0.25 + 0.4 = 2.4`

From a computational point of view, the second method is quicker and easier, yet produces exactly the same result as the first.

The NC-Value is therefore calculated for the candidate terms, using Formula 6.6:

$$NCvalue(a) = 0.8 * Cvalue(a) + 0.2 * CF(a) \tag{6.6}$$

where

a is the candidate term,

$Cvalue(a)$ is the C-Value for the candidate term,

$CF(a)$ is the context factor for the candidate term.

A list is compiled of the NC-Value weights for the candidate terms, ranked in order of decreasing value, and divided into 3, as with the C-Value list.

6.5.4 Stage 3: the SNC-Value

The third stage involves adding deeper linguistic information about the context, as described earlier. In this stage, we fine-tune the contextual information so that more of a distinction is made about the relative importance of different context words.

Acquiring a list of preliminary candidate terms

The context is defined as before by taking a window of 10 words either side of the candidate term. It is then processed using a search method which takes as input a list of candidate terms and the corpus. The list of candidate terms is compiled from the NC-Value list. At this stage, the weights allocated in the NC-Value process can be ignored, as they are no longer useful.

Extracting contextual information

The list of candidate terms is sorted into alphabetical order for ease of processing, and a search program iterates through the list, finding all the occurrences of each term in turn, together with a window of 10 words either side. A list is compiled of terms and their context words. The example below shows some of the contexts extracted with the candidate term *descemet's membrane* – they are too numerous to list here in their entirety.

```
Term: DESCEMET'S MEMBRANE
Left: ASPECT A SHORT LENGTH OF
Right: AND SOME RATHER TATTERED TRABECULAE
Left: OF THE PERIPHERAL PART OF
Right: AND THE CORNEO-SCLERAL PORTION OF
Left: THE STROMA IS SLIGHTLY OEDEMATOUS
Right: IS INTACT
Left: MEMBRANE HAS BEEN LARGELY LOST
Right: AND THE CORNEAL ENDOTHELIUM APPEAR
Left: BENEATH THIS
```

The program contains variables which can be altered to change options such as the size of the context window, and whether punctuation should be included or not. For our experiments we set the size of the context window to 10 words either side of the candidate term, unless the end of the sentence was reached. This means that a context could not exceed the boundaries of a sentence.

The linguistic filter

The context is then passed through the linguistic filter, so that only nouns, adjectives, prepositions and verbs are permitted. Those context words which do not pass through the filter are simply discarded, since we have no further use for them at this stage. Note that auxiliary verbs are tagged as a separate category from verbs, and therefore do not pass through the filter. The resulting list of candidate terms and context words derived from the previous example is shown below.

```
Term: descemet's membrane Context Word: aspect
Term: descemet's membrane Context Word: short
Term: descemet's membrane Context Word: length
Term: descemet's membrane Context Word: of
Term: descemet's membrane Context Word: tattered
Term: descemet's membrane Context Word: trabeculae
Term: descemet's membrane Context Word: of
Term: descemet's membrane Context Word: peripheral
Term: descemet's membrane Context Word: part
Term: descemet's membrane Context Word: of
Term: descemet's membrane Context Word: corneo-scleral
Term: descemet's membrane Context Word: portion
Term: descemet's membrane Context Word: of
Term: descemet's membrane Context Word: stroma
Term: descemet's membrane Context Word: oedematous
Term: descemet's membrane Context Term: intact
Term: descemet's membrane Context Word: membrane
Term: descemet's membrane Context Word: lost
Term: descemet's membrane Context Word: corneal
```

Category	Weight
Verb	1.2
Prep	1.1
Noun	0.9
Adj	0.7

Table 6.2: Syntactic weights for boundary words

```
Term: descemet's membrane Context Word: endothelium

Term: descemet's membrane Context Word: appear

Term: descemet's membrane Context Word: beneath
```

The syntactic weight

The second stage determines the boundary weight of the context for each candidate term. Boundary words are identified as being any context word immediately preceding or following a candidate term, which lies within the same sentence. Note that we do not distinguish between these two types of boundary word, as discussed in Chapter 4. The syntactic filter is the same as for context words, i.e. nouns, adjectives, verbs and prepositions are the only categories permitted. The syntactic weight for a candidate term is described as the total of the syntactic weights of all the boundary words occurring with that candidate term. The syntactic weight is calculated for each boundary word, using Table 6.2.

For each candidate term encountered, the relevant boundary words are identified, and the syntactic weights are allocated. The weights for all the boundary words occurring with that candidate term are then summed to produce a total boundary weight for the candidate term. The example shown above would thus have the following weights attached:

```
descemet's membrane | of | 1.1

descemet's membrane | of | 1.1
```

Category	Occurrences	Weight	Total Score
Verb	1447	1.2	1736.4
Prep	1223	1.1	1345.3
Noun	960	0.9	864
Adj	308	0.7	215.6
Total	3938		4160.6

Table 6.3: Syntactic weights for *descemet's membrane*

Category	Occurrences	Weight	Total Score
Verb	12	1.2	14.4
Prep	14	1.1	15.4
Noun	18	0.9	16.2
Adj	4	0.7	2.8
Total	48		48.8

Table 6.4: Syntactic weights for *macular dystrophy*

```
descemet's membrane | oedematous | 0.7
descemet's membrane | intact | 0.7
descemet's membrane | lost | 0.7
```

There are too many occurrences of the candidate term *descemet's membrane* to list all the syntactic weights here, but they are summarised in Table 6.3. For comparison, Tables 6.4 and 6.5 show the weights for the candidate terms *macular dystrophy* – another example of a good candidate term – and *acanthotic*

Category	Occurrences	Weight	Total Score
Verb	9	1.2	10.8
Prep	18	1.1	19.8
Noun	55	0.9	49.5
Adj	22	0.7	15.4
Total	104		95,5

Table 6.5: Syntactic weights for *acanthotic hyperkeratotic*

hyperkeratotic – an example of a poor candidate term[1].

Terminological status

The next stage is to determine which of the context words are also terms. The terminological status of the context words is defined by matching any context words with (single-word) terms from the UMLS thesaurus. The reason that only single-word terms are matched is two-fold. Firstly, most multi-word terms contain at least one single-word term. If we matched both single and multi-word terms, we would frequently end up counting multi-word terms twice. Secondly, each context word is allocated an individual score. Involving multi-word terms would therefore cause problems since scores would have to be allocated to lexical units, rather than individual words, which would interfere with the other stages of processing.

A program takes as input the list of candidate terms and context, and a list of single-word terms extracted from the UMLS Metathesaurus files. The context words are then matched with the list of terms, to form a separate list of candidate terms and the context terms they occur with, as shown below.

```
Term: proliferative_retinopathy Context Term: retina
%Term: proliferative_retinopathy Context Term: severe
Term: corneal_epithelium        Context Term: abnormality
Term: macular_oedema            Context Term: retinopathy
Term: fibrous_membrane          Context Term: cellular
Term: dyskeratotic_cells        Context Term: acantholysis
Term: intradermal_naevus        Context Term: bilateral
Term: fibrous_tissue            Context Term: biopsy
Term: abundant cytoplasm        Context Term: mitosis
```

[1] By "good" and "poor" candidate terms, we mean terms which respectively would be likely and unlikely to be considered valid by an expert.

Each context word found to be a context term is then allocated a score of 1 for this section, and each context word which is not a context term is allocated a score of 0 for this step.

Semantic weight

The semantic weight is only relevant to the context terms, since the remaining context words, useful though they may still be, cannot be allocated a score, as they are not found in the UMLS. We create first a list of triples, containing the candidate or context term, its semantic category from UMLS, and the matching semantic code, e.g.

```
absinthe_addiction        TB222  Disease or Syndrome
absinthium_oil            TA1412 Chemical Viewed Structurally
absolute                  TA212  Qualitative Concept
absolute_afferent_
pupillary_defect          TB222  Disease or Syndrome
absolute_dullness         TA22   Finding
absolute_glaucoma         TB222  Disease or Syndrome
absolute_hypovolemia      TA22   Finding
absorbable_cellulose      TA1412 Body Substance
absorbable_gelatin_sponge TA141  Chemical
allergic_conjunctivitis   TB222  Disease or Syndrome
amelanotic melanoma       TB222  Disease or Syndrome
anaplastic_carcinoma      TB222  Disease or Syndrome
angle_closure_glaucoma    TB222  Disease or Syndrome
```

The input for this section is the list of pairs of candidate terms and context terms, and this file. The two files are then matched and combined so that each candidate term or context term is tagged with its semantic category and code, e.g.

```
allergic_conjunctivitis TB222 conjunctivitis TA22
allergic_conjunctivitis TB222 history        TA211
allergic_conjunctivitis TB222 insufficient   TA214

amelanotic_melanoma     TB222 abnormality    TA122
amelanotic_melanoma     TB222 absolute       TA212
amelanotic_melanoma     TB222 anaplastic     TA212
amelanotic_melanoma     TB222 extreme        TA212

anaplastic_carcinoma    TB222 hyperchromatic TA214
anaplastic_carcinoma    TB222 islands        TA215|
anaplastic_carcinoma    TB222 secondary      TA213
anaplastic_carcinoma    TB222 tumour         TB222

angle-closure_glaucoma  TB222 area           TA215
angle-closure_glaucoma  TB222 drainage       TB13
angle-closure_glaucoma  TB222 gliosis        TB222
angle-closure_glaucoma  TB222 haemorrhage    TA22
```

The semantic code is used for the similarity calculation in the next step. Since there is a unique relationship between the semantic category and the semantic code, the semantic category is not crucial to the weight calculation, but it makes it easier for any manual evaluation which may take place later.

Calculating similarity

Similarity between each pair of candidate term and context term is then calculated using the semantic codes. A simple program applies the formula detailed in the previous chapter to each pair, which we repeat here for convenience:

$$sim(w1, w2) = \frac{com(w1, w2)}{pos(w1, w2)}$$

where

com(w1,w2) is the commonality weight of word 1 and word 2

pos(w1,w2) is the positional weight of word 1 and word2.

This produces a similarity weight for each combination of candidate term and context term, as in the example below.

```
allergic_conjunctivitis TB222 conjunctivitis TA22  0.2222
allergic_conjunctivitis TB222 history        TA211 0.2
allergic_conjunctivitis TB222 insufficient   TA214 0.2

amelanotic_melanoma     TB222 abnormality    TA122 0.2
amelanotic_melanoma     TB222 absolute       TA212 0.2
amelanotic_melanoma     TB222 anaplastic     TA212 0.2
amelanotic_melanoma     TB222 extreme        TA212 0.2

anaplastic_carcinoma    TB222 hyperchromatic TA214 0.2
anaplastic_carcinoma    TB222 islands        TA215 0.2
anaplastic_carcinoma    TB222 secondary      TA213 0.2

angle-closure_glaucoma  TB222 area           TA215 0.2
angle-closure_glaucoma  TB222 drainage       TB13  0.4444
angle-closure_glaucoma  TB222 gliosis        TB222 1
angle-closure_glaucoma  TB222 haemorrhage    TA22  0.2222
```

The disambiguation module

At this point, ambiguous terms and context terms are also dealt with. Following the procedure detailed in the previous chapter, any candidate term having more than one semantic category is passed to the disambiguation module, processed separately, and then re-incorporated into the results.

Calculating the importance weight

Once all the individual weights have been allocated, they can be combined to form the Importance Weight for the term. Using the file of pairs of candidate term and context word, the weights for each are added and the total importance weight is calculated according to Formula 6.4.

Finally, the NC-Value results are combined with the Importance Weight to give the final SNC-Value (according to Formula 6.5), which ranks the candidate terms. An example of the final results is shown below in Table 6.6. This compares the top 20 results from the SNC-Value list with the top 20 from the NC-Value list. The terms in italics are those which were considered by the subject experts as not valid. We shall discuss the results in more detail in Chapter 8, but we can note here two points. Firstly, the weights for the SNC-Value are substantially greater than those for the NC-Value. This, in itself, is not important, since it is the position in the list, i.e. the *relative* weight, rather than the *absolute* weight, which is important. Secondly, we can see that there are more valid terms in the SNC-Value results than in the NC-Value results. It is hard to make further judgements based on this list alone, because we cannot say whether one term is better than another, if the two terms are both valid.

Term	SNC	Term	NC
bowman's_membrane	605782	*plane_of_section*	1752.71
malignant_melanoma	231237	descemet's_membrane	1345.76
hyaline_fibrous_tissue	215843	basal_cell_carcinoma	1268.21
planes_of_section	170016	*stump_of_optic_nerve*	993.15
trabecular_meshwork	157353	basal_cell_papilloma	616.614
keratinous_debris	101644	*plane_of_section=*	506.517
bruch's_membrane	94996.2	*melanoma_of_choroid*	497.673
plane_of_section=	90109.4	*planes_of_section*	453.716
melanoma_of_choroid	71615.1	malignant_melanoma	448.591
lymphocytic_infiltration	53822	optic_nerve_head	422.211
ciliary_processes	52355.7	ciliary_processes	421.204
cellular_fibrous_tissue	51486.8	bruch's_membrane	413.027
squamous_epithelium	46928.9	keratinous_cyst	392.944
optic_nerve_head	39054.5	*ellipse_of_skin*	267.636
pupillary_border	36510.8	*wedge_of_lid_margin*	211.414
corneal_epithelium	31335.9	*scar_track*	228.217
scleral_invasion	31017.4	connective_tissue	167.053
granulation_tissue	28010.1	*vertical_plane*	167.015
stratified_squamous_epithelium	27445.5	carcinoma_of_lid	164
ocular_structures	26143.6	*excision_biopsy*	155.257

Table 6.6: Top 20 results for the SNC-Value and NC-Value

6.6 Summary

In this chapter, we have explained how the context weights are combined to form the SNC-Value, and how terms are extracted and disambiguated using this method. From an initial corpus of medical data and a domain-specific thesaurus, we extract candidate terms and rank them in order of termhood, or likeliness of being a term. Although we rely on a pre-existing thesaurus, unknown terms are still recognised, but are simply not allocated a similarity weight. This lowers their total score a little, which reflects a lower confidence in previously unknown terms, but does not prevent them being ranked highly. We leave explanation and evaluation of the results until Chapter 8. In the following chapter, we investigate how the contextual information we have acquired can be clustered and used for further improving the ranking, and for other NLP tasks such as automatic ontology building.

Chapter 7

Clustering Contexts

7.1 Introduction

The third part of our work involves expanding the contextual information we have acquired. Once semantic relationships between terms and context have been established, we can combine this with syntactic information to group contexts according to their similarity. Context clustering is useful for a variety of purposes, such as further disambiguation, further statistical processing, and in particular for ontology building (Maynard and Ananiadou, 1999c). The first part of this chapter will discuss the reasons why context clustering is useful, and will then examine some of the ways in which it can be achieved. Then we shall elaborate on our own method for clustering, using syntactic and semantic information, and discuss its effectiveness and further extensions.

7.2 Aims of Context Clustering

One of the major advantages of having clustered contexts is that they provide more information than a single context, particularly where corpus analysis and any kind of statistical processing is involved. With small corpora, each context may appear only once or a few times in the whole text. Where frequencies are low, statistical skewing can occur. If contexts can be considered as a group, frequencies will be higher, leading to more accurate statistical results. Also, contexts which behave in a similar way should not necessarily be treated as separate entities.

We identify and outline 3 main areas in which context clustering can be useful for NLP, all of which are related: disambiguation, knowledge acquisition and lexical tuning. These rely on two main advantages brought by clustering: firstly, that it enables the statistical importance of contexts to be more effectively measured, and secondly, that it provides a source of previously inaccessible information.

7.2.1 Disambiguation

Clustered contexts can be used to help disambiguate terms, particularly multi-dimensional terms, where the ambiguity is a result of multiple inheritance in a hierarchy. In the same way that different words which occur in the same context are likely to be synonymous, so different occurrences of one ambiguous word (or term) which occur in the same context are likely to have the same meaning (Grefenstette, 1994). This also means that the reverse is possible, i.e. that occurrences of terms occurring in contexts with words from different semantic families are likely to have different meanings (i.e. belong to different semantic families). This is because terms tend to appear in close proximity with terms belonging to the same semantic family. Indeed, this was one of the basic assumptions behind the context weights we developed for the SNC-Value.

In Chapter 5, we demonstrated that by comparing the semantic similarities of contexts and terms, we were able to disambiguate terms using this information. Where there was a choice of meanings for the term, the meaning with the highest similarity to the context was likely to be most appropriate. One of the problems with this method is that it relies only on information about contexts in general.

This has the drawback that we do not get any kind of measure of how important that context is, and we cannot therefore identify and reject any anomalous contexts which are not good indicators. If we find one context occurring many times with a term, and a different context occurring once with a term, clearly the first context is statistically more important, and its relationship with the term is likely to be more significant. Without using statistical information, these two contexts would be treated as equal.

7.2.2 Knowledge acquisition

Secondly, context clustering is useful for various forms of knowledge acquisition. One of the most important things about clustering is that it enables us to make use of contexts about which we would not otherwise have sufficient information. In this way, information about clusters can be used where information about individual words is not available. If we had a context containing a context term about which we have no information, we would not be able to make use of that context in recognising or disambiguating terms. If we can show that this context is similar to a set of other contexts about which we do have knowledge, this information can be extended to that context. For example, suppose we have a term, T, and a set of contexts, S, which are all semantically similar and which are closely related to term T. If we have another context, C, about which we have little information, and we can show that it is similar to our set of contexts S, we can predict that it will behave in a similar way to each of those contexts and therefore also be strongly related to term T.

A more concrete example might serve to make things clearer. Let us assume:
a term T: *proliferative retinopathy*,
a set of contexts S: {*lower third of cornea*, *lower half of iris*, *top half of retina*},
a context C: *lower section of cornea*.

Suppose we have no semantic information available about this context except that we know the word *section* is similar to the words *third* and *half*. Given this knowledge, we can predict that C is similar to the other contexts and should form part of the set S. Once we know that C behaves in a similar way to the other contexts, we can predict that the relationship between term T and context C will be similar to the relationship between T and the other members of the set S.

Clusters are also useful for applications which require other types of knowledge which are hard to obtain. For example, ontological or world knowledge is virtually impossible to acquire in sufficient amounts and detail to satisfy even a subdomain. To acquire general ontological knowledge about clusters of objects is, however, more feasible, and it is reasonably safe to assume that this type of information can be extended to all the members of a cluster of similar objects.

7.2.3 Tuning of ontologies

The third, and probably most important, use of clustering techniques is for the development and tuning of ontologies. Even though domain-specific ontologies do exist and are quite comprehensive in certain fields, it is unlikely that any pre-existing ontology will be entirely adequate for an NLP application. Apart from containing potential errors, omissions and redundancies, the ontology is also limited if it is not tailored specifically to the corpus in question. For example, a specialised medical thesaurus can be too general for a corpus of eye pathology records, and even if we had a specialised eye pathology thesaurus, this would not guarantee its appropriateness for every corpus in the field of eye pathology.

The other problem is the non-standardisation of terminology. Because terms are not always standardised, and different variations are used, this means that either ontologies have to be all-encompassing, leading to extremely complicated entries and potential massive redundancy when used with a specific corpus, or alternatively that separate ontologies are necessary in order to ensure that they are accurate enough for particular corpora. Evidently, the extent of this problem is dependent on the text type and subject field.

Although standardisation is very much a currently targeted issue, it is unlikely that the problem will be resolved overnight, and therefore methods of adapting ontologies need to be considered in order to realise their full potential in NLP applications. It is important for almost any NLP system that linguistic knowledge can be adapted to the domain (Kiyono and Tsujii, 1994), since a major problem is the potential infiniteness of knowledge required. In the past, systems have been customised manually, but it is far more desirable that this be achieved at least semi-automatically, due to the time and effort necessary for manual customisation.

Customisation, or tuning, of lexical resources to the domain is therefore a very important task for NLP. Both term recognition and clustering can help in this process by structuring the knowledge acquired. In the following section, we outline and discuss some of the methods used for this structuring process. In Section 7.4 we then describe our technique for clustering contexts using semantic frames.

7.3 Approaches to Clustering

7.3.1 Creating semantic frames

The idea of acquiring semantic information from texts using some kind of information formatting dates back to the early late 1960s with Harris' *distributional hypothesis* (Harris, 1968) and Hirschman and Sager's work in the 1970s (Sager, 1972), (Hirschman et al., 1975), which focused on determining sets of sublanguage-specific word classes using syntactic patterns from domain-specific

Subject	Verb	Object
patient	develop	mild cold
patient	develop	fever
patient	develop	slight cough
patient	develop	headache

Table 7.1: Example of an information format for the class *Sign or Symptom*

corpora. By examining sets of lexical items found in specific syntactic environments, Hirschman and Sager established semantic word classes in the field of clinical reporting. For example, by collecting instances of the lexical items found as objects of the verb *develop* together with the subject *patient*, they developed a class *sign or symptom*, consisting of lexical items such as *mild cold*, *fever*, *slight cough* etc. The simple frame created looks something like Table 7.1.

Since then, much work on semantic knowledge acquisition has followed the same approach, although the purposes for which it is used and the ways in which it is implemented differ widely. Rocha et al. capture semantic information by means of an Event Definition model (Rocha et al., 1994), which uses a case-frame like structure to represent clinical events. The frame holds slots consisting of semantic categories, and fillers consisting of the lexical items which have been assigned these categories. The Event Definition frames are not used for clustering, but as a kind of interlingual structure for translating from one controlled medical vocabulary to another. The principles used, however, are very similar. For example, the sentence *chest x-ray shows lung nodule* would be represented by the frame shown in Table 7.2. The semantic information is acquired using a manually-created lexicon, in which each term is allocated a semantic type, and identical concepts are represented by the same numerical code.

Slot	Filler
Procedure:	chest x-ray
Link:	shows
Observation:	nodule
Location:	lung

Table 7.2: Example of Rocha et al.'s Event Definition model

7.3.2 Lexical vs. structural pattern matching

There has been much debate as to whether semantic information can be more easily and reliably obtained from syntactic or from lexical patterns (Montemagni and Vanderwende, 1993; Davidson et al., 1998; Grefenstette, 1994). The general consensus seems to be that syntactic patterns produce more extensive coverage. Lexical patterns can provide some very effective results within a limited range, but it is very hard to identify all the necessary patterns for certain types of semantic information or for certain semantic relations.

Davidson et al. (Davidson et al., 1998) used both lexical and grammatical patterns to identify conceptual relations, although they found that there were many inherent problems with the lexical pattern matching. Their aim was to uncover all the lexical forms which indicated specific conceptual relations, such as hyponymy. They discovered, however, that the variability of lexical patterns made it hard to generate mappings to semantic frames, because the same semantic pattern could be expressed by a variety of surface lexical realisations, making it difficult to pinpoint individual mappings.

Condamines and Rebeyrolle (Condamines and Rebeyrolle, 1998) searched for conceptual relationships using a combination of syntactic and lexical patterns. Although their method worked well, there were some relationships (such as *has the function of*) which they could not identify reliably, because they could not find an unambiguous pattern guaranteeing the presence of the relation.

SEEK (Jouis, 1994) also used a combination of syntactic and lexical information, identifying lexico-syntactic patterns for the acquisition of conceptual relationships such as IS-A and PART-OF. Jouis claimed that general patterns of this type can be specified a priori and are domain-independent.

Paice and Jones (Paice and Jones, 1993) mapped surface forms to semantic types without the use of syntax, for the purpose of automatic abstract generation. Their aim was to select filler strings for various semantic roles, initially using manually constructed semantic categories. They were slightly more successful, mainly because they were not attempting to find every possible lexical pattern for a given semantic relation, but only the most typical ones. In global terms, therefore, they achieved high precision but low recall.

Montemagni and Vanderwende produced similar findings, claiming that for the extraction of semantic relations, only very basic semantic patterns can easily be extracted using lexical patterns. They argued that only patterns based on syntactic information are reliable, and that patterns based on lexical information are limited when it comes to capturing all the small variations:

> "...the most interesting patterns that identify the differentiae and the (possibly embedded) semantic relations expressed therein rely on complex structural information, information which cannot be expressed adequately in string patterns."(Montemagni and Vanderwende, 1993)

They also found that there was no need to use a specialised grammar to capture the syntactic patterns, and that a general broad-coverage grammar was just as effective. Again, whilst they were not trying to cluster sentences or phrases, the principles are still relevant to our work, since our clustering task requires that we group phrases by semantic similarity.

Some successful results have been achieved using only syntactic patterns. Tapanainen and Jarvinen (Tapanainen and Jarvinen, 1994) collect information from a tagged corpus using a "sentence axis", which is a syntactic pattern describing the sentence at a certain level, depending on each case. Groups of axes are formed, which are then used as a means of disambiguation for parsing.

SEXTANT (Grefenstette, 1994) also uses syntactic information as a bootstrapping mechanism for extracting semantic knowledge. Grefenstette's aim is to exploit patterns which are easily identifiable, thereby avoiding the need for the deeper forms of knowledge which only humans really have the power to make use of. Grefenstette claims that lexical syntactic approaches provide a middle ground between simple word counting and knowledge-rich approaches, and that a certain level of syntactic analysis is necessary and possible without requiring rich knowledge structures to be associated with each word. Grefenstette uses a combination of Ruge's method (Ruge, 1991), which restricts the textual window to noun phrases and Hindle's method (Hindle, 1990), which uses contexts such as noun-verb combinations, by using as the context all the nouns and adjectives that modify the word, as well as the verbs entering into relations with that word. Grefenstette's method does not deal specifically with terminology but only with general language, but again, the techniques used are relevant.

On consideration, it appears that there is a definite advantage to be gained over using syntactic rather than lexical patterns, unless only a small and relatively predictable set of semantic relations is to be acquired. The type of syntactic information used is also an important factor. Using more detailed information, particularly concerning the relationship between words or lexical units, may be beneficial, although the increase in precision may not offset the disadvantage of the more complex structures needed.

7.3.3 Information formatting in the medical field

Recent work in the medical field has focused on using domain-specific semantic information to create frames or templates which can then be used to structure the information found in patient records and other similar textual resources. One example of this is the *proximity processing* method developed by the medical informatics team at Geneva University State Hospital (Morel-Guilllemaz et al., 1990), (Baud et al., 1992). The idea is to make use of pre-defined conceptual knowledge about the domain without the need for any formal parsing or grammars, although syntactic information can be used to help disambiguation if available.

Do Amaral and Satomura (Do Amaral and Satomura, 1995) used domain-specific semantic information from SNOMED in conjunction with a formal grammar to process, structure and represent medical information in a language-independent frame. They claim that this association of medical and linguistic information is both necessary and adequate to explain the relationship between concepts and to represent the structure of the phrase. Their method involves allocating semantic classes from SNOMED to instances of terms, in conjunction with syntactic classes for verbs and some other words, to form a pattern consisting of a combination of the two. They identified the following 6 basic patterns as occurring most commonly:

```
<PROCEDURE> of the <TOPOGRAPHY> verb <RESULT>
<PROCEDURE> verb <RESULT>
<TOPOGRAPHY> <PROCEDURE> verb <RESULT>
<PROCEDURE> of the <TOPOGRAPHY> and <TOPOGRAPHY> verb <RESULT>
<TYPE> <PROCEDURE> verb <RESULT>
<TYPE> <PROCEDURE> of the <TOPOGRAPHY> verb <RESULT>
```

Using a corpus of case reports from the New England Journal of Medicine, they then produced language-independent frames for sentences. In a very simplified format, an example frame for the sentence *An ultrasonography of the abdomen revealed hepatomegaly* would be represented by:

```
[PROCEDURE: ultrasonography
 | of TOPOGRAPHY: abdomen
  | verb RESULT: heptomegaly]
```

Most of the work in the medical field revolves around creating semantic and/or syntactic frames from texts such as patient records, in an attempt to create some kind of neutral concept representation system - often for use in translation either between different languages (Do Amaral and Satomura, 1995) or between different coding systems (Rocha et al., 1994). The aim is therefore centred on identifying how the concepts in a sentence are related, and in forming a representation of the whole sentence. Any relationship between sentences tends therefore to be ignored, so that, for example, it is unimportant whether a term that occurs in two different sentences has the same meaning in each case, or whether the same concept is described in different ways in different sentences. Our approach is somewhat different in two ways. Firstly we focus on the contexts surrounding a term rather than the whole sentence. Although these contexts inevitably contain other terms, the point is that we are trying to compare contexts independently of the terms with which they appear, rather than in conjunction with them. Secondly, our aim is not to provide some kind of interlingual representation for sentences (or contexts), but rather to provide a representation which permits us to make comparisons about contexts. This means that we can make generalisations to a much greater extent. The similarity lies in the fact that in both cases, the aim is to provide a representation that will appropriately capture the meaning of the text.

7.4 Clustering Contexts Using Semantic Frames

The clustering method we adopt involves the creation of semantic frames from the contexts. Rather than using lexical pattern matching, which, as we have seen, can be problematic, we first collect syntactic patterns from the corpus. Then, after a normalisation process has been applied, the patterns are mapped to semantic frames, using the information already acquired in the SNC-Value method. One particular advantage of this method is that it caters for mutual interchangeability of both syntax and semantics, i.e. unknown words can be replaced both syntactically and semantically with an equivalent word from another sentence in the cluster.

Since clustering is an extensive topic in itself, our work here is limited to a small semi-automatic prototype rather than a completed system with a full range of syntactic and semantic patterns. This, however, should be sufficient to demonstrate the potential of the approach and its feasibility.

7.4.1 Acquiring syntactic information

The first stage in the process is to extract syntactic patterns from the corpus of contexts produced from the eye pathology records in the first part of this work – the term extraction process. For this we use a set of tools developed by researchers at the University of Pennsylvania (Joshi and Srinivas, 1994). This consists of a syntactic analyser providing part-of-speech information, a chunker which segments noun and verb phrases, and a dependency analyser which determines the relationships between these chunks. The tools are used on our corpus to segment and label the text accordingly.

Step 1: annotating words with tags and supertags

The first step is to separate the text into single words and to annotate each word with its part of speech and detailed syntactic information. The sentence
"*On one aspect a short length of descemet's membrane and some rather tattered trabeculae are seen.*"
is tagged as follows:

```
on//IN//B_Pnxs
one//CD//B_Dnx
aspect//NN//A_NXN
a//DT//B_Dnx
short//JJ//B_An
length//NN//A_NXN
of//IN//B_nxPnx
descemet's//NNP//B_Nn
membrane//NN//A_NXN
and//CC//B_nx1CONJnx2
some//DT//B_Dnx
rather//RB//B_ARBa
tattered//JJ//B_An
trabeculae//NN//A_NXN
are//VBP//B_Vvx
seen//VBN//A_nx0V
.//.//A_PU
...EOS...//...EOS...
...EOS...//...EOS...
```

The first part of each line is the word being tagged. The second part is the syntactic category – for example, NN stands for a common noun, JJ for an adjective, etc. A detailed explanation of the tags is given in the Appendix. The third part is the more detailed syntactic information (supertags). For example, A_NXN stands for a noun phrase as an argument, B_An is an adjective modifying a noun, B_ARBa is an adverb modifying an adjective, B_nx1CONJnx2 is a noun phrase conjunction, and so on. Again, detailed explanations can be found in the Appendix.

Step 2: dependency annotation

The second step involves annotating each word according to its dependency relations with the surrounding words. Each word in the sentence is allocated a number based on its position within the sentence. Three types of dependencies are possible:

- **modifiers** are annotated with "*****"
- **complements** are annotated with "**.**"
- **uninstantiated** dependencies are annotated with "**?**".

The above sentence is annotated as follows:

```
0 on B_Pnxs  2.  15*
1 one B_Dnx  2*
2 aspect A_NXN
3 a B_Dnx  5*
4 short B_An  5*
5 length A_NXN
6 of B_nxPnx  5*  8.
7 descemet's B_An  8*
8 membrane A_NXN
9 and B_nx1CONJnx2  8*  13.
```

```
10 some B_Dnx  13*
11 rather B_ARBa  12*
12 tattered B_An  13*
13 trabeculae A_NXN
14 are B_Vvx  15*
15 seen A_nx0V  5.
16 . B_sPU  15*
...EOS...
```

For example, *some* (10) and *tattered* (12) are modifiers of *trabeculae* (13), whilst *rather*(11) is a modifier of *tattered* (12). *Of* (6) is both a modifier of *length* (5) and a complement of *membrane* (8). Note that the definitions of complement and modifier may not necessarily equate to the standard definitions used in theoretical linguistics and are perhaps slightly "arcane". However, they were designed to be of practical usage and we see no especial benefit in making any changes to them. The way in which we make use of them will be explained later.

Step3: chunking

The third step is to divide the sentences into chunks or segments, which tend to be noun and verb phrases. Prepositions, adverbs, conjunctions, etc. are generally left as individual words. The original sentence is chunked as follows:

```
0 on B_Pnxs  1.  7*
1 one_aspect A_NXN
2 a_short_length A_NXN
3 of B_nxPnx  2*  4.
4 descemet's_membrane A_NXN
5 and B_nx1CONJnx2  4*  6.
6 some_rather_tattered_trabeculae A_NXN
```

```
7 are_seen A_nx0V  2.
8 . B_sPU  7*
...EOS...
```

Each chunk is allocated a number based on its position in the sentence. Then it is allocated a supertag (based on the same set of supertags as previously in Step 1) and finally dependency relations are assigned (in the same way as before, but this time relating to other chunks rather than to individual words). In the sentence above, the whole phrase *some rather tattered trabeculae* becomes one chunk and is assigned the supertag A_NXN (noun phrase as an argument). This is in many ways more useful than assigning the tag to just the noun *trabeculae*. Similarly, the dependency relations refer to chunks rather than individual words, which makes more sense. Using chunks in this way also eliminates unnecessary information and makes the whole tagging result more succinct and readable, as well as more easily manipulable.

7.4.2 Identifying syntactic patterns

Our initial aim is to identify syntactic patterns relating to chunks of context that we can make use of for clustering. Some types of regular patterns are likely to contain similar types of context. Once contexts have been identified and grouped together based on their syntax, it becomes easier to examine their semantic content.

Initially we use the chunks determined by our analysers to define 4 very general types of pattern:

1. NP + V e.g. *corneal_endothelium can_be_identified*
2. V + NP e.g. *covers trabecular_meshwork*
3. NP + P + NP + V e.g. *capillary_proliferation around pupil causing*
4. V + NP + P + NP e.g. *occupied posterior_nasal_portion of globe*

We do not make the patterns too specific at this stage because we are more interested in capturing generalities than specificities. Rather than making distinctions, our aim at this stage is more to group together similarities. The reason for making a distinction between the patterns containing a prepositional phrase and those without is because the relationship between the noun and its preceding or following verb is not the same. In the example given in (3), the dependency is between the verb and the first NP, rather than between the verb and the NP immediately preceding it. In the example given in (4), the dependency is between the verb and the second NP, rather than between the verb and the NP immediately following it. This can be seen more easily in the following examples of annotated chunks.

```
(i)   |0 corneal_endothelium A_NXN |1 can_be_identified
      A_nx0V  0.
(ii)  |3 covers A_nx0Vnx1  2. 4. |4 trabecular_meshwork
      A_NXN
(iii) |11 capillary_proliferation A_NXN |12 around B_nxPnx
      11* 13. |13 the_pupil A_NXN |14 causing A_nx0Vnx1
      11. 15.
(iv)  |1 occupied A_nx0Vnx1  0. 2. |2 posterior_nasal_
      portion A_NXN |3 of B_nxPnx  2*  4. |4 the_globe A_NXN
```

Looking at the dependency relations between Verb and NP (marked by a dot), we see the following. In (i), the NP-V example, the relation is between 0 and 1, i.e. adjacent chunks. In (iii), the NP-P-NP-V example, the relation is not between adjacent chunks (13 and 14) but between the first NP and the Verb (11 and 14). In (ii), the V-NP example, the relation is again between adjacent chunks (3 and 4). In (iv), the V-NP-P-NP example, the relation is also between adjacent chunks, although a PP is present as in example (iii), i.e. the relation is between

1 and 2 and not between 1 and 4.

For each of these patterns, we then make a distinction between active and passive verbs. This is so that they can later be normalised, as described in Section 7.4.5. Although it may seem strange that we are distinguishing between them only to normalise them again, this is because we want to treat them in a different way. We wish to normalise them so that their meaning is the same, although the surface order of the constituents may be different. For example, we wish the two phrases "enter the cornea" and "the cornea is entered" to be considered as meaning the same. At the moment, they are treated differently because of their linear order. Also, we need to distinguish between cases where the surface order is the same but the meaning is different. For example, *the central region encloses* does not have the same meaning as *the central region is enclosed*, but unless we differentiate between active and passive verbs, both will be considered as the same.

We also have the question of how gerunds and gerundives should be represented. Our guiding principle on this issue is to follow the syntactic tags determined by the analysers. The distinction between gerunds and gerundives is not always an easy one and is often the source of ambiguity. For example, the phrase *flying planes* could refer to the act of a person flying planes or to planes which are in flight. In the first, *flying* is a gerund and should be treated as a verb, and in the second it is a gerundive and should be treated as an adjective. Usually it requires context to solve this problem, and so we have left the tags as they have been assigned, so that gerunds tagged as verbs are treated in exactly the same way as other verbs, and gerundives tagged as adjectives are treated like other adjectives.

This leaves us with the following 7 syntactic frames:

NP + V(active)

NP + V(passive)

Syntax	Example
NP-V(act)	fibrous_proliferation has_occurred tumour has_invaded
NP-V(pas)	fairly_extensive_haemorrhages are_seen globe was_opened
V(act)-NP	show follicular_arrangement is_infiltrating muscle
V(pas)-NP	preserved trabeculae can_be_identified peripheal_piece
NP-P-NP-V(act)	nuclei of cells contain capillary_proliferation around pupil causing
NP-P-NP-V(pas)	cells in mitosis can_be_seen dermis around tumour is_densely_infiltrated
V(act)-NP-P-NP	shows skin with acanthosis have_reached middle of sclera

Table 7.3: Examples of syntactic frames

V(active) + NP

V(passive) + NP

NP + P + NP + V(active)

NP + P + NP + V(passive)

V(active) + NP + P + NP

The reason there is no 8th frame – V(passive) + NP + P + NP – is because no examples of this existed in our corpus, although potentially they could. There is provision for this case if necessary. Examples of each of these frames are shown in Table 7.3.

7.4.3 Identifying semantic patterns

Once the contexts have been clustered according to their syntactic patterns, the next task is to identify their semantic content. First we carried out an assessment of the verbs in the contexts. The verbs occurring in the syntactic frames

Frequency	No. of Verbs	Total Occurrences	% of Total Occurrences
223	1	223	28
123	1	123	15
90	1	90	11
70	1	70	9
23	1	23	3
18	1	18	2
17	1	17	2
12	1	12	2
11	2	22	2
10	1	10	1
9	1	9	1
8	1	8	1
7	4	28	4
6	5	30	4
5	2	10	1
4	4	16	2
3	8	24	3
2	10	20	3
1	45	45	6

Table 7.4: Distribution of verbs appearing in syntactic frames

Verb	Frequency	% of Total Occurrences
show	223	28
contain	123	15
cut	90	11
infiltrate	70	9
bear	23	3

Table 7.5: Distribution of most frequent verbs in syntactic frames

Class	Abbreviation	Examples
Affect	<aff>	change, alter, affect
Description	<desc>	occur, define
Observation	<obs>	show, feature, present, indicate
Position	<pos>	occupy, envelop, surround, penetrate
Procedure	<proc>	cut, remove, sever, apply

Table 7.6: Verb classes

were normalised to their root form, and their distribution and frequency were calculated. Table 7.4 shows the number of different verbs occurring with each frequency. For example, only one verb occurred with a frequency of 223, but 45 different verbs occurred with a frequency of 1. It also shows the total number of occurrences for each frequency, and the percentage of the total number of verb occurrences. For example we can see that 28% of the total occurrences of verbs were from one verb (show) which occurred 223 times. So this one verb accounts for a far larger proportion of the total occurrences than any other single verb, and also for a larger proportion than the total of all the verbs which only occur a single time. A complete list of verbs and their individual frequencies can be found in the Appendix. The analysis shows that there is a very small subset of verbs which are used with high frequency. Table 7.5 shows the top 5 most frequently occurring verbs, which in total account for 66% of all occurrences. This is not surprising, given that the corpus is of a specialised nature and consists of a fairly restricted vocabulary. Sublanguages tend to be restrictive in terms of both syntax and semantics, which makes the whole clustering process much more feasible.

Given that there is such a small number of verbs accounting for most of the occurrences, it is relatively simple to create semantic classes for them. Since verbs are not widely represented in the UMLS, we used the WordNet synsets for inspiration and manually created an initial set of verb classes, based on the verbs we found in the syntactic frames, as shown in Table 7.6. For convenience, each class is allocated an abbreviation, as shown in the table. We shall use these

Class	Abbreviation	Examples
Body Part	<BP>	corneal_epithelium, retina
Abnormality	<AB>	epithelium_oedema, keratinous_debris
Phenomenon	<PHEN>	spontaneous_rupture
Finding	<FIN>	ocular_haemorrhage
Virus	<VIR>	papilloma_virus
Injury	<INJ>	perforating_wound
Procedure	<PROC>	scanning_electron_microscopy
Body Substance	<BS>	cholesterol_crystals
Section	<SEC>	lower_third, piece, area
Size	<SI>	width, depth, circumference

Table 7.7: Noun classes

abbreviations where appropriate from this point.

Each class contains a list of the possible verbs which can belong to it, and each verb occurring in a frame is then automatically assigned to a semantic class. Note that the verbs are not normalised to their root form at this stage, because we wish to preserve the information relating to the inflectional morphology of the verb for later use (for example to help determine whether a verb is active or passive).

The noun phrases are then also allocated a broad semantic category based on the types found in the UMLS semantic hierarchy. This is carried out by generalising the semantic tags allocated earlier for the term extraction process. Each noun class comprises a set of semantic tags, so that the categories can be assigned in a similar way to the verb classes.

Initially, we have identified 10 broad noun classes, depicted with their abbreviations in Table 7.7. The last two classes (SECTION and SIZE) do not appear in the UMLS, but have been added to accommodate various groups of nouns which do not fit into any of the other classes but are important semantically.

7.4.4 Mapping from syntax to semantics

Armed with this information, each instance of a syntactic pattern can then be automatically mapped to a semantic pattern. For each syntactic pattern, the constituents are assigned a semantic class. If the noun phrase consists of a single noun, the semantic tag allocated previously is simply transformed into its appropriate corresponding semantic class. If the noun phrase consists of a multi-word term, the same thing applies, since the term will have had a semantic tag from UMLS allocated previously. If neither of these apply, there are two remaining possibilities. Firstly, the noun phrase might contain a term but be modified by something, such that the whole does not constitute a multi-word term. For example, the noun phrase *medium_sized_artery* is not a multi-word term although the noun *artery* on its own is. In this case, *artery* will have a semantic tag (and therefore semantic class) available, and this can be extended to the whole noun phrase since we can assume that the modifiers do not change the meaning of the noun to which they are attached (i.e. that a medium-sized artery is a kind of artery).

Secondly, the noun phrase might contain a group of words, none of which are terms. In this case, the noun is checked against a supplementary list to see if it may fit into one of the existing semantic classes. If so, it is assigned to this class; if not, the whole frame is rejected. This supplementary list was created manually by taking an initial set of frequently occurring words for which no class had been assigned and deciding whether they fitted into a particular category. The “Section” class is largely composed of these words, since they are not likely to be terms. If it becomes evident that other sets of words occur frequently for which there is no class established, these can be added.

Let us take as an example the set of contexts which fit the syntactic pattern Verb + NP + Prep + NP, and where the verb is of the *observation* category. The

Verb	NP	Prep	NP
shows<obs>	disruption <FIN>	of	lens <BP>
shows<obs>	atrophy <AB>	of	mid_zone <SEC>
shows<obs>	atrophy <AB>	of	muscle <BP>
containing<obs>	elements <BS>	of	pilary_complex
containing<obs>	crystals <BS>	in	subretinal_space <SEC>
shows<obs>	erosion <AB>	of	superficial_lamellae <BP>
shows<obs>	vascularization <AB>	of	superficial_stroma<BP>
showing<obs>	acanthosis <AB>	with	typical_molluscum_bodies<BP>
shows<obs>	reduction <FIN>	of	visual_cell_layer<BP>
shows<obs>	skin <BP>	with	acanthosis <AB>
shows<obs>	proliferation <FIN>	of	capillaries <BP>
shows<obs>	disruption <FIN>	of	epithelium <BP>
shows<obs>	disruption <FIN>	of	lens <BP>
shows<obs>	atrophy <AB>	of	muscle <BP>

Table 7.8: Strings matching the pattern V(act)<obs> N Prep N

results of the initial mapping procedure are illustrated in Table 7.8. Each noun phrase is allocated a noun class derived from its semantic category in UMLS. Some of these, e.g. *fibrous_tissue* and *muscle*, can be immediately allocated their class, since they appear in the text exactly as found in UMLS, i.e. they are classified by UMLS as a term. Others are not classed as terms in UMLS, but contain terms within the noun phrase which are. For example *superficial_lamellae* contains the term *lamellae*, and *visual_cell_layer* contains the term *cell_layer*. Others do not contain UMLS terms at all, but are nevertheless possible to classify, such as *mid-zone*. Finally, there may remain noun phrases which are impossible to classify and which are therefore likely to be either words or terms unknown by UMLS and by the user, or which are simply errors, for example *pilary_complex*.

From the 4 basic syntactic patterns we can produce a set of mappings to semantic patterns depicted in Table 7.9. For each syntactic frame, we can see that a number of semantic patterns exists, but that the syntactic categories do impose some restrictions.

Syntax	Semantic Pattern	Example of Lexical Form
NP-V(Act)	<AB><aff>	basaloid_proliferation affecting
	<BP><obs>	corneal_epithelium shows
NP-V(Pas)	<BP><desc>	corneal epithelium is_attenuated
	<BP><obs>	retina represented
	<SEC><proc>	sections were_taken
	<BP><proc>	optic_nerve cut
	<AB><proc>	basal_cell_papilloma removed
	<SEC><pos>	small_sections are_separated
	<BP><pos>	vitreous_chamber traversed
	<AB><pos>	foamy_histiocytes surrounded
V(Act)-NP	<aff><BP>	affecting epidermis
	<obs><AB>	shows laminar_splitting
	<proc><SEC>	cut edges
	<pos><AB>	traverses tumour
	<pos><BP>	crosses cornea
NP-Prep-NP-V(Act)	<SEC>Prep<BP><obs>	small_fragment of conjunctiva showing
	<SEC>Prep<BP><pos>	lower_third of cornea crossing
	<SEC>Prep<AB><pos>	tip of tumour has_invaded
V-NP-Prep-NP(Act)	<obs><SEC>Prep<BP>	shows small fragment of skin
	<obs><AB>Prep<BP>	showing proliferation of epidermis
	<pos><SEC>Prep<BP>	crossing lower_part of cornea
	<pos><AB>Prep<SEC>	shows bullous_separation in centre
	<pos><BP>Prep<BP>	to_envelop lower_half of iris
	<pos><BP>Prep<SEC>	to_envelop eye at peripheral_border

Table 7.9: Mappings from syntactic to semantic patterns

7.4.5 Creating semantic frames

The next stage is to create semantic frames which can be filled by normalised lexical forms. The semantic frame incorporates the syntactic pattern from which it was derived. In practice this just means combining the active and passive verbs such that the semantic pattern is reordered accordingly, using rules such as the following:

```
1. NP + V(Act) --> sem(NP) + sem(V)


2. V(Act) + NP --> sem(V) + sem(NP)


3. NP + V(Pas) --> sem(V) + sem(NP)
```

This translates as meaning that where we have a syntactic pattern consisting of a NP followed by an active verb, as in (1), a frame is created consisting of the semantic category for the NP followed by the semantic category for the verb (with no special feature to indicate whether it is active or passive). Where we have a syntactic pattern consisting of an active verb followed by an NP, as in (2), a frame is created consisting of the semantic category for the verb followed by the semantic category for the NP. Where we have a syntactic pattern consisting of an NP and a passive verb, as in (3), a semantic frame is created consisting of the semantic category for the verb followed by the semantic category for the NP. This has the effect of collapsing the syntactic patterns (2) and (3) into one semantic frame.

Once the frames have been created, the surface lexical forms can be transformed into slot fillers. First the verbform is normalised to its root form, and then any passive forms are reordered with the same set of rules as above. Table 7.10 depicts the original syntactic pattern, the semantic frames created and examples of the fillers for the frames (i.e. the normalised contexts).

Syntax	Semantic Frame	Slot Filler Example
NP-V(Act)	<AB><aff>	basaloid_proliferation affect
	<BP><obs>	corneal_epithelium show
NP-V(Pas)	<desc><BP>	attenuate corneal epithelium
	<obs><BP>	represent retina
	<proc><SEC>	take sections
	<proc><BP>	cut optic_nerve
	<proc><AB>	remove basal_cell_papilloma
	<pos><SEC>	separate small_sections
	<pos><BP>	traverse vitreous_chamber
	<pos><AB>	surround foamy_histiocytes
V(Act)-NP	<aff><BP>	affect epidermis
	<obs><AB>	show laminar_splitting
	<proc><SEC>	cut edges
	<pos><AB>	traverse tumour
	<pos><BP>	cross cornea
NP-Prep-NP-V(Act)	<SEC>Prep<BP><obs>	small_fragment of conjunctiva show
	<SEC>of<BP><pos>	lower_third of cornea cross
	<SEC>of<AB><pos>	tip of tumour invade
V-NP-Prep-NP(Act)	<obs><SEC>Prep<BP>	show small fragment of skin
	<obs><AB>Prep<BP>	show proliferation of epidermis
	<pos><SEC>of<BP>	cross lower_part of cornea
	<pos><AB>Prep<SEC>	show bullous_separation in centre
	<pos><BP>Prep<BP>	envelop lower_half of iris
	<pos><BP>Prep<SEC>	envelop eye at peripheral_border

Table 7.10: Examples of semantic frames with fillers

7.4.6 An example

Let us illustrate the whole procedure with an example. Consider the following contexts:

crosses cornea

vitreous_chamber traversed

traverses_tumour

optic_nerve surrounded

envelops iris

After allocating syntactic and semantic categories we get the following:

```
crosses_cornea              V(Act) + NP <pos><BP>
vitreous_chamber traversed  NP + V(Pas) <BP><pos>
traverses_tumour            V(Act) + NP <pos><BP>
optic_nerve surrounded      NP + V(Pas) <BP><pos>
envelops iris               V(Act) + NP <pos><BP>
```

Using Rules 2 and 3, we produce the following semantic frames:

```
crosses cornea             <pos><BP>
vitreous_chamber traversed <pos><BP>
traverses tumour           <pos><BP>
optic_nerve surrounded     <pos><BP>
envelops iris              <pos><BP>
```

The lexical forms are then normalised and reordered to produce the final completed frame:

```
<pos><BP> : cross cornea
```

```
<pos><BP> : traverse vitreous_chamber
<pos><BP> : traverse tumour
<pos><BP> : surround optic_nerve
<pos><BP> : envelop iris
```

Although originally the contexts appeared to have little in common, in this format it can be seen clearly that they are in fact very similar, and they will therefore be clustered together in one group.

If we now come across a context of the form “X traverses vitreous chamber”, where X is a candidate term, we can be fairly sure that it will have the same meaning as X found in the context “X crosses the cornea”. So from an initial unordered bag of contexts, we can produce clusters of contexts which have the same syntactic and semantic properties. Table 7.11 depicts an example for the semantic frame <pos><BP>, where all the contexts consist of a “position” verb and a “Body Part” noun phrase. In the left hand column is the original context, and in the right is the normalised form. Clearly the normalised versions are easier to recognise as being similar than the original contexts.

7.5 Discussion

The method we have described for context clustering is only an outline and is not a full working system, since this is only one part of our work. Its aim is rather to demonstrate the feasibility of our idea and to show how the contexts identified in the term extraction process can be put to further use given the semantic information acquired earlier and some deeper syntactic information. In this section we shall discuss some of the finer points about our particular approach, both positive and negative, and explain how it may be expanded further.

Original Context	Normalised Form
occupied posterior_nasal_portion	occupy posterior_nasal_portion
has_entered nerve_head	enter nerve_head
traverses cornea	traverse cornea
to_invade retina	invade retina
vitreous_chamber traversed	traverse vitreous_chamber
crossing limbus	cross limbus
naevoid_area is_separated	separate naevoid_area
to_enter eye	enter eye
overlying pannus	overlie pannus
have_penetrated fibrous_tissue	penetrate fibrous_tissue
crosses cornea	cross cornea
corneal_endothelium has_separated	separate corneal_endothelium
traverses peripheral_cornea	traverse peripheral_cornea
invaded macular_region	invade macular_region
occupy pupil	occupy pupil
lens is encased	encase lens

Table 7.11: Contexts and their normalised forms

7.5.1 Clustering errors

Naturally, we cannot assume that all the clusters will be perfect. A certain amount of noise and omissions is only to be expected. Like most others, we aim for accuracy over recall in this respect, since erroneous classification tends to be more harmful than incomplete classification.

Table 7.12 shows some examples of different context clusters created. The asterisked lexical strings denote errors. Three main reasons can be identified for erroneous clustering: syntactic tagging errors, chunking errors and semantic classification errors. Syntactic tagging errors result in a word or chunk being allocated the wrong syntactic tag, for example *somewhat* is identified as an NP. Chunking errors result in a word or group of words being incorrectly identified as a chunk, for example the phrase *both globes cut* is incorrectly divided into the two chunks *both* and *globes_cut*, instead of *both_globes* and *cut*. Similarly, *nasal_side_the_retina* is wrongly identified as a chunk. Semantic classification errors result in a chunk being allocated an incorrect noun or verb class. For example, *cholesterol_crystals*

Pattern	Lexical String
$< obs >< AB >$	show chronic_inflammatory_changes see atrophic_changes show vacuolar_changes show gross_liquefactive_changes see hyaline_degenerative_changes show hyaline_degenerative_changes
$< BP >< pos >$	track cross large_vessel enter track cross adventitia overlie vessel enter
$< proc >< BP >$	cut optic_nerve shed cells fold retina take membrane open globe displace lens remove cornea
$< proc >< AB >$	remove basal_cell_papilloma
$< BP > Prep < BP >< aff >$	nerve behind cribriform_plate affect **considerable_traction on ciliary_processes affect
$< aff >< AB > Prep < BP >$	show degenerative_changes in collagen show bullous_separation in axial_region **show nullous_separation of epithelium show bullous_separation in central_region
$< aff >< SEC > of < BP >$	affect lower_half of retina affect all_layers of cornea affect pigmented_epithelium of peripheral_retina affect adjacent_skin to small_extent
$< pos >< SEC > of < BP >$	occupy posterior_nasal_portion of globe traverse horizontal_axis of cornea overlie external_surface of tumour occupy temporal_third of anterior_chamber envelop lower_half of iris separate bundles of cells
$< pos >< BP > Prep < LOC >$	enter eye at peripheral_border invade angle in planes **traverse cornea in somewhat

Table 7.12: Examples of final clustered contexts

is tagged as *Body Part*, but it should really be a *Phenomenon* (in line with terms such as *calcium_deposit* which is tagged as *Phenomenon*). In this case, the error lies with the UMLS classification. Errors are also possible when a multi-word term is not found in UMLS. In this case, if the head noun is found to be present, the semantic tag for this is used for the whole term. This assumes that a term will have the same semantic category as its head noun, but this is not always the case. We did not find any examples where this type of error actually occurred, but with a more comprehensive system or a larger corpus this is certainly a possibility.

One advantage of our method, however, is that because we do not rely on lexical pattern matching, some types of errors such as typographical errors can be ignored without being rejected. In the term recognition method described earlier, this kind of error is problematic because misspelled terms will still be recognised but will be treated as separate terms from their correctly spelled versions, thereby leading to statistical differences. Although this is not such a problem where only one occurrence of a frequently occurring term is misspelled, some terms suffer from frequent misspellings or simply have variant spellings, and this can be more problematic. This is particularly the case with hyphenated terms and terms containing apostrophes, which are not always used consistently. The clustering method has the advantage of ignoring these differences. Sometimes it may ignore a misspelled word because it cannot allocate a semantic class to it, but as we have seen, there are various safety mechanisms in place to deal with precisely such cases. For example, the term *bullous_separation* occurs in the text misspelled as *nullous_separation*. Although this cannot be allocated a semantic tag initially (since it does not occur in the UMLS Metathesaurus), the term *separation* does occur and can thus be allocated a tag. As an untagged NP chunk, the whole term *nullous_separation* takes on the class allocated to the noun *separation*, and the whole context in which it occurs will be clustered into the same set as any occurrences of the correct term *bullous_separation*.

7.5.2 Granularity of clustering

The level of granularity of the clustering mechanism is an issue which needs to be considered. This is mostly dependent on what the clusters are required for. Some applications may require a very broad grouping while others may require more fine-grained distinctions. It may not even be dependent on the type of application but on the particular instance of application. For example, a clustering technique which is used for ontology building might require a different level of granularity according to the domain or text type or even for the ultimate use of the ontology. So a general mechanism for context clustering ideally needs to be flexible in its level of granularity so that it can be used for different purposes with the minimum of tuning. The method we have described is initially set at a low level, such that clustering is in broad groups. If necessary, it can be made more fine-grained by the addition of finer semantic distinctions between classes. This would be achieved by tightening restrictions on which semantic categories constitute a semantic class, for the noun phrases, and for the verbs by subdividing the semantic classes into finer distinctions. For example, the verb class *Procedure* could be subdivided into *Cutting*, *Removing*, *Inserting* and *Altering* Procedures.

7.5.3 Further expansion

As we have explained earlier, the context clustering is not a complete system, but rather an example of how such a system might be achieved. In particular, only a representative sample of syntactic patterns has been identified and implemented, focusing on verb phrases. However, other patterns also need to be considered, especially phrases modifying a term which themselves do not form part of a larger term.

The allocation of semantic classes is also not complete, nor is it fully automated, relying largely on the information from UMLS for the nouns and on the most frequently occurring verbs. Although verbs which occur infrequently are statistically less important, it would nevertheless be beneficial to include as many of these as possible. Naturally the same applies to those nouns which have not been allocated a semantic class.

Chapter 8

Evaluation

In this chapter, we examine the results of the SNC-Value and discuss both its effectiveness as a whole, and that of its individual components. We first investigate some of the methods used for evaluation in terminology, and the problems associated with these methods. Secondly, we compare two methods of evaluation. The first compares the results with manually identified terms, while the second compares them with terms from a pre-defined list (the UMLS Metathesaurus). Thirdly, we analyse each of the context weights to see how well they perform against expectation, and whether the premises on which they are based can be validated. Finally, we examine the overall performance of our system and draw some conclusions.

8.1 Evaluation of ATR

One of the problems with evaluating term recognition systems is that there is no real "gold standard" by which they can be measured. In this respect, it is similar to the problem of evaluating MT systems, for there is no one correct solution for either translation or term recognition. This problem has been discussed at length in the field of MT (Lehrberger, 1982; Hutchins and Somers, 1992; Arnold, 1993; Sager, 1994; Cole, 1996). MT systems tend therefore to be evaluated in a number of ways – not only in terms of "correctness", but in terms of factors such as usability, speed, level of automation, "readability" and domain portability. They can also be compared with either human translators or other systems. Term extraction systems are generally evaluated with respect to manual extraction but, as with MT systems, this is not always sufficient since manual techniques are subjective and prone to error.

Another problem is that the evaluation should be tailored to each system, and is therefore likely to depend on the ultimate use of the system. A distinction can be made between the evaluation of systems designed for general terminology and that of systems for applied terminology (Jouis and Elhadi, 1997). The former need to ensure complete coverage of the field, while the latter are tailored more towards the application and the end-user, and tend to be much more specific.

8.1.1 Comparison of ATR methods

Although studies have been carried out to compare different approaches, e.g. (Lauriston, 1995), (Kageura and Umino, 1996), these tend to be largely methodological, and it is rare that any comparison is carried out between different methods on the same data. For example, Bourigault and Habert (Bourigault and Habert, 1998) undertook a comparison of LEXTER (Bourigault, 1992) and AlethIP-NP[1] (Herviou-Picard, 1996). They did not, however, compare the overall performance, but were more interested in the internal workings of the tools and their analysis of NPs.

Even within ATR, there exists a variety of interpretations of "term recognition", such that the end results cannot always be compared. For some, term

[1] a tool developed by GSI-Erli as part of the Eureka GRAAL project

recognition is comparable to manual extraction of terms, in that there are no degrees of termhood, and the result is binary – a candidate term is either a term or a non-term. For others, terms are ranked according to their degree of termhood, which can be interpreted either as the extent to which they behave like terms, or the probability that they are terms. There is also disagreement over what should be accepted or recognised as a term – for example, whether variants should be classified as terms, whether they should be recognised separately, or whether they should not be recognised at all. Most ATR systems are also developed for a specific domain, and are not immediately portable to a new domain, nor will they necessarily have similar results if domain porting is possible. This means that it is virtually impossible to compare any two ATR systems quantitatively.

The issue of the validity of statistical methods, not only in ATR but generally in corpus linguistics has been debated previously (Dunning, 1993; Kageura and Umino, 1996). Kageura claims that many statistical methods or models are unfounded from a mathematical point of view, although intuitively they make sense. Similarly, we need to ensure that a distinction is made between statistical and linguistic or terminological importance. For example, it is often presumed that linguistic objects are more important if they occur more frequently. Although it is true that *statistically* they are indeed more important, it does not necessarily follow that they are *terminologically* more important. Methods for measuring similarity are often based largely on frequency, which does not cater for the possibility that infrequently occurring objects may nevertheless be similar.

8.1.2 Precision and recall

The method we use to evaluate the terms is based largely on precision and recall, although it differs slightly from traditional use of precision and recall in information retrieval.

Precision measures the accuracy of the items retrieved, regardless of whether it fails to retrieve everything it should. It is defined as

$$P = \frac{no.\ of\ relevant\ items\ retrieved}{total\ no.\ of\, items\ retrieved} \tag{8.1}$$

Recall measures the ability to retrieve all the correct items (regardless of what else is retrieved). It is defined as

$$R = \frac{no.\ of\ relevant\ items\ retrieved}{total\ no.\ of\ relevant\ items} \tag{8.2}$$

The problem with precision and recall is that they are only really useful when used in combination. If a system only retrieved one item, which happened to be correct, then even if 100 items should have been retrieved, it would still score 100% precision. Its recall, however, would be 1%, since it only retrieved one out of the 100 items.

If, on the other hand, the system retrieved 10,000 items, of which only 100 were correct, it would have perfect (100%) recall, but 1% precision, since only 100 out of the 10,000 items it retrieved were correct. So, a tradeoff is required between the two, which is not always easy to determine. Often, the exact balance depends on the user's needs and the type of application.

8.1.3 Manual evaluation of terms

The main evaluation procedure was carried out with respect to a manual assessment of the list of terms by 2 domain experts. There are, however, problems associated with such an evaluation. As we have mentioned, there is no gold s-tandard of evaluation. Also, manual evaluation is both fallible and subjective. One of the reasons why automatic term extraction is so important is because, in addition to being time-consuming and labour-intensive, manual extraction is often inconsistent and unreliable. If there were some perfect set of results against which we could compare systems, there would be less need for such systems in the first place.

To avoid this problem, we measure the performance of the system in relative terms rather than in absolute terms. We used two domain experts in order to offset the subjectivity process. A third would have been desirable, but practical limitations (the problems of finding a suitably skilled person willing and able to undertake this task) prevented us from using any more experts.

8.1.4 Evaluating terms against the UMLS Metathesaurus

The alternative to comparing the results with a manual evaluation is to compare them with some pre-established list of known terms. The most obvious means of doing this would be to use the list of terms provided by the UMLS Metathesaurus, since we are using this as our source of semantic information. However, it has some limitations in this respect. Firstly, it is not specific to the field of ophthalmology, but only to medicine in general. Secondly, it is organised such that only the preferred terms are actively present. Whilst there are links from each preferred term to alternative terms, these are not explicitly stated. Similarly, lexical variants are not consistently shown at this stage either. The Specialist Lexicon does provide details of inflectional information, but this is again not explicitly stated, and in the version of UMLS we used, it was not complete. Since we do not currently use stemming techniques or other means of conflating different lexical variants, it is awkward for us to manipulate these variants. The reason for this was largely due to the inconsistent way in which term variants are captured in the UMLS; as a consequence the nature of the problem was not discovered until a late stage. However, the number of terms affected by this was relatively small. As we discussed in Chapter 2, term variation and the lack of standardisation can be a significant problem in terminology. For example, in applications such as

Manual Evaluation	No. of terms	No. in UMLS	Precision
valid	826	165	20%
invalid	1728	109	6%

Table 8.1: Comparison of manual evaluation and UMLS

information retrieval and hypertext linking, if term variants are not recognised, this can lead to a serious loss of recall. In areas such as machine translation, it is, at the least, useful to be able to identify variants as being related to the same concept.

8.1.5 Comparing the manual evaluation with the UMLS

We first performed an experiment to compare the set of terms judged valid by our human experts with the set of terms in the UMLS Metathesaurus. The set of manually evaluated terms was selected from the candidate terms extracted by the C-Value, i.e. it is not a complete set of medical terms, but a set of medical terms found in the corpus. We expect the list to be incomplete because the C-Value is not likely to have extracted all the relevant terms. However, it would have been an impossible task to ask the manual experts to go through the entire corpus searching for terms, without some kind of automatic filtering (as the C-Value provides). The UMLS, on the other hand, provides a (supposedly) complete set of medical terms, many of which will never be found in the corpus. So there is necessarily a significant discrepancy between the two lists of terms. We therefore compared the two lists by measuring the number of terms judged valid and found in the UMLS, and the number of terms judged invalid and found in the UMLS. In the former case, theoretically the result should be 100%, and in the latter, 0%. Since neither list is perfect, this is evidently not going to be the case.

The results are shown in Table 8.1. The first column shows whether the terms were found to be valid or invalid according to the experts, the second column shows the number of terms in each of these two sets, the third column shows the number of terms in each set which were also found in UMLS, and the fourth column shows this figure as a percentage of the total for the set (i.e. the precision).

Although we had not expected a perfect match, the results were quite low. Only 20% of those terms judged as valid were to be found in UMLS. Clearly, both the UMLS is incomplete, and the manual assessment was highly over-generative. Furthermore, as we have discussed, term evaluation is highly subjective. Various criteria must be established as to what sort of term will be judged acceptable – for example, whether lexical or terminological variants will be permitted, whether verbs can be counted as terms, whether terms including prepositions are valid, and so on. We also found that of the terms judged as non-valid, 6% occurred in UMLS. This was a less surprising result, since we expect that UMLS may contain errors or anomalies, and that the human experts may have missed some terms (i.e. judged some valid terms as invalid).

The purpose of this comparison is not to identify the deficiencies of either the manual evaluation or UMLS, as such, but to highlight the subjectivity of term identification. Since we aim to show how our system improves on the C-Value and NC-Value, we use the same standard (that of the manual experts) to evaluate our results, rather than using the UMLS. Although the UMLS might seem a more appropriate measure of evaluation, since this is the resource being used for part of our extraction, it should be reiterated that the evaluation is a comparative rather than an absolute one. It is therefore more important that the evaluation standard is consistent.

List Section	No. of Valid Terms	Precision	Distribution
top	139	13%	51%
middle	67	8%	25%
bottom	66	8%	24%

Table 8.2: Evaluation of NC-Value extracted terms compared with UMLS terms

List Section	No. of Valid Terms	Precision	Distribution
top	346	41%	42%
middle	284	33%	34%
bottom	196	23%	24%

Table 8.3: Evaluation of NC-Value extracted terms compared with manually judged terms

8.1.6 Comparison of the NC-Value and SNC-Value rankings against the UMLS

As a further comparison, we performed an evaluation of the results of the NC-Value and SNC-Value against the UMLS Metathesaurus. To do this, we measured the number of terms in each section of the NC-Value list which were present in the UMLS. We carried out the same procedure measured against the manually evaluated terms, and compared the two sets of results (Tables 8.2 and 8.3). The tables depict the number of valid terms in each list section (out of a total of 851 terms in each section), the number of valid terms as a percentage of this total, and the number of valid terms in each section as a percentage of the total number of valid terms. In this and future results, we label these two percentages *Precision* and *Distribution*. We use the term Precision in its usual sense (Formula 8.1), and Distribution to mean the proportion of correct items[2]. Ideally, the distribution of terms will be such that a high proportion of valid terms will be found at the top of the list, and a low proportion at the bottom. This acts as a useful guide

[2]the total distribution for the list will therefore always add up to 100%

List Section	No. of Common Terms	Percentage	Distribution
top	86	31%	52%
middle	51	18%	31%
bottom	28	10%	17%

Table 8.4: Manually evaluated and UMLS-evaluated terms per list section

when comparing results.

From Tables 8.2 and 8.3, we can see that the precision scores are much higher for the manual evaluation than for the UMLS evaluation. This is to be expected for two reasons. Firstly we used the manual evaluation as a baseline when establishing and testing the NC-Value method. Secondly, there are many terms found in the corpus judged valid by the human experts which are not found in the UMLS. Regarding the distribution, we see that the valid terms are much more evenly spread throughout the list in the manual evaluation than in the UMLS evaluation. Again, we do not presume to make assumptions about the relative validity of the two forms of evaluation from this fact, but view it simply as another indication of the differences between them. The fact that more of the valid terms are to be found in the top of the list with the UMLS evaluation indicates that the terms which are both judged correct by the manual experts and found in UMLS tend to have high weights. To prove this, we measured the number of terms in each section of the NC-Value list which were both manually judged valid and found in UMLS. The results in Table 8.4 confirm this, there being more terms common to both in the top part of the list (52%), and fewest in the bottom part (17%).

We also carried out the same comparison, but with the results of the SNC-Value rather than the NC-Value (Tables 8.5 and 8.6. Again, we have similar findings, though the difference between the two methods of evaluation is slightly minimised here. This can be attributed to the fact that we are making use of

List Section	No. of Valid Terms	Precision	Distribution
top	140	13%	53%
middle	73	8%	28%
bottom	49	6%	19%

Table 8.5: Evaluation of SNC-Value extracted terms compared with UMLS terms

List Section	No. of Valid Terms	Precision	Distribution
top	349	41%	42%
middle	244	29%	30%
bottom	233	27%	28%

Table 8.6: Evaluation of SNC-Value extracted terms compared with manually judged terms

information from UMLS for the SNC-Value, but not for the NC-Value.

In the next three sections, we evaluate each of the three components of the importance weight individually, to see how well their initial premises are supported.

8.2 Similarity Weight

8.2.1 Expected theoretical performance

One of the problems with our method of calculating similarity is that it relies on a pre-existing lexical resource. For reasons which we have discussed earlier, this is not a perfect solution, since omissions and errors are likely.

Bearing in mind its innate inadequacies, we can nevertheless evaluate the expected theoretical performance of the measure by concerning ourselves only with what is covered by the thesaurus. This means that we assume completeness (although we know that this is not the case) and evaluate it accordingly, ignoring anything which may be missing. The overall practical performance of the method

	Term	Non-Term
top set	76%	24%
middle set	56%	44%
bottom set	49%	51%

Table 8.7: Semantic weights of terms and non-terms

will be discussed separately, taking into consideration not only what is actually covered, but also what should be covered and is not.

The semantic weight is based on the premise that the more similar a context term is to the candidate term it occurs with, the better an indicator that context term is. So the higher the total semantic weight for the candidate term, the higher the ranking of the term and the better the chance that the candidate term is a valid one. To test the performance of the semantic weight, we sorted the terms in descending order of their semantic weights and divided the list into 3, such that the top third contained the terms with the highest semantic weights, and the bottom third contained those with the lowest. We then compared how many valid and non-valid terms (according to the manual evaluation) were contained in each section of the list.

The results, depicted in Table 8.7, can be interpreted as follows. In the top third of the list, 76% were terms and 24% were non-terms, whilst in the middle third, 56% were terms and 44% were non-terms, and so on. This means that most of the valid terms are contained in the top third of the list and the fewest valid terms are contained in the bottom third of the list. Also, the proportion of terms to non-terms in the top of the list is such that there are more terms than non-terms, whereas in the bottom of the list there are more non-terms than terms. The evaluation therefore demonstrates two things:

- more of the terms with the highest semantic weights are valid, and fewer of those with the lowest semantic weights are valid;

- more valid terms have high semantic weights than non-terms, and more non-terms have lower semantic weights than valid terms.

8.2.2 Comparison with statistical methods of measuring similarity

Using a pre-defined thesaurus to calculate similarity has the pitfalls that the hierarchy is not necessarily even, especially when it has been created manually, as in the case of UMLS. So the distance between two nodes may actually reflect differing degrees of similarity, depending on which section of the thesaurus they occur in. Although we take into consideration the vertical position of the nodes, this is to account for the fact that nodes lower in the hierarchy are intuitively more similar, rather than for any discrepancies in the uniformity of the ontology. Various methods of regulating this problem have been proposed, e.g. (Resnik, 1995a), (Smeaton and Quigley, 1996). In almost all similarity measures using thesauri, statistical information is either solely used or at least incorporated into the measure. However, there is no clear evidence to suggest that adding statistical information will significantly improve the measure, where the statistical part is not the primary component, and where a relatively small corpus is used as the basis for frequency information.

The methods which intuitively seem most plausible are based on information content. The information content of a node is related to its probability of occurrence in the corpus. The more frequently it appears, the more likely it is to be important in terms of conveying information, and therefore the higher weighting it should receive. We performed experiments to compare two such methods with our similarity measure. The first considers the frequency of the MSCA of the two terms, whilst the second considers the frequency of the nodes of the terms being compared.

Calculating probability

The probability of a node is calculated by its frequency in the corpus. Since semantic types do not appear as such in the corpus, this frequency is found by summing the frequencies of all the terms in the corpus belonging to this semantic type. To estimate the probability of a semantic type, we divide its frequency of occurrence by the total frequency of all semantic types occurring in the corpus. This can be described more formally as follows. Firstly we define the probability of a term occurring in the corpus:

$$P(t_j) = \frac{f(t_j)}{\sum_{S \epsilon T} \sum_{t_j \epsilon S} f(t_j)} \tag{8.3}$$

where

T is the set of semantic types occurring in the corpus

S is a semantic type belonging to T

t_j is a term belonging to a semantic type S

$f(t_j)$ is the frequency of term t_j occurring in the corpus

The probability of a semantic type occurring can then be defined as:

$$P(S_i) = \sum_{t_j \epsilon S_i} P(t_j) \tag{8.4}$$

where

S_i is a semantic type

t_j is a term

$P(t_j)$ is the probability of t_j occurring in the corpus

Experiment 1: probability of the MSCA

The first experiment involves adding the probability of the MSCA to the similarity value of a term. For each combination of term and context term, the weight of the MSCA is added to the similarity weight to give a new score. The idea stems from the approach of (Smeaton and Quigley, 1996), which uses the probability of the MSCA alone to calculate similarity between words. Our rationale for using the probability of the MSCA is that the MSCA is the most significant part of the hierarchy when measuring similarity, because it is the lowest part of the hierarchy (and thus the most informative part) at which the two terms are similar. Below this point, the two terms begin to differ, so any additional information encountered will not be common to both. Thus it should be beneficial to the measurement of similarity to take particular notice of this "lowest informative point" at which the similarity between the two is greatest in terms of information yielded, and thus at this point that the probability should be measured.

Experiment 2: probability of the nodes

It could also be argued that it is the probability of the individual nodes being considered that is most relevant. If a particular semantic type is statistically important, then this is likely to reflect on all the terms belonging to that semantic type. It could therefore be claimed that two terms belonging to statistically important semantic types have a greater degree of similarity than two terms which do not. Care must be taken, however, to avoid blurring the distinction between theoretical notion of similarity and its practical usage.

The second experiment involves adding the information content (IC) weight to each term being considered. So when similarity is calculated between a term and a context term, the IC weight of the node under which each is positioned is added to the similarity weight.

The experiments incorporating the idea of information content into the similarity measure were evaluated by comparing the number of terms and non-terms found in the top, middle, and bottom sections of the list of terms ordered by

	Term	Non-Term
top set	76%	24%
middle set	57%	43 %
bottom set	49%	51%

Table 8.8: Experiment 1: semantic weights of terms and non-terms

	Term	Non-Term
top set	76%	24%
middle set	59%	41%
bottom set	48%	52%

Table 8.9: Experiment 2: semantic weights of terms and non-terms

similarity.

The results, shown in Tables 8.8 and 8.9, are remarkably similar, and in fact differ little from the original similarity measurement shown in Table 8.7 and discussed earlier. Figure 8.1 depicts the results of all three measurements in the form of a graph.

As we had originally suggested, neither method was significantly better than the other, nor did either show any real improvement over our original method. This does not necessarily mean that all three methods are equivalent, however. Each of the methods does produce different results in that the values attributed to the terms differ. This means that the terms are ranked differently in each of the three results, but in terms of overall performance this has no real impact, i.e. there are the same number of terms and non-terms in each section of the list in each method. Thus the alternative methods of calculating similarity do not significantly affect term extraction, though they may allocate slightly different similarity weights to the terms. They may, however, be more significant for tasks such as ontology building, since the differences might affect the classification of the terms.

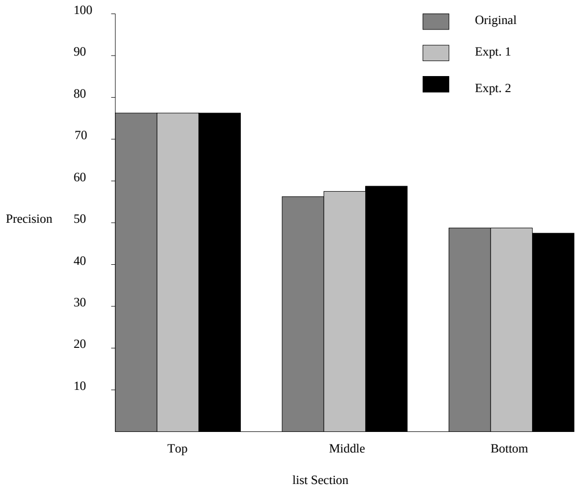

Figure 8.1: Influence of statistical methods on the similarity measurement

As we have discussed, the criteria for evaluation may be dependent on the application. The simple experiments described above show that for the purposes of our method of term ranking, any of the three similarity measures is better than none, but that there is very little difference between each one. For other applications, or with different data sets, the choice of measure might be more significant, but our purpose here is to evaluate the measures locally rather than globally, i.e. within the context of our system.

8.2.3 Practical performance

Using a thesaurus to calculate similarity has the drawback that we can only calculate similarity between two terms that exist in the hierarchy. Since we use a pre-defined thesaurus for this, it is clear that there will be problems of omission, since it cannot be expected to cover every term in our corpus. The most obvious solution is to derive a thesaurus automatically from the actual corpus. This is not, however, a trivial task and is currently beyond the scope of this work. Ushioda's statistical algorithm for the hierarchical clustering of words (Ushioda,

1996) is a method of automatically creating a dendrogram, and has been used in preliminary experiments by Mima et al. to create a hierarchy for terms (Mima et al., 1999).

Given the inherent problems of this method and the extra effort it would involve, we have not adopted this as yet, preferring to concentrate on developing a similarity measure which works, albeit on a more limited data set. Another alternative could be to use a pre-existing thesaurus but to augment it with supplementary information from the corpus, again using statistical methods. This would have the advantage of making use of the accuracy and efficiency of pre-existing knowledge, whilst ensuring that the resources are tuned to the corpus.

8.3 The Context Term Weight

We have proposed the hypothesis that context words which are terms are useful indicators of other terms, because terms do not tend to appear in isolation. This was incorporated into our measure by considering the number of times a candidate term appears with a context term in the text. The more frequently they appear together, the more useful the context term is considered to be, and the higher the weight it therefore receives.

8.3.1 Comparing context terms with context words

In the NC-Value approach, Frantzi showed that context words could be ranked according to their importance as a term indicator (Frantzi, 1998). This importance was calculated using a statistical approach, which proposed that the more frequently a context word appeared with a term, the better the indicator it would be. In the next experiment, we compare the importance of context terms with the importance of context words as indicators.

Context Words	Terms	Non-Terms
top set	56%	44%
middle set	44%	56%
bottom set	35%	65%

Table 8.10: Evaluation of context words relating to terms and non-terms

Context Terms	Terms	Non-Terms
top set	55%	45%
middle set	53%	47%
bottom set	52%	48%

Table 8.11: Evaluation of context terms relating to terms and non-terms

We performed a similar experiment with the context terms. These were sorted in order of context term weight and divided into three sets (the top set consisting of those with the highest weights and the bottom set those with the lowest). We then calculated how frequently each set of context terms appeared with both terms and non-terms. The results, depicted in Table 8.11 are similar to Frantzi's results for context words, shown in Table 8.10. Figure 8.2 depicts the combined results in a graph. The results show that the top set of context terms appeared 55% of the time with terms and 45% of the time with non-terms. So a higher proportion of the top set of context terms appeared with terms than with non-terms, and this proportion decreased slightly for the middle set of context terms and again slightly for the bottom set of context terms. This has two main implications. Firstly, the top set of context terms are better at distinguishing terms from non-terms than the bottom set of context terms, i.e. the greater the context term weight, the better an indicator of termhood it is. Secondly, all context terms appear more frequently with terms than with non-terms (although in the case of the lowest ranked context terms, this is only marginal).

It is perhaps more revealing to compare these results with Frantzi's results

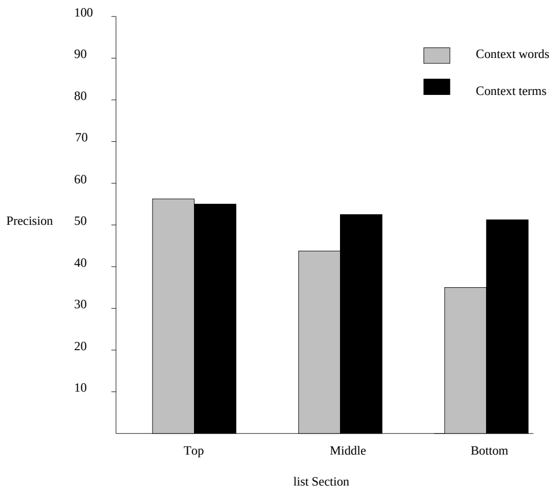

Figure 8.2: Comparison of context words and context terms

for context words. From Table 8.10 it is clear that there is much more of a discrepancy between the top and bottom sets of the context words than there is for the context terms. Whilst 56% of the top set of context words appear with terms, only 35% of the bottom set do so, a difference of 21%. For context terms, there is only a difference of 3% between the top and bottom sets. There are several contributory factors to this.

- There are fewer context terms than context words. In our corpus there are 1182 context terms, compared with 6948 context words.
- The set of context words and the set of context terms are not mutually exclusive, since context terms are simply a subset of context words.
- We are comparing just the context term weight with the context word weight in this experiment. We have not included the semantic weight, which adds to the discriminatory power of the context term.
- Since context terms are supposed to be better indicators than context words, it is to be expected that the bottom set of context terms will be better

Occurrence	No. of Context Terms	Percentage
more frequent with term	419	90%
equally frequent with either	28	6%
more frequent with non-term	17	4%

Table 8.12: Occurrences of context terms with terms and non-terms

indicators than the bottom set of context words. The fact that the bottom set of context terms appears with almost as high a proportion of terms as the top set of context words demonstrates the superiority of context terms over context words.

8.3.2 The distribution of context terms in the corpus

If we examine the behaviour of individual context terms, some interesting facts emerge. In the next experiment, we compare the frequency of occurrence of each context term with terms and with non-terms. This was achieved by taking each context term individually and calculating the number of terms and non-terms each occurred with.

For each context term, we then calculated the proportion of terms to non-terms it appeared with. The overall results depicted in Table 8.12 show that 90% of context terms appear more frequently with a term than with a non-term, and only 4% appear less frequently. The figure in the second column indicates not the total number of co-occurrences, but the number of context terms occurring in the proportion indicated by the first column (more, less or equally frequently with terms and non-terms).

Another property of context terms should be that if they appear more frequently with terms at the top of the list than terms at the bottom of the list (i.e. they occur more often with "better" terms), then they are better indicators. We therefore conducted another experiment to see how often context terms appeared

List section	No. of Co-occurrences	Proportion of Co-occurrences
top	31310	76%
middle	6019	15%
bottom	3927	9%

Table 8.13: Occurrences of context terms with terms from each part of the list

with terms from each part of the list. The C-value list was again divided into 3 equal sections and the number of occurrences of context terms with terms from each section of the list was calculated. The results depicted in Table 8.13 confirm overwhelmingly that this is the case, 76% of all occurrences being with terms from the top of the list.

8.4 Syntactic Weight

The syntactic weight aims to identify context boundary words which are particularly useful in identifying terms. We do this by allocating different values to boundary words according to their syntactic category.

8.4.1 Average syntactic weights for terms and non-terms

To evaluate the syntactic weight, we compared the average weights for terms and non-terms (as evaluated manually). We divided the list of candidate terms and their syntactic weights into two, according to whether they were terms or non-terms (according to the manual evaluation), and divided the total weight for all the candidate terms in each list by the total number of candidate terms in each list. Table 8.14 shows the total weight, the total number of candidate terms, and the average weight for each list. Although we have nothing with which we can compare this result, clearly the fact that the average weight for terms is more than twice that for non-terms demonstrates that the weight is a useful addition

	Total Weight	Total Number	Average Weight
Terms	10489.3	499	21.02
Non-Terms	8382.3	862	9.72

Table 8.14: Comparison of average syntactic weight for terms and non-terms

to our ranking since, on average, it increases the distinction between terms and non-terms.

Naturally, comparing the average weights for terms and non-terms only provides a rough guide to the usefulness of this weight, but what it does ensure is that the evaluation is not clouded by other factors such as the way in which the syntactic weight is incorporated into the SNC-Value. Our second experiment compares the total SNC-Value of the terms with and without the syntactic weight.

8.4.2 Evaluation of the addition of the syntactic weight to the SNC-Value

For this experiment, we calculate the impact that the addition of the syntactic weight has to the SNC-Value as a whole. We use the same method as for some of our previous experiments. We first recalculate the SNC-Value without the addition of the syntactic weight, and sort the list in order of descending ranking. We divide the list vertically into 3 equal sections, and calculate the percentage of terms and non-terms in each section of the list, as before. The results are compared with those obtained earlier from the SNC-Value which included the syntactic weight. Table 8.15 shows the precision and distribution values for the SNC-Value with the syntactic weight added, and Table 8.16 shows the values for the SNC-Value without the syntactic weights. We can see that in terms of precision, there is an increase of 3% in the top section of the list when the

List Section	No. of Valid Terms	Precision	Distribution
top	372	44%	45%
middle	234	27%	28%
bottom	222	26%	27%

Table 8.15: SNC-Value with the syntactic weight

List Section	No. of Valid Terms	Precision	Distribution
top	349	41%	42%
middle	244	29%	30%
bottom	233	27%	28%

Table 8.16: SNC-Value without the syntactic weight

syntactic weight is added, a decrease of 2% in the middle section, and a decrease of 1% in the bottom section. As verified by the distribution results, this means that the ranking has been improved by the addition of the syntactic weight.

If we break down the list into more sections, this becomes even clearer. Table 8.17 and Figure 8.3 show the results for precision when the list is split into 10 equal parts rather than 3. Again, we see that the range of values is higher for the list with the added syntactic weight, i.e. that the precision is higher at the top of the list and lower at the bottom of the list.

If we look more closely at the top of the list, we can again see an improvement in the precision figures. Table 8.18 and Figure 8.4 show the precision for each section of the top part of the list (which has been divided into 10 equal parts). From the graph we can clearly see that the shape of the unbroken line is much smoother than that of the dotted line. This means that the results for the SNC-Value with the syntactic weight are better in this top third of the list than for the SNC-Value without the syntactic weight.

List Section	+syn weight	-syn weight
1	64	62
2	33	35
3	35	28
4	35	37
5	30	27
6	22	28
7	26	25
8	29	29
9	27	26
10	23	27

Table 8.17: Precision of SNC-Value with and without syntactic weight added

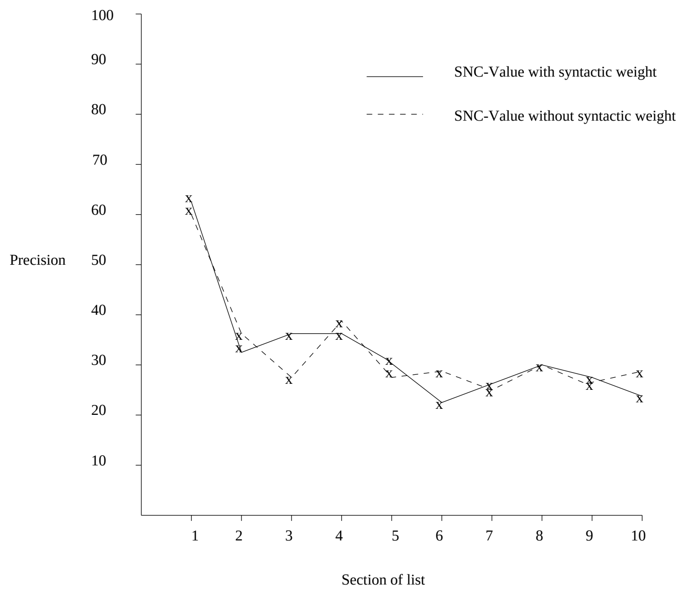

Figure 8.3: Precision of SNC-Value with and without syntactic weight

List Section	+ syn weight	- syn weight
1	84	80
2	76	80
3	68	88
4	64	64
5	72	40
6	48	68
7	52	48
8	68	48
9	52	64
10	56	52

Table 8.18: Precision of top 250 terms with and without the syntactic weight

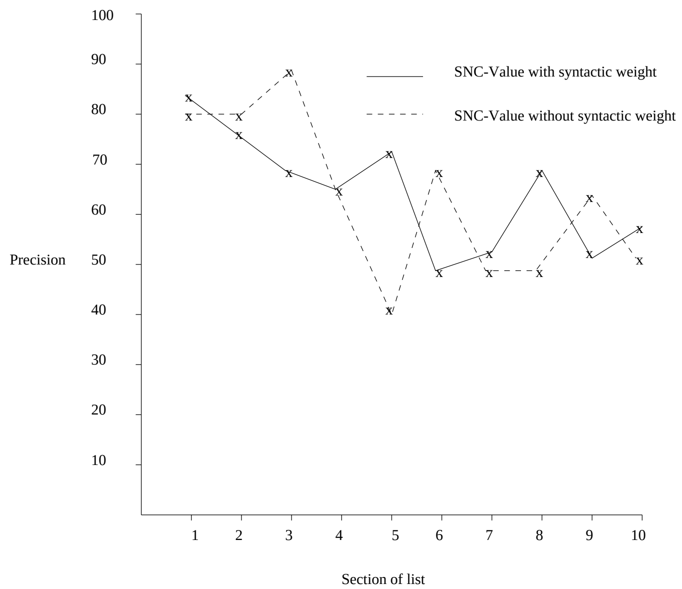

Figure 8.4: Precision of top 250 terms with and without the syntactic weight

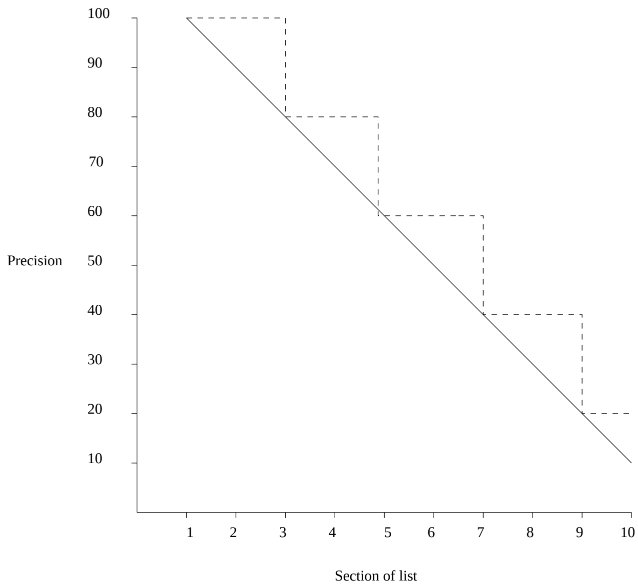

Figure 8.5: Examples of "perfect" results for precision

8.5 Overall Evaluation of the SNC-Value

Finally, we evaluate the SNC-Value as a whole, by comparing it with the NC-Value. We know already that the C-Value is improved by the addition of contextual information (the NC-Value), and expect that it will be improved still further by the addition of the deeper forms of linguistic information incorporated in the Importance Weight.

8.5.1 Comparing results: upward and downward trends

The purpose of the term ranking is to estimate how likely a candidate term is to be valid. So the higher the weight, and the further up the list the term is, the more likely it is to be valid. This means that if we divide the list into sections, the higher the section, the more valid terms it should contain, and therefore the higher the precision value.

When examining the results of the precision, there are two important factors

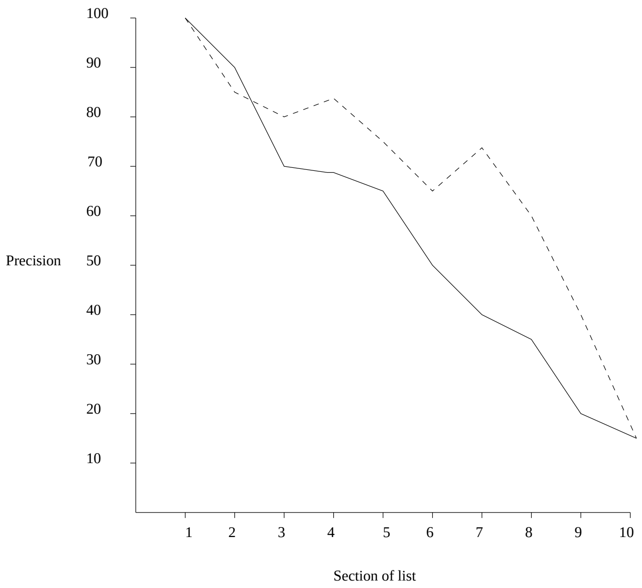

Figure 8.6: Examples of upward and downward trends

to note. Firstly, it is hard to define what the perfect result should look like. We might consider that, when plotted on a graph, the results should resemble a straight line at a 45 degree angle (the unbroken line in Figure 8.5). Or we might consider that they should resemble something more similar to a set of downward steps (the dotted line in Figure 8.5). There are also many other possibilities for perfect graphs, resembling something in between these two extremes.

Secondly, the trend should always be downwards rather than upwards. This means that as we progress down the list, the precision should never exceed the precision of the previous section. The unbroken line in Figure 8.6 depicts a downward trend, while the dotted line in the same figure depicts a mixture of upward and downward trends. Although both lines start and end at the same point, the straight line can be considered a better result than the dotted line, because the precision always decreases as we progress down the list.

We therefore compare the graphs by measuring their comparative upward trends, rather than by evaluating them against an ideal. This is calculated by dividing the total rise in precision percentage by the length of the graph. A graph with a lower upward trend will be better than a graph with a higher upward trend. The reason we do not divide the rise by the fall in percentage is that the result would then be affected by the angle of the line. A small increase in a sharply falling precision rate would be considered as better than a small increase in a gently falling precision rate. This, however, is not really a valid judgement, because it considers the gravity of the error in direct proportion to the rate of decrease in precision. This assumes that a straight line is the ideal solution, which, as we saw in Figure 8.5, is not necessarily the case.

The two graphs in Figure 8.6 would thus be evaluated as follows. The unbroken line would have an upward trend of 0, since at no point does the precision rise. The dotted line would have an upward trend of 1.5, since the total rise in precision is 15 points (5 points between intervals 3 and 4, and 10 points between intervals 6 and 7. The length of the graph is 10 intervals, so we calculate the upward trend as 15 / 10 = 1.5.

8.5.2 Precision of the SNC and NC-Values

We first compare the precision rates for the SNC and NC-Values (Table 8.19, by dividing the ranked lists into 10 equal sections. Each section contains 250 terms, marked as valid or invalid by the manual experts. We see that in the top section, the precision is higher for the SNC-Value, and in the bottom section, it is lower. This indicates that the precision span is greater for the SNC-Value, and therefore that the ranking is improved. Looking at Figure 8.7, we can see that the SNC-Value graph is smoother than that of the NC-Value. If we compare the upward trends of the two graphs, we find that the trend for the SNC-Value is 0.9, whereas the trend for the NC-Value is 2.7. This again shows that the SNC-Value

	SNC-Value		NC-Value	
Section	Valid	Precision	Valid	Precision
1	163	64%	160	62%
2	84	33%	98	38%
3	89	35%	69	27%
4	89	35%	78	30%
5	76	30%	87	34%
6	57	22%	78	30%
7	66	26%	92	36%
8	75	29%	100	39%
9	70	27%	42	16%
10	59	23%	68	27%

Table 8.19: Precision of SNC-Value and NC-Value

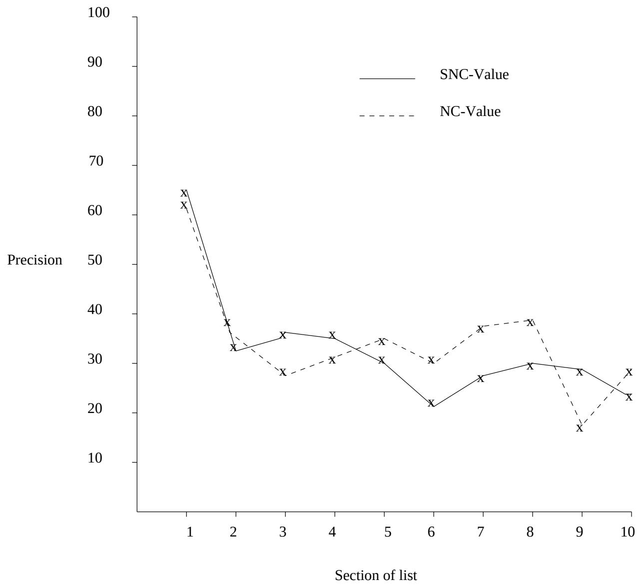

Figure 8.7: Precision of SNC-Value and NC-Value

Section	SNC-Value	NC-Value
1	64%	62%
2	48%	50%
3	44%	42%
4	42%	40%
5	39%	38%
6	36%	37%
7	35%	37%
8	34%	37%
9	33%	35%
10	32%	34%

Table 8.20: Accumulative precision of SNC-Value and NC-Value

ranking is better than the NC-Value ranking, since it is more consistent.

8.5.3 Accumulative precision of the SNC and NC-Values

We next turn our attention to the accumulative precision rates. Instead of measuring the precision from one interval to the following interval, we measure the accumulative precision from the top of the list down to each interval. This gives us another view of the consistency of the ranking, and also gives us the precision value for any sized section of the list (e.g. for the top half of the list, or for the top quarter, or the top tenth). Table 8.20 shows the results for each interval, and Figure 8.8 depicts the same information graphically. Again, the span of values for the SNC-Value is greater than for the NC-Value, which indicates a better ranking. As for the upward trend of each graph, we find that both have a trend of 0. This means that the accumulative precision always decreases from one interval to the next as we progress down the list. So in this case, there is no difference between the two. However, looking at the shape of the two graphs, it is apparent that the SNC-Value result is slightly better than the NC-Value result.

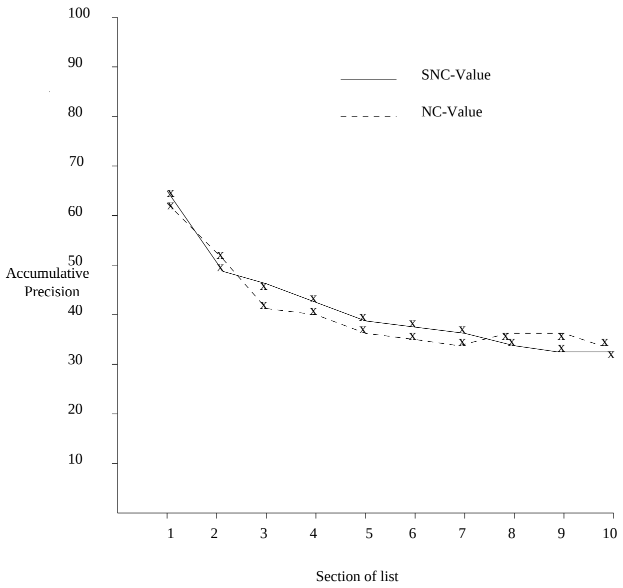

Figure 8.8: Accumulative precision of SNC-Value and NC-Value

8.5.4 Detailed precision at the top of the list

We are most interested in the ranking at the top end of the list for two reasons. Firstly, this contains the candidate terms most likely to be valid. Secondly, from the previous results (Table 8.19), it is clear that the precision rate drops very steeply below the first section of the list (the top 250 terms). So we now take a more detailed look at the precision of these terms. We perform the same procedure as before on this section of the list, by dividing it into 10 equal segments and measuring the precision of each. Each segment therefore now contains 25 terms.

From Table 8.21, we can see that the change in precision (as we progress down the list) is much smoother for the SNC-Value than for the NC-Value, although the overall precision rates are almost identical. Again, this indicates a better distribution of the valid terms, and therefore an improvement in the ranking. Looking at the graphs (Figure 8.9), we see two rather different results. The SNC-Value graph is more horizontal, and has less severe peaks and troughs than the NC-Value graph. This is significant for two reasons. Firstly, the fact that it is

	SNC-Value		NC-Value	
Section	Valid	Precision	Valid	Precision
1	21	84%	19	76%
2	19	76%	23	92%
3	17	68%	21	84%
4	16	64%	13	52%
5	18	72%	13	52%
6	12	48%	19	76%
7	13	52%	18	72%
8	17	68%	14	56%
9	13	52%	10	40%
10	14	56%	8	32%

Table 8.21: Precision of SNC-Value and NC-Value for top 250 terms

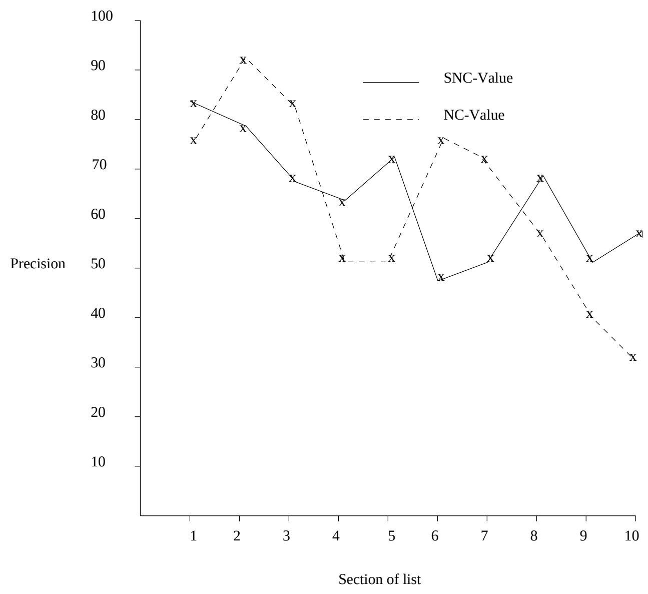

Figure 8.9: Precision of SNC-Value and NC-Value for top 250 terms

Section	SNC-Value	NC-Value
1	84%	76%
2	80%	84%
3	76%	84%
4	73%	76%
5	73%	71%
6	69%	72%
7	66%	72%
8	66%	70%
9	64%	67%
10	64%	63%

Table 8.22: Accumulative precision of SNC-Value and NC-Value for top 250 terms

flatter means that the precision at the bottom of this section is higher. This is good because, since we are only measuring terms in the top of the list, we would expect a high rate of precision, even at the bottom. Secondly, the fact that the graph is smoother is good, because it again means that the ranking is consistent, with precision gradually falling down the list. Measuring the upward trend, we find that the SNC-Value has a trend of 2.2, and the NC-Value a trend of 3.0, which again indicates the superiority of the SNC-Value.

8.5.5 Accumulative precision of the top 250 terms

As before, we also examine the accumulative precision of the top 250 terms. From Table 8.22 and Figure 8.10, we see that the SNC-Value presents a better set of results than the NC-Value, since the results for the top section are better, and the graph is much smoother. This is also evident from measuring the upward trend, which is 0 for the SNC-Value and 0.9 for the NC-Value. This means that the accumulative precision for the SNC-Value follows a smooth monotonic line downwards, and never increases at any point, whereas the accumulative precision for the NC-Value shows signs of upward movement.

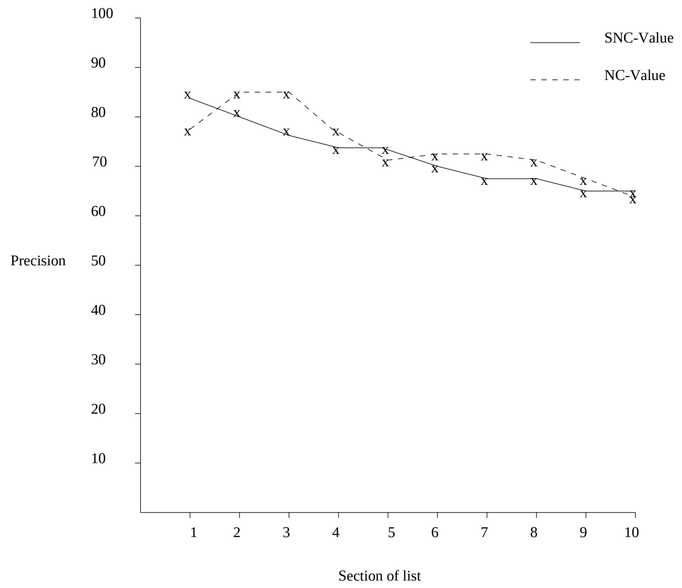

Figure 8.10: Accumulative precision of SNC-Value and NC-Value for top 250 terms

8.5.6 Distribution

Looking at the distribution figures for each of the lists gives us a better idea of how the valid terms are dispersed throughout the lists. The aim again is that a high percentage of valid terms should be found in the top part of the list, and a low percentage in the bottom part, and that the figure should gradually diminish as we progress down the list.

From Table 8.23, we can see that the distribution figures are better for the SNC-Value than for the NC-Value. A higher proportion of valid terms appears at the top of the list (20% as opposed to 18%), a lower proportion appears at the bottom of the list (7% as opposed to 8%), and that there are fewer discrepancies in the middle of the list.

Section	Valid	SNC-Value	Valid	NC-Value
1	163	20%	160	18%
2	84	10%	98	11%
3	89	11%	69	8%
4	89	11%	78	9%
5	76	9%	87	10%
6	57	7%	78	9%
7	66	8%	92	11%
8	75	9%	100	11%
9	70	8%	42	5%
10	59	7%	68	8%

Table 8.23: Distribution of SNC-Value and NC-Value

Section	Valid	SNC-Value	Valid	NC-Value
1	21	13%	19	12%
2	19	12%	23	15%
3	17	11%	21	14%
4	16	10%	13	8%
5	18	11%	13	8%
6	12	8%	19	12%
7	13	8%	18	11%
8	17	11%	14	9%
9	13	8%	10	6%
10	14	9%	8	5%

Table 8.24: Distribution of SNC-Value and NC-Value for top 250 terms

8.5.7 Distribution of the top 250 terms

For the top 250 terms, the distribution is also better for the SNC-Value than for the NC-Value, although it is not quite as immediately evident. The highest percentage of valid terms is at the top of the list for the SNC-Value, but not for the NC-Value. The terms are also more regularly dispersed for the SNC-Value.

8.6 Conclusions

In this chapter, we have evaluated the results of the SNC-Value compared to the NC-Value, in order to show the improvement it produces. Clearly, although the differences are only a matter of several percentage points, it does show an improvement in terms of the precision of the ranking, and of the distribution of valid terms. It does not show any improvement in recall, since it simply aims to reorder the list of candidate terms produced by the NC-Value, but this could be achieved by altering the thresholds at which candidate terms are accepted or rejected. This is not of vital importance, however, because we do not expect terms at the bottom of the list to be valid. If we consider the top third of the list as being valid, as with the C-Value, then we have a significantly better chance of getting valid terms from this part of the list after it has been ranked by the SNC-Value.

We have also evaluated individually each of the three main components of the Importance Weight – the terminological weight, the syntactic weight, and the semantic weight – in order to see whether the theoretical premises on which they are based are valid. From the results, it is apparent that each of these weights is indeed justified and offers some contribution towards the identification of important contextual information which is so important for term extraction.

It should be pointed out that the size of the corpus, and therefore the number of terms, is very small. This tends to affect statistical calculations and evaluation procedures negatively. The fact that we do not see very smooth graphs for the precision is largely because of this problem. Any errors are magnified when the number of terms being considered is small. The improvements seen when new weights are added may also be small, because only a small number of terms may be affected by the changes.

In the final chapter, we summarise the main points of this work, draw some conclusions about its strengths and weaknesses, and highlight the future directions which it could take.

Chapter 9

Conclusions

In this chapter, we provide a brief summary of our work, some concluding remarks and suggest ideas for future developments. We have described a method for multi-word term extraction within the medical field, based on the acquisition of contextual information from a variety of knowledge sources. In particular, it focuses on using deeper forms of linguistic information than current statistical or hybrid methods use. In addition to term extraction, disambiguation of terms and their context is also performed.

9.1 Summary

The SNC-Value measure builds on the NC-Value measure, a largely statistical method of extracting and ranking terms, by investigating the context more closely, and making distinctions about which parts of the context are most relevant to terms. It consists of three main parts:

- the *terminological* weight, which considers the importance of words in the context which are themselves terms;

- the *syntactic* weight, which considers the importance of the syntactic category of boundary context words;
- the *semantic* weight, which considers the importance of the semantic relationship between candidate terms and their context.

The contextual information acquired in this process can also be used for a number of other related tasks, such as disambiguation and clustering. At present, the semantic information is largely acquired from a pre-existing domain-specific thesaurus, but there are possibilities for either creating such a thesaurus automatically, or enhancing an existing one, using the contextual information we acquire.

Our method was applied to a corpus of eye pathology records to produce a ranked list of extracted terms. These terms had been previously evaluated by two domain experts, to determine their validity. We compared the results in several different ways. First, we evaluated each of the individual context weights according to its expected theoretical performance. Second, we compared the precision of the results of the SNC-Value method with the list of terms contained in the UMLS Metathesaurus. Third, we compared the precision and distribution of the list with the NC-Value method, to determine if and how the SNC-Value method improved on the results.

Analysis of the individual context weights showed that the methods used for extracting relevant contextual information were productive, and that they were based on sound theoretical assumptions. This has important consequences for future research. For example, the idea of using barrier methods for extracting terms is not new, but the additional benefit of distinguishing between different syntactic categories adds a new dimension to an old idea. Similarly, the idea of the terminological weight enables the acquisition of contextual information to be more selective, because advantage can be taken of the fact that different parts of the context may vary in their relevance to terminological processes.

The main part of the contextual information, the similarity measure, is particularly useful, since it not only provides information for term recognition, but also can be used for term and context disambiguation. The former, term disambiguation, has been barely touched on in the past, since methods for term extraction have disregarded the relationship between term and context, and have sought to extract terms largely according to their surface form rather than their meaning. Similarly, where contextual information has been used as a knowledge source for terminology, little account has been taken of its meaning and of any ambiguities that may exist. We have not, however, evaluated the importance of disambiguation, but in larger corpora we would expect it to have much greater significance.

Although the improvement in precision and distribution over the results of the NC-Value was minor in terms of percentage points, this is only to be expected with such a small corpus. In almost every aspect of analysis, however, benefits are gained. In particular, the overall distribution of the list was improved, such that the rate of precision decreases more smoothly from the top to the bottom of the list of ranked terms. Also, when the top section of the list is analysed separately, improved precision and distribution is found. This means that, in addition to improving knowledge about the theoretical nature of terms, incorporating deeper forms of linguistic information has an impact on the overall performance of term extraction, compared with existing statistical techniques.

9.2 Further Work

As we have already explained, the aim of this thesis is not to provide a complete solution to the problem of term extraction, but to investigate methods in which existing systems could be improved. Although our evaluation has highlighted the benefits of our approach, there are various ways in which both the evaluation techniques and the system itself could be improved:

1. The SNC-Value measure has only been tested on one small, very specific corpus. Ideally, it should be tested on a larger set of data and on a variety of corpora within the medical domain. Unfortunately, we did not have the means available to carry out this kind of testing. We would predict that on a larger corpus, the improvement over statistical methods such as the NC-Value would be magnified. On a corpus with different content, such as asthma records as opposed to eye pathology records, we would not expect to see much change, although it might highlight areas of the UMLS which are more or less complete.

2. Due to using a domain-specific thesaurus, testing is necessarily limited to corpora within the medical field. We see no a priori reason why the approach should not be valid in other domains, given a suitable thesaurus. Indeed, it would be extremely interesting to make such a comparison, both as a measure of the usefulness of the thesaurus, and as a measure of the domain portability of the method.

3. Clearly there are many other methods of evaluation possible aside from the standard information retrieval techniques which we have used. Since there are no real standards for evaluation, we have chosen the methods which seem to be the most appropriate for initial evaluation. Evaluation of terminological systems is an area that has been rather overlooked in the past, and is an area which deserves much more recognition than it currently receives. Better standards and guidelines for evaluation would almost certainly have an impact on the development of new methods in terminology, such as the incorporation of detailed linguistic information.

4. One problem with relying on a specialised thesaurus is that the results are limited by the reliability of the thesaurus. As we discussed in Chapter 8, some of the problems associated with a pre-existing thesaurus could be tackled if we were to either create a thesaurus automatically, or tune an existing thesaurus to the domain. The latter approach is perhaps more feasible, at least in the short term, and the method described in Chapter 7 for clustering contexts could be a valuable starting point, along with approaches such as that of Ushioda (Ushioda, 1996) and Mima et al. (Mima et al., 1999). This idea is discussed elsewhere (Maynard and Ananiadou, 1999c). Furthermore, an incremental approach is possible, whereby the similarity measure described in Chapter 5 is combined with statistical information to tune an existing ontology. For example, the work on disambiguation of terms is particularly suitable for application to ontology tuning.

5. Another problem specifically with the UMLS is that it is designed as an *information* resource rather than a *linguistic* one, in contrast to dictionaries and thesauri such as LDOCE, WordNet and Roget's Thesaurus. Given that we cannot realistically use a general language resource such as one of these, for reasons we have discussed earlier, we are limited to using an information resource for a linguistic application. This is another way in which tuning would be of use, since it might be possible to integrate the UMLS with a linguistic resource. Current and future developments of the UMLS are indeed pursuing this track by incorporating more detailed linguistic information, such as morphology. The problem of lexical variation is also one which should be tackled. We have already discussed its influence on our results and on terminological applications in general. Although relationships with variants are established in the UMLS, this type of information is not easy to manipulate in the same way that it is in linguistically-oriented resources.

This, too, would be a very useful topic for further investigation.

6. Finally, the work on clustering is only embryonic at this stage, and, as we have mentioned, could be integrated further with the contextual information extracted in the SNC-Value Measure. Although similar methods for extracting information and creating templates are not uncommon, it is rare that these make use of much semantic information. This is largely because of the practical difficulties associated with such integration rather than due to any fundamental theoretical issues. It is hoped that the work carried out in this thesis will therefore suggest some advances in this direction.

Appendix A

Stop List for Term Extraction

The following is the list of words used in the stop list, or lexical filter, applied to candidate strings from the corpus.

ABANDONED

ABSOLUTE

ACCIDENT

ACTIVITY

ACTUAL

ADJACENT

AMOUNT

APPROXIMATE

APPROXIMATELY

AREA

ARRANGEMENT

ARTEFACT

ASPECT

ASSOCIATED

BODIES

BODY

BROAD

BROWNISH

CHANGES

CHOCOLATE

CHRONIC

CLEAR

CLINICAL

COARSE

COMPLETE

CONSIDERABLE

CONTENT

CONTENTS

CREAMY

CRUSH

CUT

DARK

DEAL

DEEP

DEEPER

DEFINITE

DEGREE

DELICATE

DENSE

DETACHED

DIAGNOSIS

DISLOCATED

DISORGANISED

EARLY

ELEMENTS

END

ENTIRE

ENTRY

EVIDENCE

EXIT

EXTEND

EXTENSIVE

EXTRA

EXTREME

FALSE

FAT

FATTY

FEATURE

FEATURES

FEW

FINAL

FINE

FIRM

FIRST

FLAT

FLESHY

FOAMY

FOREIGN

FRAGMENT

FRESH

FULL

FURTHER

GAP

GENERAL

GIANT

GOOD

GRADE

GREAT

GREATEST

GREENISH

GREY

GREYISH

GROSS

GROWTH

GUN

HALF

HEAVY

HIGH

HISTORY

INFERIORLY

INFLAMED

INJURY

INNER

INTACT

INTENSE

IRREGULAR

IRRITATED

JUST

LARGE

LARGISH

LAYER

LEAST

LEFT

LENGTH

LEVEL

LIGHT

LIMITS

LITTLE

LOCAL

LONG

LOOKING

LOOSE

LOSS

LOW

LOWER

MAIN

MAJOR

MANY

MASS

MASSIVE

MATERIAL

MAXIMUM

MILD

MIXED

MONTH

MORE

MOST

MUCH

NARROW

NEW

NORMAL

NUMBER

NUMEROUS

OBVIOUS

OCCASIONAL

OIL

OLD

OPEN

OPPOSITE

ORDERLY

ORGANISING

ORGANIZING

ORIGIN

OTHER

OVAL

OVERLYING

PAINFUL

PALE

PAPER

PART

PARTICULAR

PATCHY

PATHOLOGICAL

PERSISTENT

PIECE

PORTION

POSITIVE

POSSIBLE

POWER

PRELIMINARY

PRESENT

PREVIOUS

PROBABLE

PROGRESSIVE

RANDOM

RECURRENT

RECENT

REPORT

RIGHT

ROAD

SECOND

SEVERAL

SEVERE

SHALLOW

SHORT

SHOW

SHOWS

SIMILAR

SIMPLE

SIZE

SLIGHT

SLOW

SMALL

SMOOTH

SOFT

SOLID

SPARSE

SPECIFIC

SPINDLY

STAGE

SUBSTANTIAL

SUPERFICIAL

SUPERIORLY

SUPPLEMENTARY

SURFACE

SWEAT

SWOLLEN

TEMPORAL

THICK

THIN

THIRD

TIERED

TINY

TOTAL

TRAFFIC

TRAUMATIC

TRAUMATISED

TYPE

TYPICAL

UNCERTAIN

UNEVEN

UNIFORM

UNUSUAL

UPPER

VARIABLE

VARIOUS

VERY

WALLED

WATERY

WHITE

WHOLE

WIDE

WIDESPREAD

YEAR

Appendix B

UMLS Semantic Hierarchy

A	Entity
A1	Physical Object
A11	Organism
A111	Plant
A1111	Alga
A112	Fungus
A113	Virus
A114	Rickettsia or Chlamydia
A115	Bacterium
A116	Animal
A1161	Invertebrate
A1162	Vertebrate
A11621	Amphibian
A11622	Bird
A11623	Fish
A11624	Reptile
A11625	Mammal
A116251	Human

A12	Anatomical Structure
A121	Embryonic Structure
A122	Anatomical Abnormality
A1221	Congenital Abnormality
A1222	Acquired Abnormality
A123	Fully Formed Anatomical Structure
A1231	Body Part, Organ or Organ Component
A1232	Tissue
A1233	Cell
A1234	Cell Component
A1235	Gene or Genome
A13	Manufactured Object
A131	Medical Device
A132	Research Device
A14	Substance
A141	Chemical
A1411	Chemical Viewed Functionally
A14111	Pharmacologic Substance
A14112	Biomedical or Dental Material
A14113	Biologically Active Substance
A141131	Neuroreactive Substance or Biogenic Amine
A141132	Hormone
A141133	Enzyme
A141134	Vitamin
A141135	Prostaglandin
A141136	Immunologic Factor
A141137	Receptor

A14114	Indicator or Reagent
A14115	Hazardous or Poisonous Substance
A1412	Chemical Viewed Structurally
A14121	Organic Chemical
A141211	Steroid
A141212	Eicosanoid
A141213	Lactam
A141214	Alkaloid
A141215	Nucleic Acid, Nucleoside, or Nucleotide
A141216	Organophosphorus Compound
A141217	Amino Acid, Peptide or Protein
A141218	Carbohydrate
A141219	Lipid
A14122	Inorganic Chemical
A141221	Element or Ion
A141222	Isotope
A141223	Inorganic Compound
A142	Body Substance
A143	Food
A2	Conceptual Entity
A21	Idea or Concept
A211	Temporal Concept
A212	Qualitative Concept
A213	Quantitative Concept
A214	Functional Concept
A2141	Body System
A215	Spatial Concept

0A2151	Body Space or Junction
A2152	Body Location or Region
A2153	Molecular Sequence
A21531	Nucleotide Sequence
A21532	Amino Acid Sequence
A21533	Carbohydrate Sequence
A2154	Geographic Area
A22	Finding
A221	Laboratory or Test Result
A222	Sign or Symptom
A23	Organism Attribute
A24	Intellectual Product
A241	Classification
A242	Regulation or Law
A25	Language
A26	Occupation or Discipline
A261	Biomedical Occupation or Discipline
A27	Organisation
A271	Healthcare Organisation
A272	Professional Society
A273	Self-Help or Relief Organisation
A28	Group Attribute
A29	Group
A291	Professional or Occupational Group
A292	Population Group
A293	Family group
A294	Age Group

A295	Patient or Disabled Group
B	Event
B1	Activity
B11	Behaviour
B111	Social Behaviour
B112	Individual Behaviour
B12	Daily or Recreational Activity
B13	Occupational Activity
B131	Health Care Activity
B1311	Laboratory Procedure
B1312	Diagnostic Procedure
B1313	Therapeutic or Preventive Procedure
B132	Research Activity
B1321	Molecular Biology Research Technique
B133	Governmental or Regulatory Activity
B134	Educational Activity
B14	Machine Activity
B2	Phenomenon or Process
B21	Human-caused Phenomenon or Process
B211	Environmental Effect of Humans
B22	Natural Phenomenon or Process
B221	Biologic Function
B2211	Physiologic Function
B2212	Organism Function
B22121	Mental Process
B2213	Organ or Tissue Function
B2214	Cell Function

B2215	Molecular Function
B22151	Genetic Function
B222	Pathologic Function
B2221	Disease or Syndrome
B22211	Mental or Behavioural Function
B22212	Neoplastic Process
B2222	Cell or Molecular Dysfunction
B2223	Experimental Model of Disease
B223	Injury or Poisoning

Appendix C

Frequency List for Verbs

The following is the frequency list for the verbs appearing in syntactic frames, as discussed in Chapter 7.

223	`show`
123	`contain`
90	`cut`
70	`infiltrate`
23	`bear`
18	`have`
17	`form`
12	`stage`
11	`diffuse`
11	`invade`
10	`overlie`
9	`undergo`
8	`traverse`
7	`suggest`
7	`see`

7	measure
7	cover
6	involve
6	extend
6	enter
6	cross
6	affect
5	obscure
5	give
4	retract
4	perforate
4	occupy
4	comprise
3	take
3	remove
3	reach
3	process
3	lack
3	funnel
3	complement
3	arrange
2	elevate
2	disperse
2	underlie
2	strip
2	spur
2	spindle

2	retain
2	present
2	incorporate
2	encase
1	stretch
1	slide
1	shed
1	shatter
1	receive
1	proliferate
1	prolapse
1	plug
1	organize
1	lift
1	heap
1	fold
1	exclude
1	envelop
1	diagnose
1	break
1	breach
1	weave
1	surround
1	separate
1	run
1	resemble
1	represent

1	replace
1	protrude
1	project
1	penetrate
1	pare
1	pack
1	make
1	lobulate
1	include
1	glisten
1	fill
1	divide
1	detect
1	constrict
1	confirm
1	complicate
1	clear
1	change
1	attach
1	accompany

Appendix D

Part-of-Speech Tags

The following is a list of the Penn Treebank Part-of-Speech tags used by the Supertagger.

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural

PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Appendix E

List of SuperTags

The following is a list of the supertags created at the University of Pennsylvania, used for the context clustering in Chapter 7.

A_ARB	Adverb as argument
A_AXA	Adjective as argument
A_AXAs	Adjective with sentence complement
A_CONJ	Conjunction as argument
A_D	Determiner as argument (in coordination)
A_N	Noun as argument (in coordination)
A_NXG	Genitive as argument
A_NXN	Noun phrase as argument
A_NXNs	Noun phrase with sentence complement
A_P	Preposition as argument (in coordination)
A_PL	Particles
A_PU	Punctuation
A_PXARBPnx	Complex preposition argument
A_PXP	Exhaustive Preposition argument
A_PXPNaPnx	Complex preposition argument

A_PXPPnx	Complex preposition argument
A_PXPnx	Preposition argument
A_by	by in passives
B_APnxs	Sentence initial complex preposition modifier
B_ARBCONJs	Sentence initial complex conjunction modifier
B_ARBPnxs	Sentence initial complex preposition modifier
B_ARBPs	Sentence initial complex preposition modifier
B_ARBa	Adverb modifying an adjective
B_ARBarb	Adverb modifying an adverb
B_ARBarbs	Sentence initial complex adverb modifier
B_ARBd	Adverb modifying a determiner
B_ARBn	Adverb modifying a noun
B_ARBnx	Adverb modifying a noun phrase
B_ARBpx	Adverb modifying a preposition phrase
B_ARBs	Adverb modifying a sentence
B_ARBvx	Adverb modifying a verb phrase
B_An	Adjective modifying a noun
B_COMPs	Complementizer
B_CONJACONJd	Complex determiner modifier
B_CONJDCONJd	Complex determiner modifier
B_CONJPnxs	Sentence initial complex preposition modifier
B_CONJarbCONJs	Sentence initial complex conjunction
B_CONJnx1conjnx2	Complex noun phrase conjunction
B_CONJs	Sentence initial conjunction
B_CONJs1conjs2	Complex sentence conjunction
B_CONJvx1conjvx2	Complex verb phrase conjunction
B_Dnx	Determiner

B_NEGarb	Negation on adverb
B_NEGnx	Negation on noun phrase
B_NEGpx	Negation on preposition phrase
B_NEGvx	Negation on verb phrase
B_Nn	Noun modifier
B_Ns	Temporal noun modifying a sentence
B_Nvx	Temporal noun modifying a verb phrase
B_PARBPnxs	Sentence initial complex preposition modifier
B_PARBarb	Complex preposition modifying an adverb
B_PARBd	Complex preposition modifying a determiner
B_PARBnx	Complex preposition modifying a noun phrase
B_PARBpx	Complex preposition modifying a preposition
B_PARBs	Complex preposition modifying a sentence
B_PARBvx	Complex preposition modifying a verb phrase
B_PDNPnxs	Sentence initial complex preposition modifier
B_PNPnxs	Sentence initial complex preposition modifier
B_PNaPnxs	Sentence initial complex preposition modifier
B_PPnxs	Sentence initial complex preposition modifier
B_PUapu	Punctuation around adjective
B_PUarbpu	Punctuation around adverb
B_PUnpu	Punctuation around noun
B_PUnxpu	Punctuation around noun phrase
B_PUpu	Punctuation on another punctuation
B_PUpx	Punctuation before preposition phrase
B_PUpxpu	Punctuation around preposition phrase
B_PUpxpunx	Punctuation around preposition phrase
B_PUs	Punctuation before a sentence

B_PUspu	Punctuation around a sentence
B_PUvpu	Punctuation around a verb
B_PUvxpu	Punctuation around a verb phrase
B_Pnxs	Sentence initial preposition modifier
B_UCONJUCONJs	Complex Sentence initial conjunction
B_Vn	Participle verb, prenominal modifier
B_Vnxpus	Parenthetical sentence complement verb
B_Vnxpuvx	Parenthetical sentence complement verb
B_Vnxs	Parenthetical sentence complement verb
B_Vs	Auxiliary inversion
B_Vvx	Auxiliary verb
B_a1CONJa2	Adjective coordination
B_aARB	Adverb modifying an adjective
B_arb1CONJarb2	Adverb coordination
B_arbARB	Adverb modifying an adverb
B_ax1CONJax2	Adjective phrase coordination
B_conjAd1conjAd2	Multi_anchor adverb coordination
B_d1CONJd2	Determiner coordination
B_n1CONJn2	Noun coordination
B_nA	Postnominal adjective modifier
B_nARB	Postnominal adverb modifier
B_nPUnx	Appositive for postal addresses without final punctuation
B_nPUnxpu	Appositives for postal addresses
B_nx1CONJARBCONJnx2	Multi_anchor noun phrase conjunction
B_nx1CONJARBnx2	Multi_anchor noun phrase conjunction
B_nx1CONJnx2	Noun phrase conjunction

B_nx1conjCONJnx2	Multi_anchor noun phrase conjunction
B_nxAPnx	Noun phrase modifying complex preposition
B_nxARB	Noun phrase modifying adverb
B_nxARBPnx	Noun phrase modifying complex preposition
B_nxGnx	Possessive 's'
B_nxN	Noun phrase modifying time noun phrase
B_nxP	Noun phrase modifying exhaustive preposition
B_nxPDNPnx	Noun phrase modifying complex preposition
B_nxPNPnx	Noun phrase modifying complex preposition
B_nxPNaPnx	Noun phrase modifying complex preposition
B_nxPPnx	Noun phrase modifying complex preposition
B_nxPUnx	Appositives without final punctuation
B_nxPUnxpu	Appositives
B_nxPUpx	Noun phrase modifying preposition
B_nxPnx	Noun phrase modifying preposition
B_p1CONJp2	Preposition coordination
B_px1CONJARBCONJpx2	Multi_anchor preposition phrase coordination
B_px1CONJpx2	Preposition phrase coordination
B_pxARB	Preposition phrase modifying adverb
B_s1CONJs2	Sentence coordination
B_sARB	Sentence modifying adverb
B_sPU	Sentence final punctuation
B_sPUnx	Sentence modifying noun phrase
B_sPUs	Punctuation connecting sentences
B_sPnx	Sentence modifying noun phrase
B_v1CONJv2	Verb coordination
B_vnxVpu	Parenthetical sentence complement verb

B_vx1CONJARBCONJvx2	Multi_anchor verb phrase coordination
B_vx1CONJvx2	Verb phrase coordination
B_vx1conjCONJvx2	Multi_anchor verb phrase coordination
B_vxAPnx	Verb phrase modifying complex preposition
B_vxARB	Post verbal adverbial modifier
B_vxARBPnx	Verb phrase modifying complex preposition
B_vxCONJPnx	Verb phrase modifying complex preposition
B_vxN	Verb phrase modifying time noun phrase
B_vxNPnx	Verb phrase modifying complex preposition
B_vxP	Verb phrase modifying exhaustive preposition
B_vxPDNPnx	Verb phrase modifying complex preposition
B_vxPNPnx	Verb phrase modifying complex preposition
B_vxPNaPnx	Verb phrase modifying complex preposition
B_vxPPnx	Verb phrase modifying complex preposition
B_vxPnx	Verb phrase modifying preposition
B_vxPs	Verb phrase modifying preposition with sentence complement
A_EW1nx1V	Wh_question on subject of ergative verb
A_Enx1V	Active ergative verb
A_Gnx0Ax1	Gerundive adjectival predicate
A_Gnx0V	Gerundive intransitive
A_Gnx0Varb1	Gerundive verb with adverb complement
A_Gnx0Vax1	Gerundive verb with adjective complement
A_Gnx0Vnx1	Gerundive transitive verb
A_Gnx0Vnx1nx2	Gerundive ditransitive verb
A_Gnx0Vnx1pl	Gerundive transitive particle verb (shifted)

A_Gnx0Vnx1s2	Gerundive verb with noun phase and sentence complement
A_Gnx0Vplnx1	Gerundive transitive particle verb (unshifted)
A_Gnx0Vpnx1	Gerundive verb with preposition complement
A_Gnx0Vs1	Gerundive verb with sentence complement
A_Inx0V	Imperative intransitive verb
A_Inx0Vax1	Imperative adjectival predicate
A_Inx0Vnx1	Imperative transitive
A_Inx0Vnx1pnx2	Imperative verb with noun phrase and preposition complement
A_Inx0Vnx1s2	Imperative verb with noun phrase and sentence complement
A_Inx0Vpl	Imperative intransitive verb particle
A_Inx0Vplnx1	Imperative transitive verb particle (unshifted)
A_W0nx0Vnx1	Wh_question on subject of transitive verb
A_W1nx0N1	Wh_question on predicate of a nominal predicate
A_nx0ARBPnx1	Complex preposition predicate
A_nx0Ax1	Adjective predicate
A_nx0Ax1pnx1	Adjective predicate with preposition complement
A_nx0Ax1s2	Adjective predicate with sentence complement
A_nx0BEnx1	Equative
A_nx0CONJPnx1	Complex preposition predicate
A_nx0N1	Nominal predicate
A_nx0N1s1	Nominal predicate with sentence complement
A_nx0PDNPnx1	Complex preposition predicate
A_nx0PNPnx1	Complex preposition predicate
A_nx0PNaPnx1	Complex preposition predicate

A_nx0PPnx1	Complex preposition predicate
A_nx0Pnx1	Preposition predicate
A_nx0Px1	Exhaustive Preposition predicate
A_nx0Px1s2	Preposition predicate with sentence complement
A_nx0V	Intransitive verb
A_nx0Varb1	Verb with adverb complement
A_nx0Vax1	Verb with adjective complement
A_nx0Vnx1	Transitive verb
A_nx0Vnx1Pnx2	Dative shifted double object verb
A_nx0Vnx1nx2	Ditransitive verb
A_nx0Vnx1pl	Transitive verb particle (shifted)
A_nx0Vnx1pnx2	Verb with noun phrase and preposition phrase complement
A_nx0Vnx2nx1	Ditransitive verb
A_nx0Vpl	Intransitive verb particle
A_nx0Vplnx1	Transitive verb particle (unshifted)
A_nx0Vplpnx1	Verb particle with preposition complement
A_nx0Vpnx1	Verb with preposition complement
B nx0Vs1	Verb with sentence complement
A_nx0lVN1	Intransitive light verb
A_nx0lVN1Pnx2	Light verb with preposition complement
A_nx1V	Transitive passive verb with no by_phrase
A_nx1Vbynx0	Transitive passive verb with by_phrase
A_nx1Vnx2	Ditransitive passive verb with no by_phrase
A_nx1Vpl	Transitive passive verb particle with no by_phrase

A_nx1Vplbynx0	Transitive passive verb particle with by_phrase
A_nx1Vpnx2	Dative passive verb with no by_phrase
A_pW1nx0Pnx1	Wh_question on object of preposition predicate
A_s0N1	Sentence subject nominal predicate
B_Inx0Vs1	Imperative verb with sentence complement
B_N0nx0Ax1	Relative clause on subject of adjective predicate
B_N0nx0Ax1s2	Relative clause on subject of adjective predicate with sentence complement
B_N0nx0BEnx1	Relative clause on subject of equative
B_N0nx0N1	Relative clause on subject of nominal predicate
B_N0nx0Pnx1	Relative clause on subject of preposition predicate
B_N0nx0Px1	Relative clause on subject of exhaustive preposition predicate
B_N0nx0V	Relative clause on subject of intransitive verb
B_N0nx0Varb1	Relative clause on subject of verb with adverb complement
B_N0nx0Vax1	Relative clause on subject of verb with adjective complement
B_N0nx0Vnx1	Relative clause on subject of transitive verb
B_N0nx0Vnx1Pnx2	Relative clause on subject of dative verb

B_N0nx0Vnx1arbx2	Relative clause on subject of verb with noun phrase and adverb complement
B_N0nx0Vnx1nx2	Relative clause on subject of double object verb
B_N0nx0Vnx1pl	Relative clause on subject of transitive verb particle (unshifted)
B_N0nx0Vnx1pnx2	Relative clause on subject of verb with noun phrase and preposition phrase complement
B_N0nx0Vnx1s2	Relative clause on subject of verb with noun phrase and sentence complement
B_N0nx0Vnx2nx1	Relative clause on subject of double object verb
B_N0nx0Vpl	Relative clause on subject of intransitive verb particle
B_N0nx0Vplnx1	Relative clause on subject of transitive verb particle
B_N0nx0Vplpnx1	Relative clause on subject of verb particle with preposition complement
B_N0nx0Vpnx1	Relative clause on subject of verb with preposition phrase complement
B_N0nx0Vs1	Relative clause on subject of verb with sentence complement
B_N1nx0BEnx1	Relative clause on object of equative verb
B_N1nx0Vnx1	Relative clause on object of transitive verb
B_N1nx0Vnx1Pnx2	Relative clause on object of dative verb
B_N1nx0Vnx1nx2	Relative clause on object of double object verb
B_N1nx0Vnx1pnx2	Relative clause on object of verb with noun phrase and preposition phrase complement

B_N1nx0Vnx1s2	Relative clause on object of verb with noun phrase and sentence complement
B_N1nx0Vplnx1	Relative clause on object of transitive verb particle
B_N1nx0Vpnx1	Relative clause on object of verb preposition complement
B_N1nx1V	Relative clause on object of passive transitive verb
B_N1nx1VPnx2	Relative clause on object of passive dative verb
B_N1nx1Vax2	Relative clause on object of passive verb with adjective complement
B_N1nx1Vbynx0	Relative clause on object of passive transitive verb with by_phrase
B_N1nx1Vnx2	Relative clause on object of passive double object verb
B_N1nx1Vpl	Relative clause on object of passive verb particle
B_N1nx1Vpnx2	Relative clause on object of passive verb with preposition phrase complement
B_N1nx1Vs2	Relative clause on object of passive verb with sentence complement
B_N2nx0Vnx1pnx2	Relative clause on indirect object of dative verb
B_N2nx1Vpnx2	Relative clause on indirect object of passive dative verb
B_Nnx0Ax1	Adjunct relative clause for an adjective predicate
B_Nnx0BEnx1	Adjunct relative clause for an equative
B_Nnx0N1	Adjunct relative clause for a nominal predicate

B_Nnx0Pnx1	Adjunct relative clause for a prepositional predicate
B_Nnx0V	Adjunct relative clause for an intransitive
B_Nnx0Vnx1	Adjunct relative clause for a transitive
B_Nnx0Vnx1nx2	Adjunct relative clause for double object verb
B_Nnx0Vnx1pl	Adjunct relative clause for a transitive verb particle
B_Nnx0Vnx1s2	Adjunct relative clause for a verb with noun phrase and sentence complement
B_Nnx0Vpl	Adjunct relative clause for a verb particle
B_Nnx0Vpnx1	Adjunct relative clause for a verb with preposition phrase complement
B_Nnx0Vs1	Adjunct relative clause for a verb with sentence complement
B_Nnx1V	Adjunct relative clause for a passive transitive verb
B_Nnx1Vpnx2	Adjunct relative clause for a passive verb with preposition phrase complement
B_Nnx1Vs2	Adjunct relative clause for a passive verb with sentence complement
B_nx0A1s1s	Presentential clausal adjunct of adjective predicate with sentence complement
B nx0Ax1s	Presentential clausal adjunct of adjective predicate
B_nx0Vnx1pnx2s	Presentential clausal adjunct of dative verb
B_nx0Vnx1s	Presentential clausal adjunct of transitive verb

B_nx0Vpls1	Presentential clausal adjunct of intransitive verb particle
B_nx0Vpnx1s	Presentential clausal adjunct of transitive verb particle
B_nx0Vs	Presentential clausal adjunct of intransitive verb
B_nx0Vs1	Verb with sentence complement
B_nx1Vbynx0s	Presentential clausal adjunct of passive transitive verb with by_phrase
B_nx1Vps	Presentential clausal adjunct of passive verb with preposition complement
B_nx1Vs	Presentential clausal adjunct of passive verb
B_nx1Vs2	Passive verb with sentence complement
B_nxVnxpus	Parenthetical sentence complement verb
B_nxVpus	Parenthetical sentence complement verb
B_nxVpuvx	Parenthetical sentence complement verb
B_nxVs	Parenthetical sentence complement verb
B_pN1nx0Vpnx1	Relative clause on the object of verb with preposition complement
B_pN2nx1Vpnx2	Relative clause on the indirect object of dative verb
B_sVnx	Parenthetical sentence complement verb
B_snxV	Parenthetical sentence complement verb
B_snxVnx	Parenthetical sentence complement verb
B_vxnx0ARBPnx1	Postsentential clausal adjunct of complex preposition predicate
B_vxnx0Ax1	Postsentential clausal adjunct of adjective predicate

B_vxnx0V	Postsentential clausal adjunct of intransitive verb
B_vxnx0Vax1	Postsentential clausal adjunct of verb with adjective complement
B_vxnx0Vnx1	Postsentential clausal adjunct of transitive verb
B_vxnx0Vnx1nx2	Postsentential clausal adjunct of double object verb
B_vxnx0Vnx1pl	Postsentential clausal adjunct of transitive verb particle (shifted)
B_vxnx0Vpnx1	Postsentential clausal adjunct of verb with preposition phrase complement
B_vxnx1V	Postsentential clausal adjunct of passive transitive verb
B_vxnx1Vbynx0	Postsentential clausal adjunct of passive transitive verb with by_phrase
B_vxnx1Vnx2	Postsentential clausal adjunct of passive ditransitive verb
B_vxnx1Vpnx2	Postsentential clausal adjunct of passive dative verb

Appendix F

Variations of the term *epilepsy*

The following is a list of the different variations of the term *epilepsy*, as found in the UMLS metathesaurus.

```
c0014544  epilepsy
c0014544  epilepsy,_nos
c0014544  epilepsy,_unspecified
c0014545  epilepsy_nec
c0014547  epilepsy,_focal
c0014547  epilepsy,_localization-related
c0014547  epilepsy,_partial
c0014547  focal_epilepsy
c0014547  localization-related_epilepsy
c0014547  partial_epilepsy
c0014548  epilepsy,_generalized
c0014548  epilepsy,_generalized_primary
c0014548  epilepsy_seizure,_generalized
c0014548  generalised_epilepsy
c0014548  generalized_epilepsy
```

```
c0014548  generalized_epilepsy,_nos
c0014548  generalized_primary_epilepsy
c0014548  primary_epilepsy,_generalized
c0014549  epilepsy,_grand_mal
c0014549  epilepsy,_major
c0014549  epilepsy,_tonic-clonic
c0014549  epilepsy,_tonic_clonic
c0014549  epilepsy_grand_mal
c0014549  familial_epilepsy
c0014549  grand_-_mal_epilepsy
c0014549  grand_mal_epilepsy
c0014549  idiopathic_epilepsy
c0014549  mal_epilepsy,_grand
c0014549  primary_generalized_epilepsy,_major
c0014549  tonic-clonic_epilepsy
c0014550  epilepsy,_myoclonic
c0014550  epilepsy,_myoclonic,_progressive
c0014550  epilepsy,_myoclonic_progressive_familial
c0014550  epilepsy,_myoclonus
c0014550  epilepsy,_progressive_myoclonic
c0014550  myoclonic_epilepsy
c0014550  myoclonic_epilepsy,_progressive
c0014550  myoclonus_epilepsy
c0014550  progressive_familial_myoclonic_epilepsy
c0014550  progressive_myoclonic_epilepsy
c0014553  absence_epilepsy
c0014553  childhood_-_juvenile_-_absence_epilepsy
```

c0014553 childhood_absence_epilepsy

c0014553 epilepsy,_absence

c0014553 epilepsy,_minor

c0014553 epilepsy,_petit_mal

c0014553 epilepsy_petit_mal

c0014553 petit-mal_epilepsy

c0014553 petit_mal_epilepsy

c0014553 primary_generalized_epilepsy,_minor

c0014553 pykno-epilepsy

c0014553 pyknoepilepsy

c0014556 abdominal_epilepsy

c0014556 epilepsy,_abdominal

c0014556 epilepsy,_temporal_lobe

c0014556 epilepsy_temporal_lobe

c0014556 temporal_lobe_epilepsy

c0014556 temporal_lobe_epilepsy_syndrome

c0014556 tle_-_temporal_lobe_epilepsy

c0014557 epilepsy,_post-traumatic

c0014557 epilepsy,_post_traumatic

c0014557 epilepsy,_traumatic

c0014557 post-traumatic_epilepsy

c0014557 traumatic_epilepsy

c0014558 epilepsy,_uncinate

c0014558 uncinate_epilepsy

c0017332 epilepsy,_generalized,_nonconvulsive

c0017332 epilepsy,_nonconvulsive,_generalized

c0017332 generalised_nonconvulsive_epilepsy c0017332 generalised_nonconvulsive_epilepsy_nos

c0017332 generalized_nonconvulsive_epilepsy

c0017332 generalized_nonconvulsive_epilepsy,_nos

c0017332 generalized_nonconvulsive_epilepsy_nos

c0019563 hippocampal_epilepsy_syndrome

c0029617 other_forms_of_epilepsy

c0029617 other_forms_of_epilepsy_nos

c0030597 partial_epilepsy_nec

c0030599 partial_epilepsy,_without_mention_of_impairment_of_consciousness

c0030599 partial_epilepsy-conscious_ok

c0030599 partial_epilepsy_without_impairment_of_consciousness

c0030599 partial_epilepsy_without_impairment_of_consciousness_nos

c0030599 partial_epilepsy_without_mention_of_impairment_of_consciousness

c0030599 partial_epilepsy_without_mention_of_impairment_of_consciousness_nos

c0037769 infantile_myoclonic_epilepsy

c0085417 complex_epilepsy,_partial

c0085417 complex_partial_epilepsy

c0085417 epilepsy,_complex_partial

c0085417 epilepsy,_partial_complex

c0085417 epilepsy,_psychic_equivalent

c0085417 epilepsy,_psychomotor

c0085417 equivalent_epilepsy,_psychic

c0085417 partial_complex_epilepsy

c0085417 partial_epilepsy,_complex

c0085417 psychic_equivalent_epilepsy

c0085417 psychomotor_epilepsy

c0085541 epilepsy,_frontal_lobe

c0085541 frontal_lobe_epilepsy

c0085541 lobe_epilepsy,_frontal

c0085543 epilepsy,_kojevnikov's

c0085543 kojevnikov's_epilepsy

c0085543 kojewnikov's_epilepsy

c0086236 atonic_epilepsy

c0086236 epilepsy,_atonic

c0086237 cryptogenic_epilepsy

c0086237 epilepsy,_cryptogenic

c0086238 epilepsy,_generalized_secondary

c0086238 generalized_secondary_epilepsy

c0086238 secondary_epilepsy,_generalized

c0086238 secondary_generalised_epilepsy

c0086238 secondary_generalized_epilepsy

c0086240 epilepsy,_simple_partial

c0086240 partial_epilepsy,_simple

c0086240 simple_partial_epilepsy

c0086241 epilepsy,_tonic

c0086241 tonic_epilepsy

c0154707 generalized_nonconvulsive_epilepsy,_without_mention_of_intractable_epilepsy

c0154708 generalized_nonconvulsive_epilepsy,_with_intractable_epilepsy c0154709 generalized_convulsive_epilepsy,_without_mention_of_intractable_epilepsy

c0154710 generalized_convulsive_epilepsy,_with_intractable_epilepsy

c0154712 partial_epilepsy,_with_impairment_of_consciousness,_without_mention_of_intractable_epilepsy

c0154712 partial_epilepsy,_without_mention_of_impairment_of_consciousness,_with_intractable_epilepsy

c0154713 partial_epilepsy,_with_impairment_of_consciousness,_with_intractable_epilepsy

c0154714 partial_epilepsy,_without_mention_of_impairment_of_consciousness,_without_mention_of_intractable_epilepsy

c0154715 infantile_spasms,_without_mention_of_intractable_epilepsy

c0154716 infantile_spasms,_with_intractable_epilepsy

c0154717 epilepsia_partialis_continua,_without_mention_of_intractable_epilepsy

c0154718 epilepsia_partialis_continua,_with_intractable_epilepsy

c0154719 other_forms_of_epilepsy,_without_mention_of_intractable_epilepsy

c0154720 other_forms_of_epilepsy,_with_intractable_epilepsy

c0154721 epilepsy,_unspecified,_without_mention_of_intractable_epilepsy

c0154722 epilepsy,_unspecified,_with_intractable_epilepsy

c0234971 epilepsy_equivalent

c0234978 epilepsy,_jacksonian

c0234978 jacksonian,_epilepsy

c0234978 jacksonian_epilepsy

c0238111 epilepsy,_petit_mal,_myoclonic_astatic

c0238111 hemiconvulsion-hemiplegia-epilepsy_syndrome

c0238111 hemiplegia_-_hemiconvulsion_-_epilepsy_syndrome

c0260337 partial_epilepsy,_with_impairment_of_consciousness

c0260337 partial_epilepsy_with_impairment_of_consciousness

c0260337 partial_epilepsy_with_impairment_of_consciousness_nos

c0262486 epilepsy,_focal_motor

c0262486 focal_motor_epilepsy

c0270816 epilepsy_and_migraine

c0270819 cursive_epilepsy

c0270819 running_epilepsy

c0270820 epilepsy,_gelastic

c0270820 gelastic_epilepsy

c0270822 centrencephalic_epilepsy

c0270824 visual_epilepsy

c0270825 visceral_epilepsy

c0270847 benign_focal_epilepsy_of_childhood

c0270849 extratemporal_epilepsy

c0270850 epilepsy_idiopathic,_generalized

c0270850 idiopath_generalised_epilepsy

c0270850 idiopath_generalized_epilepsy

c0270850 idiopathic_generalised_epilepsy

c0270850 idiopathic_generalized_epilepsy

c0270850 idiopathic_generalized_epilepsy,_nos

c0270850 primary_generalised_epilepsy

c0270850 primary_generalized_epilepsy

c0270853 impulsive_petit-mal_epilepsy

c0270853 juv_myoclonic_epilepsy_of_janz

c0270853 juvenile_myoclonic_epilepsy

c0270853 juvenile_myoclonic_epilepsy_of_janz

c0270853 myoclonic_epilepsy_of_adolesc

c0270853 myoclonic_epilepsy_of_adolescence

c0270854 epilepsy_symptomatic,_generalized

c0270854 symptomatic_generalis_epilepsy

c0270854 symptomatic_generalised_epilepsy

c0270854 symptomatic_generalized_epilepsy

c0270854 symptomatic_generalized_epilepsy,_nos

c0270856 benign_childhood_epilepsy_with_centro_-_temporal_spike

c0270856 benign_rolandic_epilepsy

c0270856 centralopathic_epilepsy

c0270856 centrotemporal_epilepsy

c0270856 temporal-central_focal_epilepsy

c0270856 temporal-centrl_focl_epilepsy

c0270857 epilepsy,_sensory-induced

c0270857 epilepsy_ass_with_spec_stimuli

c0270857 epilepsy_associated_with_specific_stimuli

c0270857 reflex_epilepsy

c0270857 sensory_induced_epilepsy

c0311334 epilepsy,_generalized,_convulsive

c0311334 epilepsy_seizure,_generalized,_convulsive

c0311334 generalized_convulsive_epilepsy

c0311334 generalized_convulsive_epilepsy,_nos

```
c0312423  epilepsy,_reading
c0312423  reading_epilepsy
c0312424  epilepsy,_musicogenic
c0312424  musicogenic_epilepsy
c0338477  merrf_-_myoclonic_epilepsy_-_ragged_red_fibers
c0338477  merrf_-_myoclonic_epilepsy_-_ragged_red_fibres
c0338477  myoclonic_epilepsy_-_ragged_red_fibers
c0338477  myoclonic_epilepsy_-_ragged_red_fibres
c0338478  epilepsy_idiopathic,_myoclonic
c0338478  idiopathic_myoclonic_epilepsy
c0338479  epilepsy_symptomatic,_myoclonic
c0338479  symptomatic_myoclonic_epilepsy
c0338661  limbic_epilepsy_persnalty_synd
c0338661  limbic_epilepsy_personality_syndrome
c0347873  epilepsy,_psychosensory
c0347873  psychosensory_epilepsy
c0347874  epilepsy,_somatosensory
c0347874  somatosensory_epilepsy
c0347875  cursive_(running)_epilepsy
c0347875  epilepsy,_cursive_running
c0376532  epilepsy,_rolandic
c0376532  epilepsy,_rolands
c0376532  rolandic_epilepsy
c0376532  rolands_epilepsy
```

Appendix G

Results of SNC-Value

The following is a list of the results of the SNC-Value measure applied to the orpus of eye pathology records. It shows the term, the SNC-Value score, the NC-Value score, and the C-Value score.

Term	SNC	NC	C
bowman's_membrane	605782	1154.09	1077
malignant_melanoma	231237	715.904	733.333
hyaline_fibrous_tissue	215843	599.398	622.89
planes_of_section	170016	448.591	413
trabecular_meshwork	157353	681.77	623
keratinous_debris	101644	273.974	238
bruch's_membrane	94996.2	413.027	364.5
plane_of_section=	90109.4	506.517	437
melanoma_of_choroid	71615.1	497.673	420.5
lymphocytic_infiltration	53822	328.383	298
ciliary_processes	52355.7	421.204	366.429
cellular_fibrous_tissue	51486.8	296.412	307.483

squamous_epithelium	46928.9	198.431	149.5
optic_nerve_head	39054.5	422.211	424.77
pupillary_border	36510.8	285.464	269
corneal_epithelium	31335.9	268.517	244.5
scleral_invasion	31017.4	252.997	201
granulation_tissue	28010.1	178.068	169
stratified_squamous_epithelium	27445.5	203.753	217.14
ocular_structures	26143.6	309.758	299.333
conspicuous_nucleoli	25009.9	130.26	124
basal_cell_carcinoma	24903.6	1268.21	1065
bullous_separation	24747	203.011	179
keratinous_cyst	23223	392.944	378
conjunctival_epithelium	22607.3	184.7	169
collagenous_tissue	20570.4	151.587	149
laminated_keratin	18235.4	65.1729	5
basal_cell_papilloma	18069	616.614	551
two-tiered_epithelium	17475.1	181.277	187
ellipse_of_skin	17048.4	267.636	318
vitreous_haemorrhage	16779.8	204.881	170
basement_membrane	15909.6	164.016	147.429
corneal_diameter	15241	93.2743	19
retinal_vessels	14046.3	161.824	155
scleral_laceration	13703.7	94.0543	62
wedge_of_lid_margin	13678.5	211.414	256.764
neovascular_glaucoma	10884.3	192.303	175
vascular_pannus	10696.8	101.2	85.5
fibrovascular_tissue	10645.5	119.21	108.75

bullous_keratopathy	10279.1	139.472	125.333
viral_papilloma	8748.94	159.944	154
scleral_extension	8648.25	102.589	56
hyperkeratotic_papilloma	8539.71	96.385	60
basaloid_papillae	7756.25	94.9357	79.8
mitotic_figures	7637.76	127.296	115
stromal_lamellae	7552.19	118.003	93
scar_track	7364.21	228.217	235
acanthotic_epidermis	7040.26	100.719	83
cortical_fibres	6782.62	127.974	111.333
central_fibrous_core	6316.57	206.424	196.535
corneal_scarring	6110.22	109.111	91
acanthotic_papillae	5887.87	87.0986	72.4
corneal_laceration	5860.85	66.3743	59
vascular_tissue	5446.42	55.1257	21.5
sub-retinal_space	5318.64	130.04	120
peripheral_cornea	5230.94	127.584	110
epithelial_cells	5150.14	91.153	73.3333
pigment_laden_macrophages	4789.46	129.096	129.967
scleral_lamellae	4664.59	125.056	102
filtration_angle	4655.53	96.1886	86
intraocular_haemorrhage	4579	106.986	97
keratotic_papilloma	4528.06	49.9235	3
vesicular_nuclei	4245.21	61.4357	54
globe_of_approx	4240.06	38.6514	4
ciliary_process	4222.39	97.29	6
ocular_haemorrhage	4142.08	53.7932	4

pyknotic_debris	4001.05	86.4157	90
vascular_spaces	3548.96	66.0886	62
connective_tissue	3541.24	167.053	103
squamous_papilloma	3504.53	122.536	111
peripheral_retina	3376.64	97.03	89
horizontal_plane	3349.83	81.5043	83
pigment_epithelium	3162.57	78.2814	69
cystoid_degeneration	3147.98	74.4203	42
surgical_excision	3126.39	57.6824	57.3333
pigment_granules	3122.65	65.4643	58
vertical_plane	3083.24	167.015	140.21
basaloid_proliferation	3028.65	60.9386	49.5
lymphocytic_response	3019.24	68.3086	59
line_of_excision	2967.24	86.0071	86
equatorial_region	2877.42	73.78	68
trabecular_spaces	2877.13	36.6514	10.5
haemorrhagic_vitreous	2811.65	62.76	58
remnants_of_lens_capsule	2776.57	75.8626	76.078
stump_of_optic_nerve	2745.63	993.15	990
polygonal_cells	2731.34	81.29	75
vitreous_chamber	2719.23	66.9761	61
amorphous_debris	2708.05	59.2571	52
squamous_papillae	2696.92	46.4186	29.5
subretinal_coagulum	2682.55	66.73	49.5
pigment_proliferation	2649.87	75.0671	67
cranial_arteritis	2608.13	74.0945	66
choroidal_haemorrhage	2563.04	43.9629	14

elastotic_degeneration	2304.05	61.6057	59
sudoriferous_cyst	2291.79	204.624	211
fibrovascular_pannus	2247.33	53.8929	51
sebaceous_glands	2231.3	54.4219	48.5
conspicuous_nucleolus	2211.46	100.521	88
excision_biopsy	2210.24	155.257	140
basal_cell_carcinoma_of_lid	2208.01	66.7071	82
diffuse_infiltration	2185.68	63.3529	67
line_of_surgical_excision	2184.57	106.047	112.532
papillomatous_lesion	2180.98	58.4714	51
scleral_boundary	2114.18	95.6643	88
limbal_laceration	2071.72	37.3957	32
epithelioid_cells	2050.56	65.513	58
opaque_corneal_disc	2033.05	82.9817	93.5128
keratotic_debris	1957.17	49.6743	38
squamous_eddies	1892.06	72.7714	64
corneoscleral_laceration	1891.72	32.3371	30
remnants_of_iris	1881.88	52.42	46
morgagnian_globules	1822.78	44.0286	42
corneal_lamellae	1766.54	52.5757	44
thrombotic_glaucoma	1766.33	75.1629	71
suprachoroidal_haemorrhage	1742.88	54.6357	48
penetrating_keratoplasty	1729.96	82.7732	80
plane_of_section	1709.77	1752.71	1377
tiered_epithelium	1698.17	47.7014	16
descemet-like_membrane	1683.92	46.5171	35
hyperkeratotic_papillae	1639.3	50.44	46.75

herpetic_keratitis	1620.62	66.1476	53
keratinising_epithelium	1600.54	26.5429	5
papillary_proliferation	1546.39	48.6286	33
squamous_differentiation	1521.45	50.2129	45
trabeculectomy_specimen	1498.63	100.579	99
benign_hyperkeratotic_papilloma	1430.6	110.046	120.457
intertrabecular_spaces	1376.57	75.6357	65
granulomatous_reaction	1367.33	58.4329	55
bulbar_conjunctiva	1275.86	48.6969	44
pseudocyst_formation	1274	39.2	31.5
anterior_peripheral_synechiae	1262.47	90.1761	85.588
corneal_dystrophy	1259.83	55.2555	47.3333
lymphoid_tissue	1249.87	43.5495	36
corneal_perforation	1239.02	56.0644	48
columnar_epithelium	1236.93	22.9486	18
electron_microscopy	1209.28	44.623	39
aggregation_of_naevoid_cells	1209.12	53.9784	57.0587
molluscum_contagiosum	1198.01	99.8343	96
prominent_nucleoli	1171.96	27.1286	26
superior_calotte	1157.46	40.7557	42
acanthotic_epithelium	1145.91	30.7214	29
abundant_cytoplasm	1136.81	53.8771	48
retinal_atrophy	1134.84	46.8943	32
expulsive_haemorrhage	1130.02	46.8886	46
cell_carcinoma_of_lid_margin	1090.74	68.6	82
histological_examination	1085.42	38.6271	37
hyaline_fibrous_pannus	1041.05	47.7548	47.5489

keratinous_pseudocysts	1030.64	38.7457	38
vascular_connective_tissue	993.14	50.6704	55.4737
peripheral_iridectomy	981.531	51.1214	48
equatorial_diameter	969.192	42.6957	35
posterior_boundary	964.01	37.8043	37
pupillary_membrane	961.528	30.9173	28
perforated_corneal_ulcer	954.418	55.4894	55.4737
keratinous_cysts	949.831	58.9957	57
nodular_excrescence	940.324	46.0943	45
lacrimal_tissue	925.797	48.2186	44.25
squamous_cell_carcinoma	921.809	49.0324	50.7188
naevus_of_lid_margin	917.87	70.0664	66.0401
remnants_of_haemorrhagic_vitreous	896.352	40.9293	39.6241
anterior_capsule	888.271	51.9457	48
globe_enucleated	873.774	56.7386	57
mature_cataract	864.738	49.6976	36
remnants_of_vitreous	858.204	34.0557	29
adherent_leucoma	855.228	48.8702	40
inward_traction	851.835	35.7914	33
stromal_atrophy	846.507	42.7529	40
scleral_rupture	833.886	38.7854	35
angle_closure_glaucoma	817.679	63.386	63.3985
anterior_boundary	814.197	37.6943	36
retinal_haemorrhages	799.783	32.9129	30
adequacy_of_excision	790.772	27.1743	25
striated_muscle_fibres	789.315	39.075	39.6241
fibroblastic_proliferation	778.96	42.8	40

angle_closure_glaucoma	776.712	60.2102	63.3985
chorio-retinal_atrophy	764.29	40.4386	36
fibrovascular_membrane	750.777	24.2186	22
papillary_structure	724.427	28.7471	24
attenuated_epithelium	715.944	38.91	37
anterior_synechiae	704.886	35.4214	22
necrotic_debris	697.923	36.1618	32
microcystoid_degeneration	692.025	35.6714	30
corneal_decompensation	687.26	34.5357	34
red_blood_cells	686.118	51.5878	47.5489
capillary_proliferation	681.148	34.4014	28
dermal_collagen	676.343	33.9871	31
cyclitic_membrane	670.083	35.0829	30
limbal_conjunctiva	665.157	45.8729	40
ciliary_epithelium	665.08	41.0543	37
band_keratopathy	659.556	53.6224	44
membranous_tissue	650.524	43.9543	45
sub-retinal_coagulum	650.507	28.2829	21.5
dyskeratotic_cells	636.918	34.2429	28
conjunctival_lesion	627.331	33.5471	33
choroidal_melanoma	625.019	36.9834	25.6667
two-_tiered_epithelium	620.821	35.8856	38.0391
clumps_of_pigment	605.522	37.1486	30
cataract_surgery	593.108	33.5089	28
vitreous_cavity	592.638	34.4557	32
vascular_fibrous_tissue	592.274	35.8954	34.8692
hyaline_sclerosis	590.04	32.78	27

sebaceous_gland	576.705	17.1129	6
hyperkeratotic_basaloid_papillae	566.394	33.7139	34.8692
remnants_of_retina	530.422	33.1514	31
remnant_of_lens_capsule	527.649	30.1514	30.1143
pseudophakic_bullous_keratopathy	527.256	52.2036	50.7188
proliferation_of_basaloid_cells	525.35	27.0799	22.982
prolapse_of_iris	520.188	46.0343	38
spindle_cell_melanoma	515.829	38.7841	38.0391
stratified_columnar_epithelium	515.345	20.4502	22.1895
gelatinous_subretinal_coagulum	513.967	49.8997	45.9639
lattice_dystrophy	505.272	41.4157	42
rubeotic_glaucoma	494.586	34.83	33
retinal_arterioles	491.523	34.6143	34
fibrous_strands	487.315	23.4286	21
hyperchromatic_cells	483.556	25.8586	24
plaque_of_hyaline_fibrous_tissue	481.6	34.4	36
diameter_corneal_disc	479.898	41.7303	44.379
subretinal_fluid	478.388	29.17	23
cellular_oedema	478.354	19.9314	7
epidermal_papillae	476.768	26.4871	19
dislocation_of_lens	468.156	43.3478	34
anterior_stroma	465.819	42.7357	35
anterior_segment_structures	464.155	33.1539	34.8692
squamous_carcinoma	457.993	24.4916	21
lens_epithelium	455.374	31.19	29
capsular_cataract	453.291	16.9772	4
vascular_adhesions	452.099	26.7514	20

sphincter_pupillae	451.905	35.0314	36
residual_vitreous	451.122	32.69	27
foci_of_necrosis	450.346	28.5029	25
balloon_cell_degeneration	445.437	23.5681	23.7744
multinucleate_cells	441.908	29.0729	27
posterior_segment	439.964	33.08	28
hyperkeratotic_lesion	438.853	17.6957	9
epithelial_cyst	438.306	25.19	21
germinal_centres	436.095	26.43	27
pupillary_region	429.351	46.6686	41
stratified_epithelium	424.959	18.2386	17
reticulin_stroma	421.677	33.2029	27
wedge_resection	420.82	35.0683	37
limbal_corneosclera	418.142	30.7457	25
histological_appearances	417.417	30.03	30
hyaline_fibrosis	417.34	29.81	28
iris_epithelium	415.136	25.6257	23
limbal_corneo-sclera	410.256	32.56	29
vascular_granulation_tissue	408.026	25.662	28.5293
fibrous_proliferation	407.569	25.1586	21
lymphocytic_reaction	403.58	32.8114	28
adnexal_structures	396.408	36.0371	31
muscular_artery	394.729	24.8257	26
inverted_follicular_keratosis	393.623	44.2273	42.794
abundant_fibrous_stroma	390.292	44.8611	45.9639
parakeratotic_debris	385	30.8	28
meibomian_glands	382.166	26.7249	26

fibrinous_exudate	379.423	24.1671	21
corneal_ulceration	378.352	25.5643	23
papillary_lesion	377.451	18.6857	14
anterior_border	375.4	26.8143	25
laden_macrophages	374.834	32.0371	4
linear_scar_track	372.259	19.49	19.0196
cataractous_lens	366.216	18.8771	18
peripheral_anterior_synechiae	362.972	29.9977	30.1143
naevus_of_conjunctiva	360.186	83.7642	61.6667
compound_cellular_naevus_of_conjunctiva	357.28	61.6	70
subretinal_exudate	355.027	32.2752	27.3333
interstitial_keratitis	352.496	44.6197	42
tarsal_conjunctiva	345.748	28.34	26
degenerative_pannus	341.549	20.21	19
supra-choroidal_haemorrhage	341.443	27.9871	26
vitreous_abscess	338.764	34.2186	28
biopsy_of_lesion	335.213	18.9386	19
globular_tumour	331.764	19.2886	21
fibrovascular_core	329.695	21.1343	16
vascular_encroachment	327.93	19.0657	10
oblique_scar_track	324.557	22.5387	22.1895
keratotic_lesion	321.183	13.3271	3
retrocorneal_membrane	320.639	23.4043	21
vascular_proliferation	319.14	19.2253	10
periphery_of_descemet's_membrane	317.573	41.2432	44.379
granulomatous_follicles	316.951	18.0086	18
intradermal_naevus_of_lid_margin	314.28	58.2	62

dermal_naevus_of_lid_margin	313.5	19	18
cystic_structure	310.896	20.32	18
angular_dermoid	306.11	41.9329	39
sprinkling_of_lymphocytes	302.828	40.3771	40
hyperchromatic_nuclei	298.773	22.8071	22
central_opacification	297.149	27.2614	25
scleral_perforation	295.13	23.9943	23
anterior_synechia	293.271	22.5593	4
anterior_uveitis	292.659	27.3513	23
increased_pigmentation	289.382	21.9229	17
hyperplastic_primary_vitreous	285.233	18.1677	17.4346
peripheral_descemet's_membrane	282.475	32.4684	38.0391
corneal_membrane	280.83	18.2357	4
posterior_cornea	277.928	47.9186	11
capsular_epithelium	276.744	23.4529	20
aqueous_drainage_veins	271.562	27.9961	25.3594
horny_excrescence	270.359	21.4571	23
pigment_deposition	269.088	24.0257	24
laceration_of_cornea	268.938	12.06	12
fibrous_trabeculae	263.489	24.3971	20
papillary_proliferations	260.328	16.5814	14
anterior_vitreous	258.844	26.9629	21
choroidal_malignant_melanoma	258.172	31.8731	33.2842
gelatinous_sub-retinal_coagulum	257.174	27.3589	25.3594
decompensated_cornea	255.678	19.82	19
primary_vitreous	255.256	15.5644	6
depletion_of_cells	253.621	20.7886	19

vascular_channels	252.256	16.8171	16
necrotic_tumour	251.992	24.23	22
eosinophilic_inclusions	251.712	15.6343	14
basisquamous_papillae	250.906	18.7243	14
lateral_canthus	250.452	24.0819	21
trabecular_remnants	248.502	24.3629	25
corneal_abscess	242.37	21.8351	19
stromal_scarring	242.225	12.2957	3
lymphoid_follicles	241.243	18.8471	20
drusen_formation	240.16	26.1043	23
capillary_haemangioma	239.597	42.0345	35
muscle_insertions	238.842	21.7129	20
mucous_membrane	238.73	17.5537	15
central_retinal_artery	238.408	24.5781	23.7744
central_ulceration	236.934	18.8043	17
sutured_laceration	235.984	13.8814	14
flakes_of_keratin	235.339	18.8271	14
vitreous_remnants	234.88	21.5486	18
pyogenic_granuloma	232.098	34.6415	32
aphakic_glaucoma	231.919	27.9421	26
multilocular_cyst	231.274	12.1086	12
pannus_formation	229.136	19.5843	17
conjunctival_cyst	226.445	31.8937	29
diabetic_retinopathy	223.68	21.3029	17
rupture_of_globe	220.367	18.0629	15
disciform_keratitis	215.734	19.6122	18
green_dichroism	215.414	15.17	17

pigment_laden_cells	214.639	22.5936	20.6045
basaloid_papilloma	213.374	18.3943	14
stromal_infiltration	212.901	21.9486	22
pilosebaceous_follicles	212.282	14.2471	13
eosinophilic_debris	212.143	15.7143	15
optic_nerve_stump	211.345	26.092	28.5293
desmoplastic_stroma	209.805	24.6829	23
pilosebaceous_units	207.348	16.9957	13
cystoid_macular_oedema	206.768	33.3496	33.2842
surgical_aphakia	206.391	32.2486	32
inferior_rectus	202.66	17.3214	15
biopsy_of_papilloma	201.681	14.2029	15
pyknotic_nuclei	201.135	14.6814	14
sub-epithelial_naevus	197.591	16.6043	16
collections_of_lymphocytes	193.069	20.1114	19
aggregation_of_naevus_cells	192.622	21.1673	22.1895
oedematous_fibrous_tissue	191.683	17.4257	19.0196
tongue_of_hyaline_fibrous_tissue	189.8	14.6	16
marginal_chalazion	188.592	33.6771	33
melanoma_of_conjunctiva	185.613	43.1657	38
keratin_pseudocyst_formation	184.518	31.2743	30.1143
anterior_staphyloma	183.024	17.9435	14
necrotic_retina	182.302	15.9914	12
pupillary_aperture	181.764	20.4229	20
cystic_swelling	181.08	14.3714	11
congenital_glaucoma	179.866	24.6392	22
pilosebaceous_follicle	179.809	9.66714	6

intracapsular_cataract	178.56	19.2	20
inferior_calotte	173.632	17.5386	18
islands_of_epithelium	172.803	18.7829	17
conjunctival_naevus	171.779	25.6386	24
acellular_fibrous_tissue	169.882	19.0879	20.6045
periphery_of_tumour	169.687	19.5043	17
deposition_of_pigment	169.339	19.4643	18
peripheral_palisading	167.194	18.5771	18
islands_of_basaloid_cells	166.926	13.1438	12.6797
basophilic_inclusions	166.772	18.9514	15
hyperkeratotic_squamous_papillae	165.217	14.1211	15.8496
fibrovascular_adhesions	164.132	18.6514	15
papillary_excrescence	162.908	12.5314	9
anterior_choroid	162.792	21.42	19
aggregations_of_lymphocytes	162.5	15.9314	13
sessile_papilloma	161.815	20.4829	18
hyaline_granules	160.8	12	2
calcification_of_bowman's_membrane	160.787	23.3024	23.7744
vascularised_cornea	160.065	22.5443	13
oedematous_connective_tissue	159.438	18.5393	20.6045
inferior_pupillary_border	157.259	19.1779	20.6045
tunica_vasculosa_lentis	155.691	19.4614	19.0196
accumulations_of_lymphocytes	152.423	19.5414	17
macular_dystrophy	151.999	19.4871	18
prominent_nucleolus	150.834	21.86	20
red_cell_debris	150.403	16.0003	14.2647
capillary_channels	150.39	11.14	11

vascular_synechiae	149.911	12.8129	9
proliferations_of_basaloid_cells	149.859	13.8758	11.0947
squamous_cell_papilloma	149.024	29.8047	28.5293
keratin_cyst_formation	147.698	28.9604	25.3594
aggregations_of_naevoid_cells	146.292	17.4157	19.0196
cupping_of_nerve_head	145.543	18.4232	14.2647
epithelial_crypts	144.466	14.7414	14
perforated_cornea	142.065	18.45	9
meibomian_ducts	139.486	13.9486	14
intracellular_oedema	138.347	23.8529	20
band_of_hyaline_fibrous_tissue	136.08	12.6	12
ringschwiele_formation	134.626	16.22	16
lymphoid_hyperplasia	133.936	12.176	6
posterior_capsule	132.283	17.4057	15
granulomatous_tissue	130.777	9.83286	9
posterior_sclera	130.204	14.4671	8
fibrous_adhesions	130.079	16.4657	14
epithelial_bullae	129.999	11.3043	8
polypoid_lesion	129.791	10.6386	11
central_perforation	129.54	18.5057	17
scar_track_traverses	128.956	14.9949	13.4722
central_opacity	128.951	20.7986	19
peripheral_synechiae	127.215	24.0029	7
suture_profiles	126.768	16.68	16
intraocular_pressure	126.139	13.004	11
hyaline_fibrous_membrane	126.083	18.8183	15.8496
hyperkeratotic_acanthotic_papillae	125.587	10.5535	9.50978

benign_hyperkeratotic_lesion	125.371	32.9925	34.8692
pigmentation_of_basal	125.052	12.26	9
keratinous_flakes	124.905	7.95571	7
attenuated_iris	124.225	16.7871	13
vascular_stroma	123.704	19.0314	13
papillary_projections	123.274	11.4143	10
fibrovascular_proliferation	122.348	15.6857	14
vitreous_semi-fluid	122.12	14.2	11
acanthotic_epidermal_papillae	121.774	13.838	12.6797
congenital_cataract	121.72	18.7261	14
basisquamous_papilloma	119.645	23.0086	20
spikey_papillae	119.196	9.46	10
lateral_margins	119.104	18.61	18
oblique_direction	117.99	20.7	20
peripheral_cystoid_degeneration	117.885	21.4336	20.6045
collagen_fibres	117.844	17.33	16
reactive_lymphoid_hyperplasia	117.085	14.4549	14.2647
foci_of_lymphocytic_infiltration	117.055	15.8183	15.8496
peripheral_stroma	116.957	19.4929	19
face_of_vitreous	116.771	23.3543	21
membraneous_tissue	116.667	18.5186	16
anterior_retina	116.606	15.3429	11
trabeculectomy_site	116.485	14.0343	13
filtration_angles	116.305	20.4043	22
papilloma_of_lid_margin	115.065	18.5588	18.2271
vitreous_remnant	114.972	14.9314	8
epithelial_melanosis	114.356	12.43	10

basophilic_granules	114.317	19.0529	15
opaque_subretinal_coagulum	112.694	14.448	12.6797
basi-squamous_papilloma	112.482	20.83	20
viable_tumour_cells	112.336	19.04	19.0196
sections_of_calotte	112.304	8.57286	9
capsular_remnant	112.196	11.6871	9
warty_excrescence	112.092	12.0529	10
vascularized_scar_track	111.745	19.6043	19.0196
cilio-choroidal_melanoma	109.896	14.46	9
carcinoma_of_lid	109,683	167.015	173
glial_proliferation	109.489	13.1914	11
seborrhoeic_wart	108.454	24.1008	26
lens_implantation	108.22	13.0386	9
strands_of_cells	107.657	13.9814	12
remnants_of_choroid	106.594	12.1129	10
pilosebaceous_unit	105.24	11.1957	6
cell_degeneration	105.008	10.5008	1
corneal_lacerations	104.245	8.83429	8
acanthotic_papilloma	103.329	8.01	1
reticular_pattern	102.674	12.8343	13
band_of_fibrous_tissue	102.545	13.3175	14.2647
meibomian_gland	102.459	7.88143	4
seborrhoeic_keratosis	102.095	19.2632	18
limbal_staphyloma	101.82	12.1214	9
rounded_packets	101.706	6.31714	4
sebaceous_carcinoma	100.244	16.9905	17
subretinal_haemorrhage	99.8868	16.6478	11

hyperkeratotic_acanthotic_papilloma	99.2067	16.2634	17.4346
trabecular_meshwork=	98.4	16.4	13
penetrating_graft	98.124	13.26	14
islands_of_epithelial_cells	97.7067	15.0318	14.2647
septic_corneal_ulcer	97.6588	8.27617	6.33985
keratin_microcyst_formation	96.3935	22.4171	19.0196
epithelial_inclusion_cyst	95.1073	23.1969	25.3594
intercellular_oedema	94.3746	15.9957	14
haemorrhagic_remnants	94.2241	10.3543	8
papillary_formation	92.6015	10.5229	7
internal_haemorrhage	92.041	14.3814	13
diffuse_scarring	91.0161	10.1129	10
spikey_acanthotic_papillae	90.9936	12.638	12.6797
band_degeneration	90.8614	13.5614	14
line_of_resection	89.9885	28.1214	27
circumferential_fibres	89.3067	14.4043	15
granulomatous_infiltration	89.1681	11.2871	10
eosinophilic_cells	88.911	9.99	9
cholesterol_crystals	88.7634	19.2964	15
posterior_polymorphous_dystrophy	88.7052	15.294	15.8496
foci_of_lymphocytes	88.1758	11.4514	11
clumps_of_cells	88.1628	10.8843	10
section_of_calotte	87.7286	8.77286	9
plasmoid_exudate	87.4789	8.41143	7
horny_papilloma	85.8	13	12
ragged_corneal_disc	84.4106	16.2328	19.0196
increased_cellularity	84.1952	8.17429	6

vitreous_anteriorly	83.6709	17.0757	13
spindle_cell_malignant_melanoma	83.52	14.4	16
anterior_vitrectomy	83.2047	13.2071	11
episcleral_nodule	83.076	12.04	11
vascularised_vitreous	82.0115	10.5143	8
fibrous_capsule	81.3059	12.9057	9
sebaceous_hyperplasia	81.12	15.6	16
choroidal_invasion	80.7544	12.0529	13
limbal_perforation	80.6603	11.5229	10
intra_ocular_haemorrhage	80.5756	16.444	15.8496
plump_spindle_cells	80.2823	10.7043	11.0947
clumps_of_tumour_cells	80.2822	14.0846	14.2647
periphery_of_base	79.8815	14.0143	13
necrotic_remnants	79.3916	10.7286	10
granular_dystrophy	79.2184	15.2343	15
perforated_hypopyon_ulcer	79.106	20.2836	20.6045
histiocytic_reaction	78.7204	13.8106	13
fibrovascular_encroachment	78.514	10.61	8
array_of_papillae	78.4181	10.8914	10
cribriform_plate	78.257	13.7293	12
remnants_of_lens_fibres	78.0054	16.9577	17.4346
hyperkeratotic_epidermis	77.9499	11.1357	8
keratinous_pseudocyst_formation	77.7935	27.7834	28.5293
basal_carcinoma	77.2871	19.8172	3
sub-retinal_fluid	77.1543	8.97143	7
cellular_fibrous_stroma	77.0628	12.8438	12.6797
hyaline_membrane	77.0617	13.0613	6

malignant_cilio-choroidal_melanoma	76.8474	16.3505	17.4346
convoluted_capsule	76.5782	8.60429	8
cellular_debris	76.4895	10.1986	9
choroidal_tumour	76.4108	7.20857	6
sprinkling_of_polymorphs	76.1557	19.5271	18
laceration_of_limbus	75.5057	6.56571	5
retina_posteriorly	75.4669	15.4014	13
epithelial_nuclei	75.4191	13.0033	9.66667
dystrophic_graft	75.4	13	11
encroachment_of_fibrovascular_tissue	75.043	10.8758	11.0947
perforating_scar_track	74.1985	9.63617	6.33985
cellular_infiltration	73.383	12.03	11
margin_of_descemet's_membrane	73.0062	21.4724	23.7744
gliosed_remnants	72.6	11	9
ulcerated_lesion	72.3395	9.27429	8
suspensory_ligament	72.0131	16.0029	15
centre_of_tumour	71.8271	10.8829	10
abundant_stroma	71.3161	14.5543	13
collar_stud_tumour	70.67	10.8723	12.6797
vascular_invasion	70.4028	8.58571	7
central_retinal_vein_occlusion	70.24	15.6089	16
herpes_zoster_ophthalmicus	70.0275	12.0737	9.50978
aggregations_of_pigment	68.9005	10.7657	9
choroidal_vessels	68.8616	14.6514	15
chorioretinal_atrophy	68.2649	10.8357	9
cryptic_epithelial	67.978	8.29	8
glaucoma_surgery	67.8932	9.98429	9

stromal_vessels	67.5679	13.2486	11
infiltration_of_lymphocytes	67.1857	17.2271	12
pre-retinal_membrane	67.1177	11.0029	9
oedematous_stroma	67.0659	11.3671	7
ghost_epithelium	66.906	11.34	11
oblique_vascularised_scar_track	64.6167	9.64429	8
perforating_wound	64.5678	10.7613	8
remnants_of_bowman's_membrane	64.5479	16.9863	17.4346
papilloma_of_caruncle	63.4584	15.8646	14
fibrous_cyclitic_membrane	62.4405	9.46068	9.50978
central_corneal_ulcer	62.1639	12.6865	12.6797
drainage_operation	61.9935	14.4171	13
retro-corneal_membrane	61.9334	15.1057	13
conjunctival_flap	61.7537	7.91714	6
plaque_of_fibrous_tissue	61.4086	11.1652	12.6797
internal_boundary	61.1868	12.4871	10
infiltration_of_stroma	60.3653	13.1229	9
pilary_structures	60.2182	8.48143	8
vitreous_haemorrhagic	60.1716	10.0286	6
lymphocytic_foci	60.1174	11.3429	10
anterior_lens_capsule	60.0314	14.2932	14.2647
convoluted_lens_capsule	59.9633	9.83005	11.0947
epithelial_remnants	59.9201	12.2286	11
anterior_margin	59.8315	13.9143	13
solar_keratosis	59.698	12.7017	12
capillary_spaces	59.466	9.01	9
haemorrhagic_exudate	59.46	9.91	8

non-pigmented_tumour	59.4459	6.83286	4
posterior_border	59.0654	12.5671	12
papillary_epidermal	58.8343	8.91429	8
equatorial_staphyloma	58.5406	10.0932	8
suprachoroidal_space	58.3373	14.2286	13
pupillary_block	57.669	13.4114	12
lymphoid_follicle	57.6086	8.72857	6
incarceration_of_iris	57.4946	12.2329	12
epidermal_proliferation	57.11	8.15857	6
acanthotic_hyperkeratotic_papillae	56.7518	10.9138	12.6797
myelinated_nerve_fibres	56.6102	11.7938	12.6797
array_of_hyperkeratotic_papillae	56.5729	9.58862	11.0947
serous_exudates	56.5536	10.8757	10
laceration_of_sclera	56.4229	7.05286	6
nests_of_naevus_cells	56.3906	13.7538	12.6797
accumulation_of_lymphocytes	56.3436	12.2486	11
peripheral_iris	56.1716	14.0429	12
pedunculated_lesion	56.1391	10.2071	10
lamellar_keratoplasty	56.0378	10.9878	11
epithelial_crypt	55.8172	5.07429	3
neovascular_proliferation	55.8112	9.79143	8
aphakic_bullous_keratopathy	55.4458	17.3268	19.0196
cellular_pannus	55.396	7.19429	5
laciform_pattern	55.3936	15.3871	13
vitreous_haemorrhages	55.2428	9.20714	8
abundant_reticulin_stroma	55.1331	18.3777	17.4346
degenerative_cyst	54.9154	7.62714	5

pars_ciliaris_retinae	54.8763	11.6758	11.0947
retinal_vein_occlusion	54.4787	8.93093	6.33985
dystrophic_corneal_graft	54.4472	13.6118	14.2647
conjunctival_nodule	54.34	7.15	6
aggregations_of_naevus_cells	54.2447	13.9089	14.2647
downward_proliferation	52.632	9.39857	6
rosette_formation	52.6041	11.6898	10
sub-retinal_exudate	51.7383	7.60857	6
spiky_acanthotic_papillae	51.3505	8.85354	9.50978
adherent_fibrous_tissue	51.3481	14.6709	12.6797
internal_fibrosis	51.0016	10.8514	9
vertical_sections	51	10	9
naevus_of_caruncle	50.9824	19.6086	15
degenerate_lens_fibres	50.8092	12.7023	12.6797
lamination_of_descemet's_membrane	50.5246	14.0346	14.2647
columns_of_nuclei	50.3893	6.71857	6
host_descemet's_membrane	50.1543	11.6638	12.6797
keratin_microcysts	49.9845	11.6243	10
spindle_cell_transformation	49.8808	15.1154	15.8496
subcapsular_epithelium	49.6544	11.0343	11
diffuse_lymphocytic_infiltration	49.5918	13.4032	14.2647
plaques_of_bone	49.5772	8.26286	7
basaloid_proliferations	49.4846	7.27714	7
staining_serous	49.4448	8.10571	7
pilary_complexes	49.416	6.96	6
basophilic_inclusion	49.2486	8.95429	3
fibro_vascular_tissue	49.0936	7.43842	7.92481

convoluted_remnants_of_lens_capsule	48.88	9.4	10
ocular_pressure	48.5258	5.51429	1
acanthotic_cells	48.448	7.57	6
peripheral_vascular_synechiae	48.4196	7.56556	7.92481
degeneration_of_macula	48.3935	6.20429	6
stromal_reaction	48.0022	8.42143	6
allergic_conjunctivitis	47.686	12.8881	12
epithelial_proliferation	47.431	12.1618	10
keratin_formation	47.2011	9.07714	5
capillary_vessels	47.201	9.63286	8
internal_structure	47.103	9.61286	9
paraffin_sections	47.0811	9.80857	10
cystic_basal_cell_carcinoma	47.0001	13.4286	12
spindle_cell_tumour	46.915	9.02211	9.50978
cap_of_keratotic_debris	46.8676	10.1886	11.0947
calcifying_epithelioma	46.86	6.6	4
adherent_pigment	46.837	9.55857	7
recurrent_chalazion	46.48	9.48571	10
epithelial_papillae	46.41	5.46	5
bundles_of_collagen	46.3133	6.34429	5
bilateral_corneal_dystrophy	46.09	10.7186	11.0947
sebaceous_cysts	45.889	14.8029	14
epithelioid_histiocytes	45.8558	6.45857	6
aggregates_of_naevoid_cells	45.8007	8.80782	9.50978
central_vein_occlusion	45.6909	10.3843	11.0947
granulomatous_lesion	45.6867	8.30667	6
squamous_metaplasia	45.6428	9.12857	9

melanocytic_proliferation	45.595	8.29	7
haemorrhage_anteriorly	45.552	17.52	15
hyperkeratotic_squamous_epithelium	45.3328	10.3029	11.0947
hyperplastic_vitreous	45.3143	7.42857	6
separation_of_epithelium	45.058	10.4786	9
islands_of_cells	45.0387	10.0086	9
ulcerated_basal_cell_carcinoma	44.928	14.04	14
focal_accumulation	44.764	7.22	5
stroma_oedematous	44.3143	19.2671	5
nodule_of_tumour	44.08	11.02	9
warty_excrescences	44.0503	6.67429	6
compound_naevus	43.9636	13.3223	12
optic_nerve_anterior_chamber	43.9618	10.4671	10
lymphoid_infiltration	43.4834	7.49714	7
gelatinous_coagulum	43.132	10.52	8
sebaceous_units	42.9246	9.33143	1
papillary_region	42.92	5.8	3
edges_of_descemet's_membrane	42.7424	11.248	12.6797
epidermal_cells	42.7114	6.47143	6
posterior_loops	42.6523	7.89857	6
fluffy_opacities	42.5106	9.24143	8
oedematous_corneal_graft	42.3392	11.7609	12.6797
posterior_vitreous	42.1264	9.36143	7
posterior_retina	42.0965	9.15142	5
conjunctival_melanoma	42.064	7.51143	6
bilateral_keratoconus	41.7555	13.0486	11
production_of_basement_membrane	41.7409	9.27576	11.0947

residual_retina	41.712	7.44857	7
necrotic_atrophy	41.58	3.78	2
peripheral_pannus	41.52	7.41429	6
acellular_collagen	41.48	6.1	6
basaloid_epithelium	41.4412	6.09429	5
calcium_deposit	41.055	5.95	2
cholesterol_crystal	40.887	8.34429	3
mouth_of_funnel	40.4611	9.19571	8
optic_nerve_cornea	40.3449	7.61225	1.58496
cellular_composition	40.1977	10.3071	10
nodule_of_basal_cell_carcinoma	39.9907	12.4971	14
misshapen_globe	39.9703	6.89143	7
polymorphous_dystrophy	39.9111	4.80857	1
amelanotic_melanoma	39.8086	12.0632	11
keratohyaline_granules	39.6618	12.3943	10
eosinophilic_cytoplasm	39.4068	10.1043	8
ciliary_staphyloma	39.3397	10.0871	10
fibroblast-like_cells	39.2786	7.85571	6
fibrous_adhesion	38.812	7.61019	3
proliferation_of_cells	38.8043	8.43571	5
opaque_gelatinous	38.5029	8.02143	6
remnant_of_vitreous	38.3017	7.36571	5
angle_recession	38.1886	9.31429	10
epithelial_cysts	38.0157	5.84857	5
rupture_of_bruch's_membrane	38.0152	14.0797	15.8496
recurrent_papilloma	38	10	9
perforation_of_cornea	37.9962	8.08429	7

islands_of_epithelioid_cells	37.8631	7.42413	7.92481
histological_grounds	37.7988	8.21714	6
scanning_electron_microscopy	37.7709	11.1091	11.0947
prominent_intercellular_bridges	37.0783	7.13045	6.33985
point_of_attachment	36.234	10.98	11
tumour_bruch's_membrane	36.1993	11.6772	11.0947
strip_of_tissue	36.106	11.6471	13
corneal_aperture	36.0927	14.4371	14
fibrous_retrocorneal_membrane	36.0926	7.07699	7.92481
inclusion_cysts	35.8611	9.43714	9
inferior_fornix	35.7598	6.74714	7
cytoplasmic_inclusions	35.7043	6.61191	3
post_mortem_change	35.576	9.36211	9.50978
hyperkeratotic_papillary_lesion	35.5088	8.66068	9.50978
multiloculate_cyst	35.388	5.61714	5
vascular_strands	35.3614	4.77857	4
proliferation_of_basement_membrane	34.9725	7.60271	7.92481
epithelial_plaque	34.872	8.30286	8
opaque_vitreous	34.8532	6.45429	5
wedge_of_tissue	34.6846	5.59429	6
fibrovascular_cores	34.6743	8.66857	7
cellular_tissue	34.5714	7.85714	3
hyper-_keratosis	34.4186	7.64857	6
cellular_collagen	34.408	3.74	3
pleomorphic_cells	34.2193	7.60429	7
pilo-sebaceous_follicles	34.144	10.67	8
anterior_subcapsular_fibrosis	34.1229	7.93556	7.92481

follicular_conjunctivitis	33.7047	7.66016	7
epithelioid_character	33.4372	7.11429	4
macular_degeneration	33.3404	5.95365	4
examination_of_sections	33.1333	5.81286	5
malignant_lymphoma	33.0291	11.0097	11
cubical_epithelium	32.7357	6.54714	6
anaplastic_carcinoma	32.5557	7.07733	6
series_of_cystic_spaces	32.5534	7.93985	7.92481
deposits_of_calcium	32.4857	8.12143	7
medial_conjunctiva	32.3149	7.34429	7
granulomatous_exudate	32.2723	7.01571	6
sub-retinal_haemorrhage	32.1917	8.25429	7
hyaline_fibrous_plaque	31.2492	6.79331	6.33985
melanoma_of_iris	31.248	26.04	25
epithelial_denudation	31.03	8.86571	7
serosanguineous_exudate	30.996	6.32571	5
fibrocellular_pannus	30.7941	7.16143	6
hyperkeratosis_of_basket	30.5909	5.88286	3
peripheral_region	30.5743	5.09571	1
papillary_pattern	30.4928	6.09857	5
serous_subretinal_exudate	30.4787	12.1915	11.0947
scleral_aperture	30.2287	7.37286	7
scleral_continuity	30.149	6.15286	6
minimal_stromal	30.0052	6.52286	5
dystrophic_calcification	29.9342	9.07098	6
retinitis_pigmentosa	29.6462	6.73778	4
degeneration_of_retina	29.574	7.04143	4

corneal_vascularisation	29.3801	6.6773	5
central_corneal_epithelium	29.1588	9.11211	9.50978
recurrence_of_basal_cell_carcinoma	28.98	13.8	16
intra_ocular_lens	28.6167	9.86782	9.50978
carcinoma_cells	28.5391	4.83714	4
calcium_deposits	28.4629	8.37143	6
downward_proliferations	28.3808	7.27714	6
fibroblastic_tissue	28.3454	4.88714	5
hyperkeratotic_acanthotic_epidermis	28.1164	5.85759	6.33985
ruptured_capsule	28.0949	5.40286	3
papillary_formations	28.0937	8.26286	7
residual_stroma	28.04	7.01	6
cellular_reaction	28.0041	12.1757	11
neutrophil_polymorphs	28	7	7
central_stromal_core	27.9601	9.98576	11.0947
vacuolar_degeneration	27.8897	5.46857	5
fibrous_scar_tissue	27.7808	5.05105	4.75489
scleral_lacerations	27.7493	5.23571	4
sclera_posteriorly	27.748	9.91	8
birefringent_suture	27.726	6.60143	6
vascular_stromal_cores	27.5934	9.51497	9.50978
deposition_of_calcium	27.3188	8.53714	8
subcapsular_fibrosis_anteriorly	27.2508	8.25782	9.50978
proliferation_of_capillaries	26.97	6.42143	6
systemic_steroids	26.91	6.40714	6
hyaline_fibrous_proliferation	26.8123	5.70474	6.33985
granulomatous_focus	26.7994	7.44429	7

acanthotic_basaloid	26.784	5.58	4
perforated_ulcer	26.7359	7.86349	7
capsular_rupture	26.6072	10.6429	10
basi-squamous_papillae	26.6043	4.83714	5
sutured_scleral_laceration	26.588	7.38556	7.92481
degenerative_cysts	26.4386	8.81286	8
additional_information	26.3791	7.75857	8
clump_of_pigment	26.3099	6.11857	5
ulceration_of_cornea	26.2479	5.83286	5
rim_of_pannus_round_limbus	26.2308	13.8057	14
vascular_scar_track	26.0904	5.92963	4.75489
sympathetic_ophthalmia	25.9286	8.64286	8
peripheral_fibres	25.8068	6.61714	6
apparent_ulceration	25.2911	7.43857	5
intra_epithelial	24.868	8.88143	7
adenoid_cystic_basal_cell_carcinoma	24.8329	11.2877	11.6096
haemorrhage_posteriorly	24.8221	8.00714	6
suprachoroidal_effusions	24.7243	5.49429	5
aggregate_of_naevoid_cells	24.718	6.50474	6.33985
conjunctival_vessels	24.6586	6.01429	4
sympathetic_ophthalmitis	24.596	9.46	10
peripapillary_atrophy	24.5	7	7
edge_of_descemet's_membrane	24.3517	9.74068	9.50978
nodule_of_lacrimal_tissue	24.3484	7.16128	7.92481
cylinder_of_tissue	24.2444	3.61857	4
capillary_blood_vessels	24.1707	7.79699	7.92481
clumps_of_lymphocytes	24.14	6.89714	7

papillary_processes	24.1354	6.18857	6
stromal_fibrosis	23.8024	7.21286	7
plaques_of_hyaline_fibrous_tissue	23.68	7.4	8
subcapsular_cataract	23.6362	9.84842	5
subretinal_haematoma	23.5929	5.24286	4
ciliary_muscle_fibres	23.5343	8.7164	9.50978
fibrous_nodules	23.4928	5.10714	4
accumulation_of_naevus_cells	23.2819	5.54331	6.33985
peripapillary_region	23.2277	8.60286	8
descemets_membrane	23.0309	6.39746	6
phacolytic_glaucoma	22.7732	10.8444	11
identification_of_muscle	22.61	4.61429	3
pigment_dispersion	22.4872	6.81429	5
descemt's_membrane	22.306	5.87	6
epidermal_cysts	22.2467	6.74143	6
fibrovascular_connective_tissue	22.2146	5.4182	4.75489
benign_epithelial_melanosis	22.1399	9.22497	9.50978
groups_of_pigment	22.1143	5.52857	3
epithelial_oedema	22.0709	6.49143	3
epidermoid_cyst	22.066	10.5076	8
atrophy_of_ciliary_muscle	22.0016	6.28617	6.33985
tags_of_fibrous_tissue	21.9516	6.27188	6.33985
sessile_papillae	21.8972	6.84286	4
retinopathy_of_prematurity	21.8791	7.05778	6
crystal_profiles	21.8186	4.84857	4
stitch_granuloma	21.6572	10.8286	11
retinal_remnants	21.6187	5.27286	4

sector_iridectomy	21.6	7.2	7
glassy_appearance	21.564	7.70143	6
sutured_limbal_laceration	21.5019	5.51331	6.33985
viral_papillomata	21.4857	8.59429	7
corneal_scleral	21.414	4.98	5
cataract_section_scar	21.4062	8.91925	9.50978
hyperkeratotic_cap	21.2937	6.26286	6
occlusion_of_central_retinal_artery	21.28	7.6	8
lymphatic_channels	21.2785	4.25571	4
plica_semilunaris	21.16	4.6	4
fibrous_tissue_anteriorly	21.0281	6.18474	6.33985
plaque_of_tumour	20.8759	7.19857	4
reticulin_framework	20.7771	8.65714	9
recurrent_basal_cell_carcinoma	20.7	9	10
plate_of_fibrous_tissue	20.5964	9.80782	9.50978
scattered_pigment	20.5663	6.63429	3
reduced_calibre	20.3503	5.65286	2
umbilicated_lesion	20.3154	4.61714	4
histological_staining	20.3097	3.90571	4
polypoid_nodule	20.2926	4.41143	5
laceration_of_globe	20.258	4.13429	3
anterior_subcapsular_cataract	20.2574	12.6609	12.6797
subepithelial_pannus	20.2126	5.46286	5
cottage_loaf-shaped_tumour	20.1362	6.10188	6.33985
encroachment_of_pannus	20.0891	5.90857	5
dermal_aggregation	20.0749	5.28286	3
processes_of_basaloid_cells	20.0628	6.47188	6.33985

orbital_cellulitis	19.9765	4.16177	3
quadrant_of_cornea	19.9064	6.42143	4
proliferation_of_hyaline_fibrous_tissue	19.8	6	6
scarring_of_sclera	19.5306	5.74429	5
linear_deposits	19.476	5.41	5
sebaceous_cells	19.3692	6.05286	5
circumscribed_packets	19.3457	6.44857	5
indented_cornea	19.342	5.09	5
ovoid_excrescence	19.1137	7.35143	7
dermal_fibrosis	19.0299	4.64143	4
remnant_of_iris	19.016	6.79143	5
infero-_nasally	18.9071	7.56286	6
capsular_remnants	18.9	7	6
intraocular_structures	18.8983	7.26857	7
lymphoid_aggregates	18.888	4.49714	4
sprinkling_of_plasma_cells	18.6685	10.9815	11.0947
descemet's_membrane_splits	18.6287	5.47902	6.33985
cupping_of_optic_nerve_head	18.6	12.4	12
basophilic_curlicues	18.5491	3.56714	3
dermal_scarring	18.522	3.78	3
polymorph_infiltration	18.5214	3.70429	2
fibrocytic_cells	18.4734	4.50571	4
degenerate_remnants	18.4026	6.34571	6
basement_membrane-like	18.3074	2.20571	1
proliferation_of_fibrovascular_tissue	18.1885	6.27188	6.33985
eccentric_ulcer	18.1416	6.25571	6
inferior_leaflet	17.952	5.61	3

granulomatous_foci	17.82	8.1	7
strands_of_fibrous_tissue	17.8017	7.73985	7.92481
remnants_of_trabeculae	17.7678	7.10714	8
leash_of_vessels	17.7351	4.66714	5
ellipse_of_conjunctiva	17.7297	4.66571	5
membrane_stretches	17.7043	3.93429	3
fibrous_membrane_stretches	17.6271	5.68617	6.33985
factitious_damage	17.6	4.4	5
conjunctival_crypt	17.4536	6.46429	3
opacification_of_cornea	17.388	6.21	5
squamous_keratosis	17.364	4.13429	3
demodex_folliculorum	17.3575	5.42421	4
junctional_naevus	17.3099	9.11047	7
remnants_of_capsule	17.253	8.21571	7
conjunctival_scar	16.9572	4.84492	3
individual_cell	16.875	3.75	3
intraepidermal_carcinoma	16.87	5.62333	5
anterior_stromal	16.6126	7.22286	6
membrane_anteriorly	16.6	6.64	6
vessels_of_capillary	16.5699	4.04143	4
equatorial_lens_fibres	16.5129	5.00391	4.75489
depigmentation_of_iris	16.5094	5.00286	4
lobules_of_lacrimal_tissue	16.4413	5.87188	6.33985
accumulation_of_polymorphs	16.368	7.79429	7
parakeratotic_squames	16.338	3.89	4
reticular_dermis	16.1392	8.49429	7
migration_of_epithelial_nuclei	16.1321	5.20391	4.75489

cortical_equatorial_fibres	16.0443	9.43782	9.50978
sprinkling_of_eosinophils	16.0371	8.01857	9
mixture_of_red_cells	16.0125	5.00391	4.75489
severin_implant	16	4	3
basisquamous_epithelium	15.9709	4.20286	3
cell_infiltration	15.9493	3.54428	1
central_elevation	15.8306	4.79714	4
conjunctival_melanosis	15.8245	9.89032	8
fibrocytic_proliferation	15.8169	8.78714	8
fibro_papilloma	15.7634	6.06286	6
posterior_limit	15.741	5.83	5
anterior_sub-capsular	15.7063	6.54429	5
christie_hospital	15.64	6.8	7
keratinous_pseudo-cysts	15.6	6.24	6
purulent_exudate	15.5917	7.08714	5
brown_discolouration	15.5292	4.85286	4
posterior_uveitis	15.5266	8.62587	7
remnants_of_conjunctival_epithelium	15.4959	4.84248	4.75489
lid_papillomata	15.4069	4.53143	3
vasculosa_lentis	15.36	4.8	1
anaplastic_tumour	15.2511	8.47286	8
aggregations_of_polymorphs	15.1303	6.30429	5
invasion_of_scleral_lamellae	15.1211	4.72534	4.75489
retinal_venules	15.075	6.03	6
tumour_projects	15.0086	4.41429	4
tongue_of_fibrous_tissue	14.9839	4.68248	4.75489
basisquamous_fibropapilloma	14.96	4.4	3

parakeratotic_scale	14.9503	7.86857	5
scarring_of_cornea	14.9074	6.21143	4
cup-shaped_lesion	14.8514	3.22857	3
bilateral_cataract	14.84	4.24	4
cell_proliferation	14.8076	4.48714	3
episcleral_reaction	14.7947	3.99857	4
mushroom-shaped_tumour	14.784	4.62	5
fibro_vascular_pannus	14.7679	6.15331	6.33985
vascularised_corneal_scar	14.747	5.26677	4.75489
internal_calcification	14.595	9.73	10
bowmans_membrane	14.58	5.4	4
hyaline_fibrous_scar_tissue	14.5643	5.82571	6
plain_of_section	14.5414	5.01429	4
epithelium_basement_membrane	14.4941	5.3682	4.75489
collagenous_wall	14.4928	5.79714	6
scleral_canal_mid-way	14.4253	6.27188	6.33985
globular_non-pigmented_tumour	14.3387	6.51759	6.33985
ulcer_of_cornea	14.3356	6.23286	5
collection_of_cells	14.3053	4.93286	4
tongue_of_tumour	14.24	4.45	4
spiky_excrescence	14.2036	5.68143	6
admixture_of_epithelioid_cells	14.1794	4.43105	4.75489
wolter_preparation	14.08	4.4	4
scanty_haemorrhage	14.04	5.85	4
malignant_transformation	14.04	3.6	3
interstitial_haemorrhage	13.9623	3.67429	3
polymorphic_infiltration	13.92	5.8	5

inconspicuous_nucleoli	13.872	4.08	4
accumulations_of_polymorphs	13.8349	6.28857	6
vitreous_retraction	13.7584	5.09571	4
bizarre_mitoses	13.7311	4.03857	3
intraocular_haemorhage	13.7036	5.48143	4
degenerate_cells	13.6554	2.96857	3
centrocytic_lymphoma	13.6092	4.25286	4
conjunctival_plaque	13.5	5	4
optic_nerve_corneal_epithelium	13.482	6.42	6
occlusion_of_pupil	13.3926	5.82286	5
retinal_pigment	13.3236	5.79286	4
lines_of_resection	13.32	3.6	4
minimal_junctional	13.22	9.44286	9
limbus_indented	13.2094	4.00286	3
rupture_of_sclera	13.0972	3.27429	3
congenital_syphilis	13.0295	6.51476	6
spongy_meshwork_of_vascular_spaces	13.02	6.2	6
connective_tissue_stroma	13.0037	5.4182	4.75489
fibrous_plaques	12.9764	3.50714	3
brown_pigmentation	12.952	4.62571	5
focal_lymphocytic_infiltration	12.9099	6.14759	6.33985
cornea_vascularised	12.8304	6.75286	6
vascularized_hyaline_fibrous_tissue	12.76	5.8	6
vessel_formation	12.7414	4.24714	4
double_perforation	12.7386	3.44286	3
cellular_fibrous_nodule	12.7227	6.69617	6.33985
hippel-lindau_disease	12.7147	5.29778	4

cholesterol_profiles	12.604	4.50143	4
lesion_of_medial	12.5714	5.02857	5
whorls_of_keratin	12.5383	5.45143	5
basophilic_stippling	12.512	5.44	4
remnants_of_vascularised_vitreous	12.4894	5.20391	4.75489
chorio-_capillaris	12.4834	5.67429	4
posterior_stroma	12.4576	4.01857	3
atrophic_cupping_of_optic_nerve	12.42	5.4	6
remnants_of_ciliary_processes	12.3438	6.17188	6.33985
granulomatous_stroma	12.1754	5.53429	5
vessels_attenuated	12.0586	5.24286	5
capsular_nuclei	12.0323	5.23143	3
remnants_of_epithelium	12.0021	8.00143	7
posterior_polar	11.9841	4.43857	4
cellular_components	11.9494	4.42571	4
diffuse_stromal	11.9351	3.22571	3
admixture_of_pigment	11.9014	3.96714	3
focus_of_polymorphs	11.8417	5.14857	4
spinous_processes	11.832	4.93	3
clusters_of_cells	11.7269	3.78286	3
stromal_sclerosis	11.6957	3.89857	4
dysplastic_epithelium	11.6911	4.03143	4
macrocystoid_degeneration	11.6569	5.29857	4
multi-nucleate_cells	11.6281	5.05571	4
sub-capsular_fibrosis_anteriorly	11.5719	6.09045	6.33985
skeletal_muscle_fibres	11.5534	4.81391	4.75489
prussian_blue_staining	11.509	5.00391	4.75489

rheumatoid_arthritis	11.488	6.38222	5
pseudo_horn_cysts	11.4324	8.79413	7.92481
choroid_posteriorly	11.392	14.24	13
remnant_of_descemet's_membrane	11.3691	6.31617	6.33985
vein_of_reduced_calibre	11.3482	5.40391	4.75489
papillary_nodule	11.2644	3.88429	4
keratohyalin_granules	11.2426	5.91714	5
central_necrosis	11.22	8.01429	7
central_corneal_perforation	11.0621	4.80963	4.75489
sub-_retinal_coagulum	11.0581	6.50474	6.33985
zoster_ophthalmicus	11.0475	5.02159	1
rectus_insertion	11.0214	4.40857	3
diffuse_opacification	11.0057	5.50286	4
senile_macular_degeneration	11.0022	6.47188	6.33985
iris_including_sphincter	10.9856	5.78188	6.33985
cap_of_parakeratotic_debris	10.9583	4.98105	4.75489
proliferative_retinopathy	10.8231	4.00857	3
exudate_of_polymorphs	10.81	4.7	4
aggregation_of_lymphocytes	10.6586	5.92143	6
pearl_formation	10.626	4.83	4
posterior_protruberances	10.604	4.82	5
vein_thrombosis	10.54	6.2	5
hyaline_basement_membrane	10.5262	5.01248	4.75489
nuclei_posteriorly	10.4387	4.53857	3
herpetic_kerato_uveitis	10.4121	6.50759	6.33985
suture_filaments	10.3714	3.45714	3
profile_of_corneo-sclera	10.3641	3.83857	3

equatorial_staphylomata	10.3414	3.44714	3
region_bowman's_membrane	10.3183	3.68511	1.58496
sub-epithelial_vascular_pannus	10.3	5.42105	4.75489
pseudotumour_of_orbit	10.2214	3.40714	3
filtration_bleb	10.1948	5.36571	5
sebaceous_adenoma	10.1829	7.27352	5
keratinised_horn	10.1594	4.41714	3
chromatin_pattern	10.0857	4.03429	4
zone_of_necrosis	10.0569	5.58714	5
intra_ocular_structures	10.0526	5.91331	6.33985
posterior_scleral_staphyloma	10.0414	7.72413	7.92481
plate_of_tissue	10.0133	3.45286	3
aggregates_of_lymphocytes	9.95428	4.97714	4
reticular_strands	9.76628	3.61714	3
remnants_of_hair	9.75301	4.64429	4
pigment_aggregations	9.74871	4.23857	3
posterior_extension	9.728	3.04	3
lens_cataractous	9.709	5.11	4
subepithelial_naevus	9.66658	4.20286	4
series_of_cysts	9.58656	3.68714	3
radial_opacities	9.57872	5.04143	4
peripheral_break	9.537	4.54143	3
migration_of_nuclei_posteriorly	9.50743	5.00391	4.75489
cells_of_medium	9.49142	4.74571	4
remnants_of_lacrimal	9.48256	3.64714	3
lens_iris_diaphragm	9.47573	7.28902	6.33985
calcium_deposition	9.46514	5.91571	6

ischaemic_retinopathy	9.36	7.2	6
indrawn_ciliary_processes	9.32782	4.66391	4.75489
minimal_acanthosis	9.28	5.8	3
metaherpetic_keratitis	9.20914	5.41714	4
horizontal_scar	9.2	4	4
indentation_of_cornea	9.192	3.83	3
posterior_synechiae	9.17217	4.82746	4
cortical_cataract	9.16765	18.3353	17
domed_excrescence	9.14743	3.81143	4
sero-sanguineous_exudate	9.08915	7.57429	6
intraocular_tension	9.08799	3.24571	3
residual_epithelium	9.07999	6.48571	6
follicle_formation	9.06715	4.12143	3
conjunctival_margin	9.05544	3.48286	3
recurrent_lesion	9.05142	5.65714	5
cytological_characteristics	9.03215	3.61286	3
angle_deformity	9	3.6	4
pedunculated_papilloma	8.97371	5.60857	5
bullous_elevation	8.95372	4.97429	4
pigment_epithelial	8.934	6.38143	6
corneal_epithelial	8.91785	8.10714	6
peripheral_opacification	8.81872	4.64143	4
collection_of_polymorphs	8.76229	5.15429	4
lymphocytic_aggregates	8.75	3.5	3
excision_margin	8.72	4.36	3
rounded_profile	8.718	4.15143	3
bilateral_fuch's	8.70858	5.44286	6

translucent_appearance	8.70685	4.83714	3
feet_of_posterior_loops	8.70586	5.80391	4.75489
cap_of_keratinous_debris	8.67344	6.67188	6.33985
rubeosis_iridis	8.652	4.12	3
polypoid_basisquamous_papilloma	8.64704	4.80391	4.75489
supra-choroidal_space	8.64685	7.20571	6
nodular_excrescences	8.62858	3.45143	3
fibroglial_membrane	8.56	4.28	3
collections_of_polymorphs	8.55999	3.05714	3
dysthyroid_eye_disease	8.5562	6.58169	4.75489
cystic_degeneration	8.54743	4.74857	4
atrophic_corneal_graft	8.51384	7.73985	7.92481
lid_squamous_papilloma	8.48504	4.71391	4.75489
staining_reactions	8.482	6.05857	6
granuloma_of_conjunctiva	8.48054	6.52349	5
marfan's_syndrome	8.43505	7.02921	6
stratified_non-keratinizing	8.38515	3.81143	3
plane_of_section-	8.36	4.4	3
nodular_elevations	8.35715	4.64286	4
fibrous_subepithelial_pannus	8.32046	4.62248	4.75489
pre-equatorial_region	8.32	3.2	3
conical_horny_excrescence	8.31898	7.56271	7.92481
congestive_glaucoma	8.303	3.61	3
vitreous_strands	8.28038	3.94304	3
collagenous_bands	8.22	4.11	4
horizontal_section	8.20715	3.28286	3
hemispherical_tumour	8.144	2.90857	3

epithelial_dysplasia	8.12558	3.53286	3
serosanguinous_exudate	8.11172	7.37429	6
corneal_infection	8.106	3.86	3
haemorrhages_anteriorly	8.09571	5.39714	4
corneal_curvature	8.08228	3.10857	3
punctate_haemorrhages	8.04285	3.21714	3
sections_of_calottes	8.01599	3.81714	4
emissary_channel	7.95471	3.45857	3
confluent_haemorrhages	7.9337	3.30571	3
cyst_of_conjunctiva	7.92458	4.95286	5
choroidal_naevus	7.882	11.26	9
polypoid_excrescence	7.83086	3.26286	3
centre_of_globe	7.65258	4.78286	3
benign_hyperkeratotic_papillae	7.62157	5.08105	4.75489
lymphocytic_focus	7.60971	3.30857	3
cornea_indented	7.54972	6.29143	6
lines_of_surgical_excision	7.50122	4.41248	4.75489
recurrent_papillomatous_lesion	7.46226	4.66391	4.75489
region_of_cornea	7.44557	8.27286	6
amelanotic_tumour	7.44	3.1	3
base_posteriorly	7.40101	3.52429	3
atrophy_of_retina	7.34572	3.67286	3
posterior_cataract	7.32085	4.06714	4
aggregates_of_foam_cells	7.30947	5.22105	4.75489
myxomatous_degeneration	7.20555	1.67571	1.43571
equator_infero-nasally	7.2	6	4
conjunctival_oedema	7.17228	5.51714	4

capsular_epithelial_nuclei	7.16993	2.86797	1.58496
central_vacuole	7.14285	2.85714	3
penetrating_scar	7.10571	4.73714	4
coloboma_of_choroid	7.0849	5.06064	5
replacement_fibrosis	7.03028	3.90571	3
peripheral_vascularization	6.98058	4.36286	3
intracapsule_cataract	6.94929	3.02143	3
glial_adhesions	6.83643	4.02143	3
post_mortem_autolysis	6.77808	5.21391	4.75489
central_aperture	6.74771	3.06714	3
sclerosis_of_processes	6.74657	3.96857	3
depigmented_patches	6.74057	2.80857	3
necrotic_malignant_melanoma_of_choroid	6.72	33.6	40
corneo_scleral_laceration	6.64658	4.43105	4.75489
tissue_incorporating_remnants	6.63372	4.42248	4.75489
museum_specimen	6.54957	3.44714	3
collagenous_scar	6.51429	4.07143	3
plaque_of_fibrovascular_tissue	6.50508	5.00391	4.75489
orientation_of_globe	6.494	3.82	3
ductular_structures	6.48	3.6	3
herpetic_infection	6.46	3.4	3
dystrophy_of_cornea	6.46	3.4	3
intracellular_inclusions	6.4	3.2	3
sprinkling_of_pigment	6.347	5.77	5
bilateral_proptosis	6.34286	4.22857	3
vitreous_surgery	6.31285	4.20857	3
posterior_ciliary_vessels	6.2748	4.82677	4.75489

pseudoexfoliative_deposit	6.20485	3.26571	3
choroidal_lesion	6.14143	4.09429	4
vitreous_aspirate	6.12	3.4	3
unknown_aetiology	6.12	3.6	3
conjunctival_papillomata	6.01	6.01	6
proliferation_of_acanthotic	5.94	6.6	4
pseudo_exfoliative	5.88185	3.09571	3
fibres_of_orbicularis	5.87657	4.89714	4
schisis_of_retina	5.81867	4.84889	3
crystalline_dystrophy	5.8	5.8	6
peripheral_chorio-retinal	5.78742	3.61714	3
membrane_splits	5.78742	1.80857	1
necrotic_optic_nerve	5.78637	4.45105	4.75489
hyaline_collagenous	5.778	3.21	3
series_of_hyperkeratotic	5.748	4.79	5
postmortem_change	5.71685	3.17603	3
opacification_of_lens	5.67514	6.30571	6
crystalline_keratopathy	5.668	8.09714	8
space_posteriorly	5.66614	2.09857	1
anterior_haemorrhage	5.56428	5.56428	3
extent_of_laceration	5.54058	3.46286	3
stromal_margins	5.52857	3.07143	3
ischaemic_optic_neuropathy	5.50869	6.12077	6.33985
rounded_elevation	5.47429	3.42143	3
collections_of_red_cells	5.46508	4.20391	4.75489
accumulations_of_pigment	5.44429	5.44429	4
fluid_posteriorly	5.42857	5.42857	4

naso-lacrimal_duct	5.408	4.16	4
convoluted_profile	5.39743	5.99714	5
corneal_rupture	5.36914	3.35571	3
epithelioid_component	5.26171	3.28857	3
intractable_pain	5.25215	3.50143	3
optic_nerve_stump=	5.10469	5.67188	6.33985
halves_sectioned	5.05286	5.61429	5
cyst_of_apocrine	5.04458	3.15286	3
cystoid_limbal_scar	5.02534	5.02534	4.75489
side_of_nerve_head	5.01248	5.01248	4.75489
papilloma_of_skin	5.01119	3.85476	3
degeneration_of_cornea	5	5	5
lobular_excrescence	4.96786	3.82143	4
globe_sectioned	4.961	4.51	4
dyskeratotic_plaque	4.94	3.8	3
fibrous_scarring	4.87686	3.75143	3
granuloma_of_conjunctiva	4.87026	3.74635	1.52857
starch_granules	4.84275	3.72519	3
internal_contraction	4.84158	3.72429	3
indurated_ulcer	4.82786	3.21857	3
syringocystadenoma_papilliferum	4.8	3.2	3
nodular_structure	4.71729	5.24143	5
pilosebaceous_cyst	4.68514	5.20571	6
keratinous_horn	4.64571	4.64571	5
supra-_choroidal	4.63857	4.63857	3
nodules_of_lacrimal	4.60342	2.87714	3
intraocular_haemorrhages	4.58915	3.82429	3

optic_neuropathy	4.56	2.4	2
fibrous_histiocytoma	4.53911	6.48444	6
useful_information	4.51658	2.82286	3
suprachoroidal_oedema	4.5	4.5	4
supraciliary_oedema	4.43143	4.02857	4
tissue_posteriorly	4.42457	3.68714	3
epithelial_edges	4.416	3.68	3
reticulin_fibres	4.414	3.15286	3
eccentric_opacity	4.41	4.41	4
nodular_hidradenoma	4.36266	5.45333	4
keratin_pseudocysts	4.35	4.35	3
recurrent_nodule	4.16	3.2	3
recipient_button	4.16	5.2	5
internal_structures	4.149	4.61	4
stroma_anteriorly	4.13143	4.13143	3
central_corneal_opacity	4.08542	5.10677	4.75489
residual_ciliary_processes	4.06256	5.0782	4.75489
collections_of_cells	4.06	4.06	4
lobulate_appearance	4.022	2.87286	3
weeks_gestation	4	5	4
traces_of_mucin	3.97885	3.31571	3
complement_of_ganglion_cells	3.96352	4.40391	4.75489
bluish_discolouration	3.88714	3.88714	4
acanthotic_epithelial	3.87257	3.22714	3
plasmoid_vitreous	3.83115	3.48286	3
superfical_stroma	3.83	3.83	3
araldite_sections	3.751	3.41	3

lens_capsule_posteriorly	3.74255	6.23759	6.33985
kerato_acanthoma	3.72743	3.38857	3
sub-epithelial_fibrovascular	3.66457	3.33143	3
lens_anteriorly	3.654	5.22	3
chamber_implant	3.64	2.8	2
granulomatous_lesions	3.61571	3.61571	3
jaw_claudication	3.60622	9.01556	6
precipitation_of_calcium	3.59228	3.26571	3
calcification_posteriorly	3.57714	4.47143	4
ophthalmic_artery	3.456	3.84	3
formation_posteriorly	3.45085	2.87571	1
hypertensive_retinopathy	3.432	3.81333	3
paracentral_region	3.42143	3.42143	3
tissue_anteriorly	3.41228	1.89571	1
anterior_suface	3.3	3	3
mature_sebaceous	3.27657	4.09571	4
branches_of_retinal_artery	3.12235	5.20391	4.75489
plaque_formation	3.121	4.45857	4
post_contusional	3.10971	3.88714	3
stromal_vascularisation	3.10171	3.87714	3
adherent_debris	3.10143	3.10143	3
diffuse_cloudiness	2.94043	3.26714	3
chondroid_syringoma	2.928	4.88	4
oblique_linear_scar	2.92692	4.8782	4.75489
lamellar_pattern	2.71929	3.02143	3
myxomatous_stroma	2.59843	2.88714	3
invasion_of_muscle	2.34714	4.69429	4

dermal_calcium_deposit	2.24156	1.86797	1.58496
conjunctival_dermoid	2.2	4.4	4
vascularized_cornea	1.83714	4.59286	3
clumps_of_epithelioid	1.56	3.9	3
orbital_pseudotumour	1.38	4.6	4
accumulation_of_lymphoid	1.26343	3.15857	3
biopsy_of_conjunctival	1.09671	3.65571	3
vitreous_liquid	0.78514	3.92571	3
fibrous_cataract	0.61057	6.10571	1
apparent_limbal	0	4.49857	4
squamous_papillomata	-2.88	7.2	7
basal_cell_papillomata	-3.82352	38.2352	42.794
carcinoma_of_lid_margin	-22.9438	5.73594	3.16993
peripheral_synechia	-65.782	21.22	6

Bibliography

Aitchison, J. (1994). *Words in the Mind: an introduction to the mental lexicon.* Blackwell, Oxford.

Ananiadou, S. (1988). *Towards a methodology for automatic term recognition.* PhD thesis, University of Manchester, UK.

Ananiadou, S. (1994). A methodology for automatic term recognition. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan.

Arnold, E. (1993). Special issue on evaluation of MT systems. *Machine Translation*, 8(1-2):1–126.

Basili, R., Della Rocca, M., Pazienza, M., and Velardi, P. (1995). Contexts and categories: Tuning a general purpose verb classifi cation to sublanguages. In *Proc. of International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

Baud, R., Rassinoux, A.-M., and Scherrer, J.-R. (1992). Natural language processing and medical records. In *Proc. of Medinfo '92*, pages 1362–1367.

Blois, M. S. (1988). Medicine and the nature of vertical reasoning. *New England Journal of Medicine*, 318(13):847–851.

Boguraev, B. and Kennedy, C. (1997). Technical terminology for domain specification and content characterisation. In Pazienza, M., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Lecture Notes in Artificial Intelligence No. 1299. Springer.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of 14th International Conference on Computational Linguistics (COLING)*, pages 977–981, Nantes, France.

Bourigault, D. and Habert, B. (1998). Evaluation of terminology extractors: Principles and experiments. In *Proc. of 1st International Conference on Language Resources and Evaluation (LREC)*, pages 299–305, Granada, Spain.

Bowden, P. R., Evett, L., and Halstead, P. (1998). Automatic acronym acquisition in a knowledge extraction program. In *Proc. of 1st Workshop on Computational Terminology (Computerm '98)*, pages 43–49, Montreal, Canada. COLING-ACL.

Bowker, L. (1995). *A multidimensional approach to classification in Terminology: Working with a computational framework*. PhD thesis, University of Manchester, England.

Bowman, C., Michaud, D., and Suonuuti, H. (1997). Do's and don'ts of terminology management. In Wright, S. and Budin, G., editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 2.2.4, pages 215–217. John Benjamins, Amsterdam.

Braden-Hader, L. (1993). Sense disambiguation using online dictionaries. In Jenon, K., Heidhorn, G., and Richardson, S., editors, *Natural Language Processing: the PLNLP Approach*, pages 247–261. Kluwer Academic Publishers, Dordrecht.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proc. of 3rd Conference of Applied Natural Language Processing.*

Brill, E. (1994). *A Corpus-Based Approach to Language Learning.* PhD thesis, Dept. of Computer and Information Science, University of Pennsylvania.

Bruce, R. and Wiebe, J. (1994). Word sense disambiguation using decomposable models. In *Proc. of 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, USA.

Campbell, K., Oliver, D. E., Spackmann, K. A., and Shortliffe, E. H. (1998). Representing thoughts, words and things in the UMLS. *Journal of the American Medical Informatics Association*, 5(5):421–431.

Chai, J. and Biermann, A. (1997). The use of WordNet in automatic information extraction. In *Proc. of ACL Workshop on Word Sense Disambiguation*, Washington D.C.

Chandrasekar, R. and Bangalore, S. (1998). Knowing a word by the company it keeps using local information in a maximum entropy model for word sense disambiguation. Submitted for publication.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Cimino, J., Hripcsak, G., Johnson, S., and Clayton, P. (1989). Designing an introspective, controlled medical vocabulary. In Kingsland, L., editor, *Proc. of 14th Annual Symposium on Computer Applications in Medical Care*, pages 513–518. IEEE Computer Society Press.

Clark, E. (1973). What's in a word? On the child's acquisition of semantics in his first language. In Moore, T., editor, *Cognitive development and the acquisition of language.* Academic Press, New York.

Cole, R., editor (1996). *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.

Coltheart, M., Besner, D., Jonasson, J., and Davelaar, E. (1979). Phonological encoding in a lexical decision task. *Quarterly Journal of Experimental Psychology*, 31:489–507.

Condamines, A. and Rebeyrolle, J. (1998). CTKB: a corpus-based approach to a terminological knowledge base. In *Proc. of 1st Workshop on Computational Terminology (Computerm '98)*, pages 29–35, Montreal, Canada. COLING-ACL.

Connine, C., Ferreira, F., Jones, C., Clifton, C., and Frazier, L. (1984). Verb frame preferences - descriptive norms. *Journal of Pyscholinguistic Research*, 13(4):307–319.

Côté, R., editor (1993). *Systematized Nomenclature of Medicine - SNOMED International*. College of American Pathologists.

Cowie, J., Guthrie, J., and Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In *Proc. of DARPA Workshop on Speech and Natural Language*, pages 238–242, New York.

Crookshank, F. (1923). The importance of a theory of signs and a critique of language in the study of medicine. In Ogden, C. and Richards, I., editors, *The Meaning of Meaning*, pages 337–355. Routledge & Kegan Paul Ltd. supplementary essay.

Dagan, I., Itai, A., and Schwall, U. (1991). Two languages are more informative than one. In *Proc. of 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 130–137, Berkeley, California.

Dagan, I., Pereira, F., and Lee, L. (1994). Similarity-based estimation of word co-occurrence probabilities. In *Proc. of the 32nd Annual Meeting of the Association for COmputational Linguistics (ACL'94)*, pages 272–278.

Dahlgren, K. (1988). *Naive semantics for natural language understanding*. Kluwer, Boston, USA.

Daille, B., Gaussier, E., and Langé, J. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, pages 515–521.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433–447.

Davidson, L., Kavanagh, J., Mackintosh, K., Meyer, I., and Skuce, D. (1998). Semi-automatic extraction of knowledge-rich contexts from corpora. In *Proc. of 1st Workshop on Computational Terminology (Computerm '98)*, pages 50–56, Montreal, Canada. COLING-ACL.

De Vries Robbé, P. and Flier, F. (1994). Towards a common vocabulary for classification and definition. *Terminology*, 1(1):97–102.

deRoeck, A., Kruschwitz, U., Neal, P., Scott, P., Steel, S., Turner, R., and Webb, N. (1998). An intelligent directory enquiry assistant. *BT Technology Journal*, 16(3).

Do Amaral, M. and Satomura, Y. (1995). Associating semantic grammars with the SNOMED: Processing medical language and representing clinical facts into a language-independent frame. In *Proc. of 8th World Congress on Medical Informatics (Medinfo)*, pages 18–22, Edmonton, Alberta.

Dorr, B. and Jones, D. (1996). Role of word sense disambiguation in lexical acquisition: predicting semantics from syntactic cues. In *Proc. of 16th International Conference on Computational Linguistics (COLING)*, Copenhagen.

Dubuc, R. and Lauriston, A. (1997). Terms and contexts. In Wright, S. and Budin, G., editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 1.3.3, pages 80–87. John Benjamins, Amsterdam.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

EAGLES Computational Lexicons Working Group (1996). Preliminary recommendations on semantic encoding. Technical report, EAGLES. http://www.ilc.pi.cnr.it/EAGLES96/rep2/.

Ellman, J. (1997). Agent assisted internet resource discovery by generic document profiling. Unpublished report. http://osiris.sunderland.ac.uk/ cs0jel/research/index.htm.

Enguehard, C. and Pantera, L. (1994). Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27–32.

Evans, D. A. and Zhai, C. (1996). Noun phrase analysis in unrestricted text for information retrieval. In *Proc. of 34th Annual Meeting of the Assocation for Computational Linguistics (ACL'96)*, pages 17–24, Santa Cruz, USA.

Felber, H. (1984). *Terminology Manual*. Unesco and Infoterm, Paris.

Fowler, H., Fowler, F., and Pearsall, J., editors (1999). *The Concise Oxford English Dictionar*. Oxford University Press.

Frantzi, K. (1998). *Automatic Recognition of Multi-Word Terms*. PhD thesis, Manchester Metropolitan University, England.

Frantzi, K. and Ananiadou, S. (1995). Statistical measures for terminological extraction. In *Proc. of 3rd International Conference on Statistical Analysis of Textual Data (JADT '95)*, pages 297–308.

Frantzi, K. and Ananiadou, S. (1996). Extracting nested collocations. In *Proc. of the 16th International Conference on Computational Linguistics (COLING)*, pages 41–46, Copenhagen, Denmark.

Frantzi, K. and Ananiadou, S. (1997). Automatic term recognition using contextual cues. In *Proceedings of 3rd DELOS Workshop*, Zurich, Switzerland.

Frantzi, K. and Ananiadou, S. (1999). The C-Value/NC-Value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

Frege, G. (1892). Uber Sinn und Bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik*, 100:25–50. (Translated by M. Black under the title "On Sense and Reference", in P. Geach and M. Black, Translations from the Philosophical Writings of Gottlob Frege, Oxford, 1952.).

Gale, W., Church, K., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In Zernik, U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum Associates, Hillsdale, N.J.

Gale, W., Church, K., and Yarowsky, D. (1992). Work on statistical methods for word sense disambiguation. In *Probabilistic Approaches to Natural Language: Papers from the 1992 Fall Symposium*, pages 54–60, Cambridge, Massachussetts. AAAI Press.

Gardiner, J. and Java, R. (1990). Recollective experience in word and nonword recognition. *Memory and Cognition*, 18:23–30.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers.

Harris, Z. (1968). *Mathematical Structures of Language.* Wiley (Interscience), New York.

Herviou-Picard, M.-L. (1996). Les outils d'indexation AlethIP issus du projet GRAAL: principes et utilisation. Technical Report HN-46/96/022, Direction des Etudes et Recherches, EDF, Clamart, France.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proc. of 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, Pennsylvania.

Hirschman, L., Grishman, R., and Sager, N. (1975). Grammatically based automatic word class formation. *Information Processing and Retrieval*, 11:39–57.

Howe, D. (1998). The free on-line dictionary of computing. http://wombat.doc.ic.ac.uk/foldoc/.

Hutchins, W. and Somers, H. (1992). *An Introduction to Machine Translation.* Academic Press.

ISO-704 (1987). International organisation for standardisation (ISO) - principles and methods of terminology.

Jacquemin, C. (1996). What is the tree that we see through the window: A linguistic approach to windowing and term variation. *Information Processing and Management*, 32(4):445–458.

Jacquemin, C. and Royaute, J. (1994). Retrieving terms and their variants in a lexicalised unification-based framework. In *Proc. of 13th Annual International ACM-SIGIR Conference in Research and Development in Information Retrieval*, pages 132–141, Dublin.

Joshi, A. and Srinivas, B. (1994). Disambiguation of super parts of speech (or supertags): Almost parsing. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan.

Jouis, C. (1994). Contextual approach: SEEK, a linguistic and computational tool for use in knowledge acquisition. In *Proc. of 1st European Conference in Cognitive Science in Industry.*

Jouis, C. and Elhadi, W. (1997). Aupelf project: Term and semantic relation extraction tools evaluation paradigm. In *Proc. of SALT Workshop on Evaluation in Speech and Language Technology*, Sheffield, UK.

Justeson, J. and Katz, S. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition. *Terminology*, 3(2):259–289.

Karlsson, F. and Karttunen, L. (1996). Sub-sentential processing. In Cole, R., editor, *Survey of the State of the Art in Human Language Technology.* Cambridge University Press.

Keil, F. C. (1979). *Semantic and conceptual development: an ontological perspective.* Harvard University Press, Cambridge, MA, USA.

Kilgarriff, A. (1997a). Evaluating word sense disambiguation programs: progress report. In *Proc. of SALT Workshop in Speech and Language Technology*, pages 114–120, Sheffield University, UK.

Kilgarriff, A. (1997b). Foreground and background lexicons and word sense disambiguation for information extraction. In *Proc. of International Workshop on Lexically Driven Information Extraction*, pages 51–62, Frascati, Italy.

Kilgarriff, A. (1997c). "I don't believe in word senses". *Computing and the Humanities*, 31(2):91–113.

Kiyono, M. and Tsujii, J.-I. (1994). Hypothesis selection in grammar acquisition. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan.

Kolodner, J. (1993). *Case-Based Reasoning.* Morgan Kaufmann Publishers Inc., San Mateo, California.

Krovetz, R. (1992). Sense-linking in a machine readable dictionary. In *Proc. of the 30th Annual Meeting of the Association for Computational Lingustics*, pages 330–332.

Krovetz, R. and Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Lauriston, A. (1995). Criteria for measuring term recognition. In *Proc. of 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 17–22.

Lauriston, A. (1996). *Automatic term recognition: performance of linguistic and statistical learning techniques.* PhD thesis, UMIST, Manchester, UK.

Lawson, V. (1997). The terms and arts of patentese: wolves in sheep's clothing. In Wright, S. and Budin, G., editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 2.1.3, pages 171–183. John Benjamins, Amsterdam.

Lehrberger, J. (1982). Automatic translation and the concept of sublanguage. In Kittredge, R. and Lehrberger, J., editors, *Sublanguage: Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

Lesk, M. (1986). Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. In *Proc. of 5th International Conference on Systems Documentation (ACM-SIGDOC)*, pages 24–26, New York.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proc. of ACL/EACL*, pages 64–71, Madrid, Spain.

Lin, D. (1998). Extracting collocations from text corpora. In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 57–63, Montreal, Canada. COLING-ACL.

Litkowski, K. C. (1999). Towards a meaning-full comparison of lexical resources. In *Proceedings of the Association for Computational Linguistics Special Interest Group on the Lexicon*, College Park. MD.

Lynch, C. (1997). Medical terminology management. In Wright, S. and Budin, G., editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 2.1.2, pages 160–170. John Benjamins, Amsterdam.

Magadur, J.-Y. and Tabuteau, G. (1996). Semantic disambiguation in an information retrieval system. In *NLP+IA 96*, pages 148–154, Moncton, Canada.

Marcus, Mitchell, P., Santorini, B., and Marcinkiewitz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2).

Marshall, P. and Bandar, Z. (1999). Working towards connectionist modelling of term formation. In *Proc. of Computer Intelligence Theory and Applications 6th Fuzzy Days International Conference*, Lecture Notes in Computer Science, Dortmund, Germany. Springer-Verlag.

Masterman, M. (1956). The potentialities of a mechanical thesaurus. *Machine Translation*.

Maynard, D. and Ananiadou, S. (1998a). Acquiring contextual information for term disambiguation. In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 86–90, Montreal, Canada.

Maynard, D. and Ananiadou, S. (1998b). Term disambiguation using the UMLS medical thesaurus. In *Proc. of 1st Computational Linguistics UK Research Colloquium (CLUK2)*, University of Sunderland, UK.

Maynard, D. and Ananiadou, S. (1998c). Term sense disambiguation using a domain-specific thesaurus. In *Proc. of 1st International Conference on Language Resources and Evaluation (LREC)*, pages 681–687, Granada, Spain.

Maynard, D. and Ananiadou, S. (1999a). Identifying contextual information for term extraction. In *Proc. of 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, pages 212–221, Innsbruck, Austria.

Maynard, D. and Ananiadou, S. (1999b). Incorporating linguistic information for multi-word term extraction. In *Proc. of 2nd Computational Linguistics UK Research Colloquium (CLUK2)*, University of Essex, UK.

Maynard, D. and Ananiadou, S. (1999c). A linguistic approach to context clustering. In *Proc.of Natural Language Processing Pacific Rim Symposium (NLPRS)*, pages 346–351, Beijing, China.

Maynard, D. and Ananiadou, S. (1999d). Term extraction using a similarity-based approach. In *Recent Advances in Computational Terminology*. John Benjamins. to appear.

Maynard, D. G. (1997). "With". Master's thesis, University of Manchester, UK.

McClelland, J. and Rumelhart, D. (1981). An interactive activation model of context effect in letter perception: Part 1, an account of the basic findings. *Psychological Review*, 88(5):375–407.

McRoy, S. (1992). Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30.

Merriam-Webster (1999). Hypertext Webster Dictionary. http://www.fin.gov.nt.ca/webster.htm.

Meyer, I., Mackintosh, K., Barrière, C., and Morgan, T. (1999). Conceptual sampling for terminographical corpus analysis. In *Proc. of 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, pages 256–267, Innsbruck, Austria.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Five papers on WordNet. Technical Report CSL Report 43, Cognitive Science Laboratory. Princeton University.

Mima, H., Ananiadou, S., and Tsujii, J. (1999). A web-based integrated knowledge mining aid system using term-oriented natural language processing. In *Proc. of Natural Language Processing Pacific Rim Symposium (NLPRS)*, Beijing, China. 13-18.

Montemagni, S. (1995). *Subject and Object in Italian Sentence Processing*. PhD thesis, UMIST, Manchester, UK.

Montemagni, S. and Vanderwende, L. (1993). Structural patterns vs. string patterns for extracting semantic information from dictionaries. In Jensen, K., Heidhorn, G., and Richardson, S., editors, *Natural Language Processing: the PLNLP Approach*, chapter 12, pages 149–159. Kluwer Academic Publishers, Dordrecht.

Morel-Guilllemaz, A., Baud, R., and Scherrer, J. (1990). Proximity processing of medical texts. In *Proceedings of MIE 1990*, pages 625–630. Springer-Verlag.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Mountain Data Systems (1988). Acronym finder. http://www.AcronymFinder.com.

Nakagawa, H. and Mori, T. (1998). Nested collocation and compound noun for term extraction. In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 64–70, Montreal, Canada.

Nelson, S., Olson, N., Fuller, L., Tuttle, M., Cole, W., and Sherertz, D. (1995). Identifying concepts in medical knowledge. In *Proc. of 8th World Congress on Medical Informatics (MEDINFO)*, pages 33–36.

Ng, H. and Lee, H.-B. (1996). Integrating multiple knowledge sources to disambiguate word sense. In *Proc. of ACL '96*.

Ng, H. and Zelle, J. (1997). Corpus-based approaches to semantic interpretation in NLP. *AI Magazine*, 18(4):45–64.

NHS Information Authority (1999a). Coding and classification: Clinical terms (the READ codes). http://www.coding.nhsia.nhs.uk.

NHS Information Authority (1999b). Global standard for healthcare terminology. Unpublished report. Ref: 1999-IA-80.

NLM (1997). *UMLS Knowledge Sources*. National Library of Medicine, U.S. Dept. of Health and Human Services, 8th edition.

NLM (1998). *UMLS Knowledge Sources*. National Library of Medicine, U.S. Dept. of Health and Human Services, 9th edition.

Oakes, M. P. and Paice, C. D. (1998). Term extraction for automatic abstracting. In *Proc. of 1st Workshop on Computational Terminology, Computerm '98*, pages 91–95, Montreal, Canada. COLING/ACL.

Paice, C. D. and Jones, P. A. (1993). The identification of important concepts in highly structured technical papers. In *Proc. of ACM-SIGIR '93*, pages 69–77, Pittsburgh, PA.

Pearson, J. (1998). *Terms in Context*. John Benjamins, Amsterdam/Philadelphia.

Picht, H. and Draskau, J. (1985). *Terminology: an introduction*. University of Surrey, Guildford, UK.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Proctor, P., editor (1979). *The Longman Dictionary of Contemporary English*. Longman, Harlow.

Radford, I., Ananiadou, S., Arranz, M. V., and Tsujii, J. (1997). Term disambiguation by adding structural constraints to lexically based context matching techniques. In *Proc. of Natural Language Processing Pacific Rim Symposium (NLPRS)*, Phuket, Thailand.

Rayson, P. and Wilson, A. (1996). The ACAMRIT semantic tagging system: progress report. In Evett, W. and Rose, I., editors, *Proc. of AISB-96 Workshop on Language Engineering for Document Analysis and Recognition*, pages 13–20, Nottingham Trent University, England.

Reinert, M. (1986). Un logiciel d'analyse lexicale (Alceste). *Cahiers de l'Analyse des Donnees*, 4.

Resnik, P. (1995a). Disambiguating noun groupings with respect to WordNet senses. In *Proc. of 3rd Workshop on Very Large Corpora*. MIT.

Resnik, P. (1995b). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal, Canada.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proc. of ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What and How?"*, Washington, D.C.

Rey, A. (1979). *La terminologie: noms et notions*. Que sais-je? PUF.

Rigau, G., Atserias, J., and Agirre, E. (1997). Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proc. of ACL/EACL*, pages 48–55, Madrid, Spain.

Riloff, E. and Lehnert, W. (1992). Classifying texts using relevancy signatures. In *Proc. of AAAI*.

Riloff, E. and Lehnert, W. (1994). Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, pages 296–233.

Rocha, R., Rocha, B., and Huff, S. (1994). Automated translation between medical vocabularies using a frame-based interlingua. In *Proc. of SCAMC '94*, pages 690–694.

Rondeau, G. (1987). Introduction to terminology. In Sager, J. C., editor, *Readings in Terminology*. CCL, UMIST. (English version of Chapters 1-5, edited by J.C. Sager 1980).

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*. Erlbaum, Hilsdale, NJ.

Royauté, J. and Jacquemin, C. (1993). Indexation automatique et recherche des noms composés sous leurs différentes variations. In *Actes, Colloque Informatique et Langue Naturelle (ILIN)*, IRIN, Nantes, France.

Ruge, G. (1991). Experiments on linguistically-based term associations. In *Proc. of RIAO '91*, pages 528–545, Barcelona.

Sager, J. (1990). *A Practical Course in Terminology Processing*. John Benjamins.

Sager, J. (1994). *Language Engineering and Translation: consequences of automation*. John Benjamins.

Sager, J. C. (1997). Term formation. In Wright, S. and Budin, G., editors, *Handbook of Terminology Management*, volume 1: Basic Aspects of Terminology Management, chapter 1.2.1, pages 25–41. John Benjamins, Amsterdam.

Sager, N. (1972). Syntactic formatting of scientific information. In *Proc. of 1972 Fall Joint Computer Conference*, volume 41 of *AFIPS Conf. Proc.*, pages 791–800, Montvale, NJ.

Schulze, E. (1987). The term, its properties and nature and its delimitation from other types of lexemes. In Sager, J., editor, *Readings in Terminology*. CCL, UMIST. English version.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Smeaton, A. and Quigley, I. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proc. of 19th International Conference on Research and Development in Information Retrieval*, Zurich, Switzerland.

Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W. (1995). CRYSTAL: Inducing a conceptual dictionary. In *Proc. of the 14th International Joint Conference on Artificial Intelligence.*

Sparck Jones, K. (1985). Compound noun interpretation problems. In Fallside, F. and Woods, W., editors, *Computer Speech Processing*, pages 363–381. Prentice Hall International, London.

Sparck Jones, K. and Tait, J. (1984). Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.

Stairmand, M. A. and Black, W. J. (1996). Conceptual and contextual indexing of documents using WordNet-derived lexical chains. In *Proc. of 18th BCS-IRSG Annual Colloquium on Information Retrieval Research.*

Sumita, E. and Iida, H. (1991). Experiments and prospects of example-based machine translation. In *Proc. of 29th Annual Meeting of the Association for Computational Linguistics*, pages 185–192, Berkeley, California.

Sundheim, B., editor (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD. ARPA, Morgan Kaufmann.

Tapanainen, P. and Jarvinen, T. (1994). Syntactic analysis of natural language using linguistic rules and corpus-based patterns. In *Proc. of 15th International Conference on Computational Linguistics (COLING)*, volume 1, pages 629–634, Kyoto, Japan.

Tersmette, K., Scott, A., Moore, G., Matheson, N., and Miller, R. (1988). Barrier word method for detecting molecular biology multiple word terms. In Greenes, R., editor, *Proc. of the 12th Annual Symposium on Computer Applications in Medical Care (SCAMC)*, pages 207–211, New York. IEEE.

Tsutsumi, T. (1993). Natural language processing: the PLNLP approach. In Jenon, K., Heidhorn, G., and Richardson, S., editors, *Word Sense Disambiguation by Examples*, pages 263–272. Kluwer Academic Publishers, Dordrecht.

Uramoto, N. (1994). A best-match algorithm for broad-coverage example-based disambiguation. In *Proc. of 15th International Conference on Computational Linguistics*, volume 2, pages 717–721, Kyoto, Japan.

Ushioda, A. (1996). Hierarchical clustering of words. In *Proc. of 16th International Conference on Computational Linguistics (COLING)*, pages 1159–1162.

van der Eijk, P. (1993). Automating the acquisition of bilingual terminology. In *Proc. of 6th Conference of the European Chapter of the Associaiton for Computational Linguistics, EACL '93*, pages 113–119.

Vanderwende, L. (1995). Ambiguity in the acquisition of lexical information. In *Proc. of the AAAI 1995 Spring Symposium Series: Working notes of the symposium on representation and acquisition of lexical knowledge*, pages 174–179.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proc. of the DELOS workshop on Cross-Language Information Retrieval*, Zurich.

Vossen, P., Peters, W., and Diez-Orzas, P. (1997). The multilingual design of the EuroWordNet database. In Mahesh, K., editor, *Ontologies and multilingual NLP, Proceedings of workshop at IJCAI-97*, Nagoya, Japan.

Walker, D. E. and Amsler, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. In Grishman, R. and Kittredge, R., editors, *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, pages 69–83. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wilks, Y. (1997). Language processing and the thesaurus. Technical report, University of Sheffield, Department of Computer Science. CS-97-13.

Wilks, Y., Slator, B. M., and Guthrie, L. M. (1996). *Electric Words: Dictionaries, Computers and Meanings*. MIT Press.

Wilks, Y. and Stevenson, M. (1997). Combining independent knowledge sources for word sense disambiguation. In *Proceedings of Recent Advances in Natural Language ProcessingConference (RANLP)*, pages 1–7, Bulgaria.

Wilms, G. J. (1996). Using a hybrid system of corpus- and knowledge-based techniques to automate the induction of a lexical sublanguage grammar. In *Proc. of 16th International Conference on Computational Linguistics*, Copenhagen, Denmark.

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. of 14th International Conference on Computational Linguistics*, pages 454–460.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of 33rd Annual Meeting of the Association for Computational Linguistics, ACL'95*, pages 189–196.

Zhao, G. (1996). *Analogical Translator: Experience-Guided Transfer in Machine Translation.* PhD thesis, Dept. of Language Engineering, UMIST, Manchester, England.