MIND THE GAP: BRIDGING FROM TEXT TO ONTOLOGICAL KNOWLEDGE

CHRISTOPHER ARTHUR BREWSTER

Thesis submitted in fulfilment of the requirements for the degree of Doctorate in Philosophy
Supervised by Professor Yorick Wilks
Department of Computer Science

University of Sheffield

July 2008
This thesis is dedicated to my father, Harry Brewster (1909-1999),
whose many journeys have led me here and will lead me beyond.
The objective of this thesis is to identify new ways to build ontologies from textual corpora automatically. A basic challenge we identify is that the ontology to a large extent is absent from set of domain specific texts because it forms part of the background knowledge the reader brings to the act of reading. The knowledge taken for granted by the writer is rarely made explicit, and thus it is computationally inaccessible to a significant degree.

We argue that the objectives of ontology learning need to be revised. It is a task which should be made both weaker than current assumptions (which appear to promise to derive fully specified formal ontologies using NLP techniques), and more sophisticated (in that we view ontology learning as a far more complex process than previously suggested). We propose to make the task weaker by acknowledging that perfect knowledge is impossible in many cases. We suggest that there are degrees of precision in a piece of knowledge and degrees of confidence. We also view the task of ontology learning from texts as more sophisticated by proposing mechanisms whereby the knowledge missing from a given domain corpus can be identified and the specific piece of knowledge needed can be sought, obtained and verified.

We present the Abraxas model as a model of what ontology learning systems should be like and the Abraxas system as a specific instantiation of this model. Abraxas is an uncertainty-based, dynamic, data-driven model of ontology learning. A key aspect of the Abraxas model is the concept of equilibrium between three resources (the corpus of documents, the set of extraction patterns, and the ontology or set of knowledge triples). Changes in this equilibrium result in a learning process which may add more knowledge triples, or add further documents or learn further extraction patterns. The model is probabilistic in that each resource item (knowledge triple, document, extraction pattern) has an associated confidence level which continually changes as the system iterates. The system is dynamic in its ability to change its confidence level in any item over time or further iterations. It is a data-driven model because the confidence the system has in each resource depends on the its relationship to the other existing resources. So, for example, evidence from the corpora and the extraction patterns provide confidence in any given knowledge triple.
ACKNOWLEDGMENTS

There are a great many people who have helped me in completing this thesis. First and foremost I wish to thank Professor Yorick Wilks, whose support, inspiration, suggestions and confidence have been fundamental to my efforts. Yorick has provided an immensely stimulating and exciting environment in which to study and undertake research. He has been a great source of inspiration and outstanding advice throughout the thesis writing process. Conversations with him have led to many of the major points in this thesis. Above all he has repeatedly shown confidence in me when I flagged and doubted my own capacity. I am very grateful to have had him as my supervisor and to have been able to work in his research group at Sheffield the past few years.

I would like to thank the other members of my committee. Fabio Ciravegna has been very supportive and has on a number of occasions helped me move forward when I needed a helping hand. Louise Guthrie has repeatedly posed questions which provided food for thought.

My colleague and friend José Iria has helped through long conversations turn instinctive half-baked ideas into reality. He has been instrumental together with Ziqi Zhang in turning the ideas presented here into the fully functioning system, Abraxas. In addition, I wish to thank all my colleagues in the Natural Language Processing Group at the University of Sheffield.

At a more personal level, I would like to thank Joanne Luciano for inspiring my efforts to finish. She played a key role in making parts of the text substantially more concrete than they would have been without her questions and suggestions. Most importantly she ‘snuck’ out of me the final pieces of text and for which I am extremely grateful. Another person I wish to thank does not wish to be named, so I hope she/he will take this sentence as an expression of my gratitude.

I would like to thank Lynne Watson for supporting me through difficult times, challenging me with existential questions and helping to give birth to the ideas presented in this text. Furthermore, I am immensely grateful for my children, Christos and Maria, who have taught me innumerable lessons of life and provided great joy as well.

Finally, I want to mention my parents. My father set me out on this journey which finds its completion in this thesis but who unfortunately is not with us today to witness the event. My mother I am grateful for just for being herself.
## CONTENTS

1 MIND THE GAP  

1 Introduction 3  
1.1 Structure of the Thesis 5  
1.2 Note to the Reader 6  

2 Knowledge and Its Representation 9  
2.1 The Nature of Knowledge 9  
2.1.1 Philosophical Perspectives 9  
2.1.2 The Knowledge Management View 13  
2.1.3 The AI View 16  
2.2 The Discovery and Transmission of Knowledge 18  
2.3 Forms of Knowledge Representation 23  
2.4 Knowledge Acquisition 27  

3 Thesauri, Taxonomies and Ontologies 31  
3.1 Historical Background 31  
3.1.1 Aristotle 32  
3.2 Thesauri 33  
3.2.1 Wilkins 33  
3.2.2 Roget 34  
3.2.3 Thesauri in Natural Language Processing 36  
3.3 Taxonomies 37  
3.4 Ontologies 38  
3.4.1 What is an Ontology? 39  
3.4.2 Knowledge Representation in Ontologies 40  
3.4.3 Types of Ontologies 41  
3.4.4 Uses of Ontologies 43  
3.5 Capturing the World 45  

4 The Commercial Context 47  
4.1 Purpose 47  
4.2 Current Practice 48  
4.2.1 Building the Taxonomy/Ontology 48  
4.2.2 Maintaining the Taxonomy/Ontology 49  
4.2.3 Associating Content to the Taxonomy/Ontology 49  
4.2.4 Confidence and Evaluation 50  
4.3 Current Products 51  

5 Anatomy of Building an Ontology 55  
5.1 Corpus Selection 57  
5.2 Associating Terms 57  
5.3 Building a Hierarchy 58  
5.4 Labelling Relations 58  
5.5 Methodological Criteria 58  
5.5.1 Coherence 59  
5.5.2 Multiplicity/ Multiple Inheritance 60  
5.5.3 Ease of Computation 61
6 Existing Approaches 63
   6.1 Associating Terms 63
      6.1.1 Simple Word Association 63
      6.1.2 Quasi-thesauri 71
   6.2 Building a Hierarchy 78
      6.2.1 Pathfinder Networks 78
      6.2.2 Document Subsumption 82
   6.3 Labelling Relations 83
      6.3.1 Synonymy or Substitutability 83
      6.3.2 Antonymy 85
      6.3.3 Hyponymy 85
   6.4 Other Systems and Approaches 87
      6.4.1 CAIULA 87
      6.4.2 ASIUM 89
      6.4.3 Ontolearn 90
      6.4.4 KnowItAll 91
      6.4.5 Heterogeneous Evidence Approach 93
   6.5 From Syntax to Semantics 94
   6.6 Conclusion 95

7 Explicit Knowledge and Lexico-Syntactic patterns 97
   7.1 Explicit Knowledge in Text 97
   7.2 A Provisional Catalogue of some Lexico-Syntactic Patterns 99
      7.2.1 Patterns for identifying hyponymy or the ISA relation 100
      7.2.2 Patterns for identifying SIBLINGS 103
      7.2.3 Patterns for identifying ATTRIBUTES 103
   7.3 The Reliability of Lexico-Syntactic Patterns 104
   7.4 Extraction Patterns 108
      7.4.1 Learning of Extraction Patterns 110
   7.5 Conclusions 114

8 The Role of Background Knowledge 115
   8.1 Ontologies and Text 115
      8.1.1 Ontologies and Background Knowledge 115
      8.1.2 Text and Knowledge Maintenance 117
   8.2 The Assumed Knowledge Hypothesis 119
      8.2.1 Experimental Evidence 120
   8.3 Sources of Ontological Knowledge 123
   8.4 Conclusions 125

9 Knowledge and the Textual Landscape 127
   9.1 Language Variety: Synchronic and Diachronic Variation 129
   9.2 The Synchronic Dimension 130
      9.2.1 Genre or language for a purpose 130
      9.2.2 Categories of Text 131
   9.3 The Diachronic Dimension 134
      9.3.1 The Borrowing or Migration of Terms 135
9.3.2 New Terminology 137
9.4 Conclusions 138
10 A Model of Knowledge Acquisition 139
  10.1 The Abraxas Approach 139
  10.2 A Tripartite Model 140
    10.2.1 The Corpus 142
    10.2.2 The Ontology 143
    10.2.3 The Extraction Patterns 145
    10.2.4 The Fourth Dimension: The System User 146
  10.3 The Basic Method 148
  10.4 Metrics for Measuring In-equilibrium 152
    10.4.1 The Resource Confidence Measure 152
    10.4.2 The Knowledge Gap Measure 160
  10.5 A Worked Example 161
    10.5.1 Calculating the measures 167
  10.6 Conclusions 178
  10.7 Glossary of terms 179
11 Ontology Evaluation and Selection 181
  11.1 Introduction 181
  11.2 Types of Evaluation and the Evaluation of Knowledge 182
  11.3 The Evaluation of a Representation of Knowledge 185
  11.4 Approaches to Ontology Evaluation 188
    11.4.1 Structural and Methodological Evaluation 188
    11.4.2 Evaluation in Use 192
    11.4.3 Data-driven Ontology Evaluation 194
    11.4.4 Gold Standard Ontology Evaluation 197
  11.5 Ontology Ranking 200
    11.5.1 AKTiveRank 201
    11.5.2 The Ranking Measures 202
    11.5.3 Experiment 206
    11.5.4 Ontology Ranking: Conclusions 210
  11.6 Evaluation of Ontology Learning Systems 211
  11.7 Evaluating Abraxas 213
    11.7.1 Abraxas as a Metamodel 213
    11.7.2 Abraxas as a Running System 216
  11.8 Conclusion 220
12 Conclusion and Future Work 221
  12.1 Conclusions 221
  12.2 Future Work 224

II APPENDIX 229
A Appendix A - Sentential Context Experiment 231
B Appendix B - Knowledge Gap and Explicit Knowledge Gap Formulas 235
  B.1 Knowledge Gap 235
  B.2 Explicit Knowledge Gap 237
C Appendix C - The Abraxas Animals Ontology 239

BIBLIOGRAPHY 253
LIST OF FIGURES

Figure 1 Categories of Knowledge [Nonaka and Takeuchi, 1995]  14
Figure 2 The taxonomy in Open Directory under the heading food. Numbers indicate the number of web pages to which the user can navigate.  21
Figure 3 A Semantic Network, from Woods (1975)  40
Figure 4  45
Figure 5 A fragment of an abstract ontology/taxonomy  57
Figure 6 subtree from Brown et al. (1992:474)  73
Figure 7 Detail of many noun-like words (McMahon and Smith 1996:226)  76
Figure 8 From McDonald et al. [1990]  80
Figure 9 PFNET derived from term co-occurrence in PNAS, from White et al. [2004]  81
Figure 10 From Sanderson and Croft (1999:4)  83
Figure 11 precision and recall of KnowItAll [Etzioni et al., 2004]  93
Figure 12 The result of applying the Level 0 Extraction Pattern (Term Recognition)  109
Figure 13 The result of applying the Level 1 Extraction Pattern (Term Association)  110
Figure 14 The result of applying the Level 2 Extraction Patterns (Ontological Relation Recognition)  111
Figure 15 Basic design of SnowBall.  112
Figure 16 Example from Thor et al. [1997]  118
Figure 17 Example from Schwartz and Christiansen [1997]  118
Figure 18 Definitions for Enzyme from the web.  124
Figure 19 Definition from the web.  125
Figure 20 Categories used by the BNC (genre is not clearly defined).  132
Figure 21 The Knowledge Acquisition Balance: Seeking Equilibrium  141
Figure 22 The central core resources surrounded by layers of diminishing resource confidence.  143
Figure 23 An abstract view of the Abraxas system.  148
Figure 24 The conceptual space over which the RC measure is calculated.  156
Figure 25 The corresponding space expressing in usual NLP terms. A list of measures used in NLP, clinical research and other research domains is provided as well.  157
Figure 26 The output of ontology normalisation  167
Figure 27 Assessment of people vs. assessment of ontologies  184
Figure 28  AKTiveRank Architecture  201
Figure 29  student + university.  208
Figure 30  Ranks based on CMM.  209
Figure 31  Ranks based on DEM  209
Figure 32  Ranks based on SSM  209
Figure 33  Ranks based on BEM  210
Figure 34  Evaluation measures (LP, LR, TP, etc.) plotted against the sequentially produced ontologies from the iterative process, for experiment 1.  218
Figure 35  Comparison of final evaluation measures for the different seed cases (Experiments 1-4).  219
Figure 36  Part of the ‘Animal’ ontology visualised using OWLviz  239
Figure 37  Another part of the ‘Animal’ ontology visualised using OWLviz  239

LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>KM definitions of Knowledge [- adapted from Acker, 2003]</td>
</tr>
<tr>
<td>Table 2</td>
<td>Wilkins’ classification of ‘stone’.</td>
</tr>
<tr>
<td>Table 3</td>
<td>Roget’s Plan of Classification.</td>
</tr>
<tr>
<td>Table 4</td>
<td>Associated Words for cheese using Encyclopaedia Britannica texts</td>
</tr>
<tr>
<td>Table 5</td>
<td>Associated Words for cheese using random web texts</td>
</tr>
<tr>
<td>Table 6</td>
<td>From Grefenstette 1994: 51</td>
</tr>
<tr>
<td>Table 7</td>
<td>From Grefenstette 1994: 67</td>
</tr>
<tr>
<td>Table 8</td>
<td>“Semantically sticky” clusters (Brown et al. 1992:478)</td>
</tr>
<tr>
<td>Table 9</td>
<td>Number of n-grams with various frequencies in 365,893,263 words of running text. Numbers absolute counts. (Brown et al. 1992:470)</td>
</tr>
<tr>
<td>Table 10</td>
<td>Classes from a 260,741-word vocabulary (from Brown et al. 1992:475)</td>
</tr>
<tr>
<td>Table 11</td>
<td>Some other sample classes</td>
</tr>
<tr>
<td>Table 12</td>
<td>Relatedness functions from McDonald et al. [1990]</td>
</tr>
<tr>
<td>Table 13</td>
<td>Words most strongly related to bank from McDonald et al. [1990]</td>
</tr>
<tr>
<td>Table 14</td>
<td>Substitutable terms</td>
</tr>
<tr>
<td>Table 15</td>
<td>Extract from a table of possible substitutes [Church et al., 1994]</td>
</tr>
<tr>
<td>Table 16</td>
<td>Example classes from the Basili et al. approach</td>
</tr>
<tr>
<td>Table 17</td>
<td>Table showing hit counts for various instantiated lexi-co-syntactic patterns. From Yahoo, June 2005</td>
</tr>
</tbody>
</table>
Table 18  Table ranking the precision and probability of false positives for lexico-syntactic patterns, based on the data presented in Table 17.  106
Table 19  Examples of instantiated lexico-syntactic patterns used as queries on Yahoo.  106
Table 20  Precision of lexico-syntactic phrases based on random sample.  107
Table 21  Recall for the lexico-syntactic patterns based on BNC/WordNet based Gold Standard.  108
Table 22  Summary Results  121
Table 23  Frequency Distribution of the Terms Selected.  122
Table 24  List of text categories used in the Brown and LOB corpora  131
Table 25  The 27 different starting states the Abraxas system could be in.  147
Table 26  Document d1 - part of the entry on Hedgehog from Wikipedia.  162
Table 27  Top keywords from the hedgehog text above. Number reflect relative ‘keyness’ to the BNC.  163
Table 28  Instantiated extraction patterns.  164
Table 29  Two instantiated extraction patterns with corresponding Yahoo document counts.  165
Table 30  Document d2, from the Tehran Museum website.  165
Table 31  The calculation of the KG measure  168
Table 32  The calculation of the EKG measure  169
Table 33  Calculating the RC of d2  171
Table 34  Resource confidence for the items in the different resource sets, iterations 1-3.  172
Table 35  Calculating the RC of t2  175
Table 36  Calculating the RC of p10 cf. Table 34  177
Table 37  Ontology Learning Errors (adapted from Porzel and Malaka [2005])  194
Table 38  Vector comparison of five ontologies  196
Table 39  Order of search result for “student university” as returned by Swoogle. Duplicates were removed and are indicated by a dash.  206
Table 40  AKTiveRank results  207
Table 41  Key findings concerning knowledge and text.  215
Table 42  Extraction patterns used: NP = noun phrase, sg = singular, pl = plural.  217
Table 43  Results obtained for experiment 1.  217
Table 44  Four hypothetical cases showing the calculation of the Knowledge Gap measure  236
Table 45  Four hypothetical cases showing the calculation of the Explicit Knowledge Gap measure  238
Part I

MIND THE GAP
INTRODUCTION

Πάντα ῥεῖ καὶ οὐδὲν μένει.
Everything flows, nothing stands still.
Heracleitus, according to Plato - Cratylus 401d

One of the key issues in Artificial Intelligence since its inception has been the extent to which computers can have internal representations of the world. If they can, and assuming that internal representations are sufficient, then it should be possible for AI applications to function successfully and interact with the world. If they cannot, then the central AI agenda is in need of severe revision. Assuming the former is true, where can one obtain such a representation of the world, a representation which is not just of a toy world such as SHRDLU [Winograd, 1972]. There is a long history in AI of efforts to develop representations of the world but, inevitably, as these have all been built by hand, they have been limited in their scope and coverage. The tradition has been to build ‘knowledge bases’ using the information provided from the analysis of structured interview, by means of so called ‘protocol analysis’, a process which is both time consuming and costly [Ericsson and Simon, 1996]. In the last decade or so, researchers have stopped speaking of knowledge bases but instead prefer the term ‘ontologies’, thereby possibly laying claim to greater philosophical rigour. With the rise of the Semantic Web [Berners-Lee et al., 2001, Fensel et al., 2003], ontologies have become even more fashionable if no more rigourous. In the Semantic Web view of the world, there will be a multiplicity of ontologies, which will make it possible for agents to communicate with each other to sort out scheduling conflicts and whether the milk in the fridge has gone off.

An ontology, at the simplest level, is a set of concepts concerning some domain of human knowledge and a specification of the relationships between those concepts. It is a structured representation of the world. Gruber famously defined an ontology as “a formal, explicit specification of a shared conceptualisation” [Gruber, 1993]. The word ontology will be used in this research to cover a whole range of structures claimed to encode knowledge, ranging from very informal thesauri to highly logically structured ontologies (cf. below Sections 3.2-3.4 for a more detailed discussion).

Apart from the AI community mentioned above, especially as expressed in the Semantic Web, there are a number of other research communities which have an interest in ontologies. One is that of Knowledge Management [Davenport, 1998, Prusak, 2001], especially those researchers interested in supporting the use of computers to facilitate corporate or institutional knowledge management [Dieng et al., 1999, Hendricks and Vriens, 1999]. The fact that an ontology is
supposed to encode shared conceptualisations and therefore shared knowledge is one reason why the Business world and the Knowledge Management community have had a growing interest in these ideas [Maedche and Staab, 2002]. With the decline of manufacturing and the rise of hi-tech companies, less and less company value is attributable to bricks and mortar, and more and more to the quality of a company’s workforce. As the value of a company has come to rest so much in what lies in the minds of its employees, so has it become imperative to find means to distil that ‘know-how’ and store it in some form which cannot just walk out. This is the holy grail of Knowledge Management, and it is in this context that ‘ontologies’ have been seen as a means of representing, of ‘capturing’ knowledge.

In the world of lexicography, an ultimate objective has been to construct dictionary entries directly from raw data i.e. large collections of text. Ideally, given sufficient exemplars of a word, a computer system would be able to construct an accurate representation of its usage, meaning etc. in a manner which would actually be useful for a human or machine user. A less ambitious version of this challenge might be to organise the words in a language into a thesaurus which would represent the significant semantic relationships between words or terms.

These three disparate communities share a number of features in common, both in their objectives and the means at their disposal. The cost of manually constructing knowledge bases has made the AI community desperately in need of other sources of ‘knowledge’. Similarly, in the commercial world of Knowledge Management, it is extremely costly to lose an expert during the construction of a knowledge base. More importantly perhaps, the sheer quantity of data such tasks require is such that manual techniques are becoming increasingly impracticable. In addition, ‘knowledge’ is changing so rapidly that it is extremely difficult to keep any knowledge base up to date. Thus, Knowledge Management faces the old problem of lexicographers, that the dictionary was out of date as soon as it was published. It is the revolution in lexicography, in turning to large text corpora [Sinclair, 1987], which has shown us the way forward. What the AI, Knowledge Management and lexicographic communities have in common is access to large quantities of unstructured and semi-structured texts from a number of sources. In AI, it is usually possible to obtain a substantial corpus of texts concerning the domain that one is trying to build an application for. Companies have vast repositories of reports, manuals, emails and letters, all of which reflect in some manner the knowledge and competencies of the author-employees. It follows naturally that one would want to find a means to analyse these texts and from them construct a representation of the knowledge contained in them.

Two major research questions arise. The first is: Is it possible to construct an ontology (in some form) from a collection of texts? Tackling this question places this thesis clearly within the sub-field of Artificial Intelligence called Natural Language Processing (NLP), and the
discussion which follows is largely written from the perspective and using the methodologies of this field. The second question is: Does any ontology so constructed constitute an adequate representation of the knowledge available to a human reader of those texts, i.e. can an ontology based on texts answer similar questions to a human being who has read those texts? The research presented here concentrates largely on the first of these. Working within the field of NLP, our intention has been, first, to consider the range of computational techniques which can contribute to this enterprise and secondly, to push the boundaries forward by proposing new methods, so as to gain a deeper understanding of the relationship between text and knowledge.

Our fundamental assumption has been that it must be possible to learn ontologies from texts. Since an ontology represents our knowledge of a domain, and since human beings most of the time successfully communicate their knowledge through the medium of texts, we have assumed, together with many other practitioners in the field, that it should be possible to automate the construction of ontologies by processing large quantities of texts. In practice, we have had significantly to revise our understanding of what knowledge lies in texts as well as both the ambition and scope of any attempt to learn ontologies.

1.1 STRUCTURE OF THE THESIS

This thesis is structured as follows: In Chapter 2, we discuss the nature of knowledge in general and how it has been perceived by philosophers, researchers in AI and practitioners of Knowledge Management. We note that a large proportion of what people consider to be knowledge cannot be represented in the knowledge representation structures familiar to AI. In Chapter 3, we present a brief historical account of some important intellectual ancestry to the current work on ontologies. Aristotle, John Wilkins and Peter Roget all form part of the foundations on which we are now building. We also discuss in greater detail the difference between thesauri, taxonomies and ontologies and support the view that they are on a continuum of logical rigour.

Chapter 4 presents an account of the needs of commercial enterprises in terms of ontologies for institutional or corporate use. We discuss the objectives and common practices as well as survey some of the commercial offerings which claim to provide taxonomies or similar structures for industry. In Chapter 5, we consider the process of ontology construction in detail, especially construction of ontologies using texts as the main data source. We present a set of methodological criteria for ontology learning which concern the type of data, the type of output and other facets of any chosen methodology. In Chapter 6, we survey a range of methods in Natural Language Processing which can be used to perform some of the steps identified in the previous chapter. At the end of this chapter we describe some of the complete systems in the literature which attempt to perform ontology learning from texts. It is clear from the work described in this chapter that ontology learning from texts is a major challenge and can only be
partially successful.

In Chapter 7, we consider the way knowledge is explicitly expressed in texts and how this can be captured computationally. A number of lexico-syntactic phrases or extraction patterns are discussed and their accuracy is evaluated. Chapter 8 makes the point that in fact a large amount of the knowledge we want to capture in an ontology is in fact assumed by the writer and forms part of the ‘background knowledge’ that the reader is expected to bring to the act of reading. Consequently, the explicit knowledge is not there! We discuss some initial approaches to how to circumvent this technical challenge in the later part of this chapter and in Chapter 9 present further arguments that specific types of text (certain genres) and certain textual circumstances (historical first use or borrowing into a new domain) will result in writers being explicit in the definition of terms and thus enable the application of extraction patterns. The problem lies in the automatic recognition of these contexts and lack of appropriate diachronic resources.

In the light of the discussion in the previous chapters, in Chapter 10 we present Abraxas, our model for ontology learning. It is a model which takes into account the varying degrees of evidence for different instances of ontological knowledge, allows for continuous and dynamic change in the ontology and provides for a confidence level to be ascribed to each piece of knowledge. It is based on the notion that the system seeks an equilibrium between ontological knowledge triples, extraction patterns and the corpus which provides evidence. These terms will be defined at an appropriate stage. Imbalances in the equilibrium trigger the learning of more knowledge, the extension of the corpus and potentially the acquisition of further extraction patterns. Chapter 11 discusses the difficulties and challenges in evaluating knowledge representations like ontologies. We consider a number of approaches to ontology evaluation and also describe out work on ranking ontologies with respect to each other. We present a qualitative evaluation of the Abraxas approach. The Conclusion rounds off the thesis by re-capping on the lessons learned and identifying a number of avenues for future research.

1.2 NOTE TO THE READER

The reader should bear in mind that this text is both philosophical in nature and empirical. It is philosophical in that at the core of the enterprise of knowledge representation lies a fundamental unknown: we are unable to define or delimit human knowledge in a formal or rigorous manner which allows for its full representation in computers. This has been traditionally interpreted as a problem which needs to be overcome. Our position is empirical in that we have been led by the observation of facts both concerning the phenomenon of language and the limitations of NLP methods and techniques. Our position is to view these challenges as a feature of language and knowledge, and not as a bug. It is the vagueness, the ambiguity, the imprecision of both knowledge and its expression in human language which
gives it its power. It is because of these features that we can handle metaphorical extensions and terminological change in meaning so effectively. It is because of this that we can make intellectual discoveries and communicate them. Otherwise Shannon would have been unable to take the thermodynamic concept of entropy\(^1\) and transform it into a statistical concept in information theory.

Consequently our perspective has been that it is our models that have to change at a more fundamental level, rather than wrestling with language in the hope of making a round object fit a square hole. Our hope has been that we have made a small contribution to a fundamental re-think in AI and NLP as to the nature of knowledge and the relationship between knowledge, language and text.

\(^1\) Cf. footnote on page 138
The great end of life is not knowledge but action.
Thomas Henry Huxley [1877]

The message is that there are no “knowns”. There are things we know that we know. There are known unknowns. That is to say there are things that we now know we don’t know. But there are also unknown unknowns. There are things we don’t know we don’t know.
Donald Rumsfeld [2002]

Ontologies are widely perceived to play a central role both in Artificial Intelligence and Knowledge Management. Ontologies are considered to be representations of knowledge [Lassila and McGuinness, 2001, McGuiness, 2003], and they are the outcome of a process of knowledge capture which transfers knowledge from human beings to a more or less formal representation. They have multiple purposes and applications but in the abstract they are an attempt to facilitate the communication between people and computer systems, between different computer systems, and (often) between different people [Gruber, 1993, Uschold and Gruninger, 1996]. This emphasis on the need to facilitate communication is often expressed in terms of “shared understanding” or “shared conceptualisation” because that ‘conceptualisation’ reflects a specific world view shared by a particular community, and consequently reflects a particular way of expressing or communicating knowledge.

If ontologies are the technical structures to be used to represent knowledge, then an effort must be made to understand what it is that these structures represent. Only by understanding what it is that they are supposed to represent can we understand how to construct, use and maintain them.

2.1 THE NATURE OF KNOWLEDGE

2.1.1 Philosophical Perspectives

Historically it has been philosophers who have analysed the notion of knowledge in the greatest detail. However, the dominant concern of conventional epistemology (and by extension all research influenced by that paradigm) has been propositional knowledge, otherwise referred
to as ‘knowing that’\(^1\) or ‘factual knowledge.’ The traditional view of
knowledge is that it is “justified true belief”, a view we owe to Plato’s
Theaetetus. According to this analysis, there are three individually
necessary conditions which are jointly sufficient i.e. justification, truth
and belief. Thus for someone to know that “Pierre is French” then first
Pierre must indeed be French (truth), the individual must believe this
to be the case (belief) and there must be some evidential support or
deductive argument to justify the belief. This was the accepted view
until the twentieth century when philosophers like Russell [1948] and
Gettier [1963] showed that, as the justification could be based on a
false premise, we needed to reconsider our understanding of what
it was to know. Russell, for example, hypothesised a situation where
where someone notes the time from a supposedly reliable clock which
however had stopped exactly 24 hours previously [Russell, 1948, :170].
In such a case, the belief is justified and true but the person cannot
be said to ‘know’ the right time. Most responses attempted to make
the process of justification more rigorous (e.g. by hypothesising ‘unde-
feated’ or ‘conclusive’ justification cf. Bernecker and Dretske
2000, :3-6) thereby limiting even further what is considered ‘knowledge’ proper.
Much philosophical ink has been spilled discussing this problem, but
no one has come up with an alternative that has gained much accep-
tance [O’Hara, 2002]. The view that knowledge is propositional has
been the dominant view in AI and underpins the use of ontologies and
forms part of the widely accepted ‘physical system symbol hypothesis’
of Newell and Simon [1976]\(^2\).

Philosophers have recognised other forms of knowledge, but largely
to dismiss them because propositional knowledge “is fundamental to
human cognition and required both for theoretical speculation and
practical judgement” [Bernecker and Dretske, 2000, :3]. The other forms
that were recognised were ‘knowing how’ which concerns practical or
skill-based knowledge, and ‘know who/what/where’ which concerns
knowledge by acquaintance, but philosophical analyses of these topics
appear to have had little impact on the course of AI. The exception
to this has been goal-based languages such as PLANNER and Prolog where “knowing that” became a matter of “knowing how to prove”
Hewitt [1969], Colmerauer and Roussel [1993].

Knowledge ‘of someone’ or ‘of something’ is often quoted as the
prototypical case of ‘knowledge by acquaintance’, a theory of knowl-
dge put forward by Bertrand Russell [1912]. This approach is also
described as knowledge concerning an object which is neither true nor
false in contrast with knowledge of propositions which bear a truth
value. Knowledge by acquaintance is supposed to provide a founda-
tion for propositional knowledge but some philosophers cannot accept

\(^1\) In contrast to “knowing how” (non-propositional knowledge) cf. below and [Ryle,
1949].
\(^2\) An example of this sort of attempt in AI is that of Cyc. Lenat, the creator of Cyc, states
that “One of the principle[sic] tenets of AI, is that one can explicitly, declaratively
represent a body of knowledge, such that it can be used by a number of programs”
[Lenat et al., 1994].
acquaintance as an indefinable and atomic concept [Fumerton, 1998].

‘Knowing how’ was brought to the fore by the work of Gilbert Ryle [1949] working under the influence of Ludwig Wittgenstein [1953]. Ryle believed that much of what is usually characterised as thought is in fact the manifestation of skill, and that theoretical or propositional thinking is of secondary significance. Thus, for Ryle, knowledge is primarily the capacity to know how to do things, including knowing how to talk. Ryle’s views have been vigorously attacked by many philosophers as being a form of philosophical behaviourism. Dretske has criticised the distinction between ‘know how’ and ‘know that’ arguing that ‘know how’ involves more than just a technical or physical knowledge. It also involves knowing when to do something, what to do to obtain a certain effect and the implication of this is that ‘know how’ is closely bound up with propositional knowledge [Dretske, 1999].

This contrast is closely related to two other philosophical distinctions which are of interest to us. First, there is the contrast between what philosophers call explicit and implicit knowledge. Explicit knowledge is usually taken to be propositional knowledge or knowledge about which the knower can make a verbal statement: “Pierre is French.” Any other form of knowledge is implicit. However, this leaves a rather big field for implicit knowledge. According to Dummett [1991], for example, explicit knowledge includes knowledge which can be derived by inference from the propositions an individual already knows, even before the inferences have been made. All that is needed is that it is potentially possible to express it verbally. Thus, in Dummett’s case, the fact that $235 + 147 = 382$ is explicit even though we use the principles of arithmetic to work it out and it is not a stored fact. In contrast, Dennett states that “for information to be represented implicitly, we shall mean that it is implied logically by something that is stored explicitly” [Dennett, 1983]. Thus in this case, $235 + 147 = 382$ is an implicit fact and not an explicit one. As Davies notes, for any given definition of explicit knowledge there exists a whole range of possible types of implicit knowledge depending on the resources allowed to draw out the inferences from the propositionally stored explicit knowledge [Davies, 2001].

The other important contrast with explicit knowledge is tacit knowledge, a concept which has played an important role in philosophy, psychology and knowledge management, and which derives from the work of Michael Polanyi. Assuming that explicit knowledge is that which can be formulated in words, tacit knowledge is knowledge that has not been and perhaps cannot be formulated explicitly in words. Polanyi was of the opinion that “we can know more than we can tell” [1966, 4]. Influenced by Gestalt psychology, he made a distinction between subsidiary and focal awareness. Thus an individual could be perceptually aware of the whole while basing this on a subsidiary awareness of parts. Knowledge consists of a continuous process of integration from the subsidiary clues to a focal whole. Tacit knowledge was not another kind of knowledge because all knowledge was either
tacit or grounded on tacit knowledge\textsuperscript{3}. Furthermore knowledge is both a social construct in that it is dependent on the accepted background of social practice, and also very personal in that it is dependent on our personal experience. Part of our tacit knowledge is learnt through the process of socialisation, and it is something we are unaware of. One consequence of Polanyi’s views is that knowledge can never be made wholly explicit, and that the objectivist ideal of detached, impersonal knowledge cannot be achieved.

\textsuperscript{3} The concept of tacit knowledge became very important in linguistics where Chomsky [1965] considered knowledge of language was a form of tacit knowledge since speakers have no grasp of the complex linguistic rules which underlie their capacity to understand an infinite number of sentences. However, Chomsky’s understanding seems to be that linguistic tacit knowledge is in fact propositional because a speaker’s grammar can interact by inference with other systems of knowledge and belief [Chomsky, 1980, 1986]. Consequently Chomsky seems to view this type of tacit knowledge as similar to Dennett’s ‘implicit’ knowledge.
2.1 The Knowledge Management View

A known mistake is better than an unknown truth.
Arabic proverb

Real knowledge is to know the extent of one’s ignorance
Confucius

A somewhat different approach to knowledge can be found in the field of Knowledge Management. I review these approaches briefly because ontologies and taxonomies are viewed as playing a key role in knowledge management.

Knowledge Management arose as a discipline in the early 1990’s as a response to a number of changes in the worlds of business and technology [Prusak, 2001]. One factor was the emergence of globalisation, where both the volume and speed of business grew enormously, and where in order to remain competitive on a world scale it became essential for companies to identify what they know, who knows it, and what do they not know. Another factor was the ubiquitous presence of computers which allowed information to be easily moved, to any place, and at almost no cost. This resulted in an increased awareness of “cognitive skills still unreplicable by silicon” and made commercial enterprises value these knowledge intensive skills ever more (ibid.). Another important factor has been the growth in the so-called ‘knowledge centric’ view of the firm. Winter argued that “firms are organisations that know how to do things. . . . Firms perform their function[s] as repositories of knowledge” [Winter, 1988, :175-77]. It is widely accepted that many companies’ assets lie more in the knowledge present in the minds of its employees rather than in externally tangible objects. All these factors have made knowledge management of great importance in the commercial world.

Clearly under the influence of the philosophical tradition, knowledge is commonly analysed as consisting of three types: explicit, implicit and tacit, as in this example, reflecting common consensus:

- Explicit, or codified information represented in data and text. (Word, Web, PDF documents)
- Implicit, or uncodified but explicable norms and experiences. (undocumented/non-captured know-how)
- Tacit, or unarticulated and inexplicable abilities and intuitions. (musician, poet, intuition, genius)

Table 1. KM definitions of Knowledge [- adapted from Acker, 2003]
Implicit knowledge in this view accords with Dennett’s cited above, while ‘tacit’ knowledge appears to be defined to reflect Polanyi’s perspective. However, Stenmark [2002] has noted that the use of the term ‘tacit knowledge’ in Knowledge Management is largely due to Nonaka and Takeuchi [1995]. Although Nonaka cited Polanyi, essentially he used the term to refer to knowledge that is difficult or impossible to express, which is a much narrower definition than Polanyi’s.

Nonaka presents a highly influential model, categorising types of knowledge and the nature of their interaction cf. Table 1 [Nonaka, 1994, Nonaka and Takeuchi, 1995]. The purpose of this model is not only to show the types of knowledge but also how knowledge is dynamically created (the focus of Nonaka is corporations and other organisations). Explicit knowledge is contrasted with tacit knowledge and defined as in Table 1 (implicit knowledge is not discussed). The process of socialisation involves the acquisition of tacit knowledge through social contact such as apprenticeship or exposure to a community of practice, by means of imitation, observation and practice – a process that is possible without the use of language (according to Nonaka). Externalisation is the process of transforming tacit knowledge into explicit knowledge, often through the use of metaphors due to the difficulty of expressing tacit knowledge in language. Nonaka emphasises imagination and intuitive learning through symbols [1994, :21].

The process of internalisation corresponds most closely to traditional learning and involves the transformation of explicit knowledge into internalised tacit knowledge. Finally combination concerns the aggregation of diverse forms of explicit knowledge so as to permit the creation of new knowledge (for example in meetings or conferences). Nonaka gives an extended example in the process by which Matsushita developed the first ‘Home bakery’ (i.e. bread making machine). The first machine designed did not produce bread of sufficient quality so a software engineer was sent to train as a baker (socialisation). She then attempted to convey the technique of kneading to the mechanical engineers. To do this she
used metaphors describing the kneading as “twisting stretch” (externalisation) and this enabled the mechanical engineers to modify their design (combination with existing engineering knowledge) [Nonaka and Takeuchi, 1995:103-6].

This approach has been viewed as too mechanistic and simplistic from a knowledge management perspective [Ram, 2002]. However, the most significant critique from our perspective has come from Tsoukas [2003] whose analysis is also inspired by Polanyi, and who argues that Nonaka’s idea that tacit knowledge is awaiting ‘translation’ or ‘conversion’ into explicit knowledge is incorrect because this misses the ‘ineffability of tacit knowledge.’ Tsoukas notes that Nonaka does not view the knowledge acquired concerning kneading as qualitatively different from knowledge acquired through reading manuals since both can be expressed as explicit rules. Tsoukas rejects this as a very impoverished view of knowledge, since it reduces knowledge to merely what is articulable. “Skillful knowing contains an ineffable element; it is based on an act of personal insight that is essentially inarticulable” (ibid.). Tacit knowledge according to Tsoukas cannot be captured because “even the most explicit kind of knowledge is underlain by tacit knowledge”. Tsoukas does not reject discussing knowledge but rather than view this as a ‘conversion’ into an explicit format, this should be viewed more as a ‘reminder’ in Wittgenstein’s sense [Wittgenstein, 1953:§89].

Another significant idea concerning knowledge discussed in the Knowledge Management literature is the concept of distributed knowledge. This idea originates in the work of Hayek who was concerned with the analysis of rational economic systems. He saw the key problem in economic policy or planning as being how to utilise knowledge which is dispersed among many individuals [Hayek, 1945] and this appears to have been the first acknowledgement that knowledge could indeed be conceived of as distributed. Hayek’s approach is that price is the means for conveying that type of knowledge i.e. it acts as a heuristic or summariser of the relevant knowledge which in fact is disseminated across many individuals. Distributed knowledge came to the fore again with the work of Tsoukas [1996] who proposed a conception of the firm as a distributed knowledge system. In a sense this had also been foreshadowed by [Cyert and March, 1963] who proposed that organisations could ‘learn’ from experience and adapt to changes in their environment. Distributed knowledge is defined as knowledge which is known collectively by a set of agents but no single agent is in possession of the complete knowledge. This is in contrast to common knowledge which is defined as being held in toto by each agent [Foss and Foss, 2003]. Thus it is that a team of engineers can be brought together to work collectively on a problem because their combined distributed knowledge is productive and creative. In such a context there must be some degree of knowledge overlap i.e.
awareness of what the other agent knows.

Distributed knowledge is closely related to organisational memory or corporate memory, since the knowledge possessed by a firm is obviously distributed across its employees. However, different writers have perceived corporate memory in different terms. Walsh and Ungson [1991] identify six different ‘repositories’ for an organisation’s memories: individuals, culture, transformations, structure, ecology and external archives. These reflect to some extent the range of perspectives. Some writers focus on the archives and define the corporate memory as “the documents and other information kept by an organisation for later re-use” (Megill 1997, xviii, cf. also Gray 2002). This is in accord with viewing corporate memory as consisting of information which can be stored, transmitted etc. [Terrett, 1998, Hendricks and Vriens, 1999] and for which consequently appropriate software systems can be constructed (e.g. Dieng et al. 1999, Simon 1996). For example, Van Heijst, Van der Spek and Kruizinga define it as an “explicit, disembodied, persistent representation of knowledge and information in an organization” [van Heijst et al., 1996]. It is especially from this perspective that ontologies have been proposed as structures which can function as corporate memories [Staab et al., 2001, Vasconcelos et al., 2003]. In contrast, Kay views organisational knowledge as the essence of a firm, “more than the sum of expertise of those who work in the firm” [Kay, 1993].

The view one takes on the nature of knowledge and especially organisational knowledge significantly affects the role that information technology is seen to play. Some authors have equated knowledge management with the software systems built to support knowledge management activities. Especially culpable in this regard are the promoters of document management systems, database and groupware products, as well as many academic researchers. This has been criticised because it persists in the fallacy that “information technologies can store human intelligence and experience” and also that “Information technologies can distribute human intelligence” [Hildebrand, 1999]. Both these assumptions appear to take basic challenges of AI for granted and assume they have been solved.

2.1.3 The AI View

Circumventing these problems, Artificial Intelligence (more specifically ‘Good Old Fashioned Artificial Intelligence’) has usually taken a more practical view. The acute need for representing knowledge was apparent in the early attempts at machine translation which failed for lack of general knowledge [Russell and Norvig, 2003, 21]. An awareness of this drove the efforts to construct ‘expert systems’ or ‘knowledge-based systems’ and some attempts were extremely successful in a narrow domain e.g. DENDRAL which interpreted the

---

5 The following discussion does not address approaches other than GOFAI such as connectionism.
output of a mass spectrometer [Feigenbaum et al., 1971]. In the course of time, researchers in AI sought to find ways to standardise both the formalisms used in knowledge representation and to create general and re-usable knowledge components (hence the interest in ontologies in recent years and efforts such as CYC [Lenat and Guha, 1990] and the Standard Upper Ontology [IEEE, 2002]). The important point, however, is that knowledge has been viewed in most part as that which is expressible in propositional or first order logic. In this sense it has taken on the perspectives of the philosophical tradition wholesale, especially traditions concerned with the development of formal languages. In AI’s defence, obviously knowledge which is expressible in some form of logic is formalised and consequently usable by a computer. Although there is substantial AI work which is not connected to a symbolic approach nor to logic (e.g. neural networks, data mining, etc.), these efforts are only of relevance as tools in order to construct knowledge representations and not to the tradition in AI which has given rise to the current interest in ontologies. A major focus of AI is to construct ‘agents’ [Russell and Norvig, 2003] and these necessarily have an internal representation of their domain usually expressed as some form of an ontology. These representations focus exclusively on ‘facts’ about the world (e.g. ‘John loves May’ or ‘The box is on the chair’) so in effect such research equates knowledge entirely with facts.

\[ \text{knowledge} \equiv \text{facts} \]

Some writers distinguish ‘data’ which is the symbolic representations in computer systems, irrespective of whether they are interpretable or not from ‘information’ defined as data which is interpretable either by a machine or to a human being [O’Hara, 2002, AKT, 2001]. Therefore, a web page or a database is to be considered information. Knowledge is that subset of information which is “usable for the purpose of supporting or suggesting action” [O’Hara, 2002, :48]. To take an example from data mining, supermarkets collect huge amounts of data concerning their customers purchasing activities. This data is held in enormous databases and cannot be read or used by human beings in its raw form. However, it is possible to apply a variety of data mining techniques to turn the data into usable knowledge on the basis of which automated or human purchasing decisions could be made.

This is a very pragmatic perspective whose focus is clearly on what works and what is needed. The niceties of philosophical distinctions and the abstract proofs and counter-proofs to be found in epistemology are of no interest to the practical needs of modern day decision makers overwhelmed by the data at their disposal. This still does not come to terms with the fact that a great deal of knowledge that is of immense practical and commercial significance is non-propositional and thus is not ‘information’ whether useful or not. For example, an aeronautical engineer explained to me that when his company received some new jet engines, he had to inspect them. He immediately was concerned by the length of the fuel pipe which was unsupported and in his opinion
would wear out due to vibrations. This proved to be the case, but there was no propositional rule or piece of knowledge in his internal representation which stated, “If fuel pipes are longer than x, they pose a danger.” Thus, while this kind of approach is clearly a step forward from the abstractions of traditional epistemology, it is far from an adequate solution because so much knowledge, so much that is of immense scientific and commercial significance, is not only intangible but also non-propositional. Other writers, from the Upanishads to the present day, have stressed that quite apart from scientific rational knowledge, there are forms of knowledge which could be described as intuitive. Given that good decisions are often based on intuition or ‘gut feeling’, there is a broad area of knowledge which has to be recognised as important even if it is not possible to capture or represent it.

2.2 THE DISCOVERY AND TRANSMISSION OF KNOWLEDGE

Two major concerns from the perspectives of science, education and knowledge management are how to discover new knowledge and how to transmit knowledge from one person to another. The process of discovery of new knowledge has become ever more important in a ‘knowledge economy’ as that is the prime competitive advantage which companies are said to have [Locke, 1999], [Hitt et al., 2000]. As more and more commercial value is placed not on the physical assets of a company but on the intellectual capital present in its employees, so the need to transmit knowledge from one individual to another becomes paramount. To understand what role ontologies might have in this regard, we briefly consider some ideas concerning knowledge discovery and transmission.

DISCOVERY It was Sir Francis Bacon in the early 17th century who, reflecting the spirit of his age, described the inductive or scientific method as a means for acquiring knowledge [Bacon, 1605, 1620]. This was in contrast with the scholastic and Aristotelian traditions of deduction depending on existing axioms (usually the Bible and Aristotle’s works). Bacon believed that natural knowledge is cumulative, the result of a process of discovery. This lay the foundation for the use of experimentation in science. It also underlay the assumption that all (scientific) knowledge was the result of cumulative observation, experimentation and induction. Like Bacon, Descartes, a few years later, fundamentally questioned knowledge derived from tradition, but in contrast argued for a deductive approach to the acquisition of knowledge, laying his faith in the use of reason [Descartes, 1637/1991]. Descartes’ approach to knowledge is based on his famous statement

---

6 It could be argued that this case results from “implicit knowledge” in Dennet’s sense i.e. derived from a complex body of propositional knowledge available to the engineer, but such an approach essentially founders in face of the Frame Problem because both the totality of that body of knowledge is currently impossible to specify and the inferences which could be drawn are open-ended and take too long to make [Haugeland, 1985].
“Cogito ergo sum” which makes the individual’s perception of truth (and consequently knowledge) privileged over tradition, any concept of objective truth or shared truth (Descartes was an internalist Newman [2000]). Equally important was Descartes’ distinction between mind and body and his scepticism concerning sensory experience, because our understanding of knowledge since then has been predominately couched in terms of mental activity i.e. beliefs. Many philosophers have taken beliefs for granted and disagreed on how to explain any connection to sensory or embodied experience. In contrast, Bacon’s perspective on knowledge discovery was to be built on by Locke who took the view that knowledge arose from sense perception and ‘reflection’ and did not depend on innate principles or maxims [Locke, 1690/1975]. While all knowledge was founded on sensory perception, for Locke, reflection and the interpretative mind play an important role in how those sensory inputs are understood and interpreted.

An important alternative to the cumulative view is falsificationism, which has been an extremely influential account of the discovery of knowledge in the past century and was put forward by Karl Popper in his The Logic of Scientific Discovery [1959]. He believed that knowledge arose from hypotheses which were open to refutation. Human beings have a natural tendency to make guesses about their environment and these hypotheses or theories can then be falsified. Only theories which could potentially be falsified are considered scientific. Scientific progress and consequently the acquisition of new knowledge is dependent on the scientist continuously seeking to refute their theory. Crucially scientific hypotheses cannot be conclusively confirmed in Popper’s view.

The importance of the interpretive mind in knowledge discovery came to the fore in Thomas Kuhn’s work on process of scientific discovery [1962]. His well-known view is that scientific knowledge is not the result of a steady, cumulative collection of items of knowledge but rather that there were long periods where one scientific ‘paradigm’ or ‘disciplinary matrix’ dominated with brief ‘revolutions’ which shattered the prevailing mind set. Normal science (i.e. during the non-revolutionary periods) does make progress (in contrast to Popper’s view) and scientists need to adhere to the prevailing shared beliefs, methods and techniques for this to succeed. Typically scientists do no reject a theory when they encounter falsifying evidence, unless there is another alternative theory available. Kuhn in fact viewed sciences with many contrasting theoretical paradigms as sciences which were in their infancy. There is for Kuhn a tension because on the one hand a scientist has necessarily to be conservative in order to work within a paradigm and yet innovation might trigger a revolution [Kuhn, 1977]. A concomitant of Kuhn’s approach was that competing scientific theories were ‘incommensurable’ i.e. that there was no way to compare them or evaluate one with respect to another because the meaning of basic terms or concepts was affected.

---

7 Recent examples are the ideas of Putnam concerning Natural Kinds [1975] and Kripke’s rigid designators [1972]
by the theories they were embedded in. This is a view Kuhn shared with other philosophers including Feyerabend [1975]. Their views concerning the determination of meaning by the theory one is using can be seen as a philosophical version of linguistic determinism i.e. the Whorf-Sapir hypothesis [Whorf, 1956, Sapir, 1949]. The trouble with linguistic determinism is that it implies that concepts outside the “Weltanschauung” are beyond reach [von Humboldt, 1836/1999], and consequently from the perspective of knowledge discovery that there are definite limits to what can be thought up, discovered or invented.

Ontologies can be viewed on the one hand as structures on which to hook the knowledge that is discovered, much as the Linnaean system (which provides a fundamental inspiration and intellectual precedent for modern ontologies) provided an organised cognitive space for biologists to ‘place’ their discoveries within. This is in keeping with a cumulative view of knowledge. However, inevitably the ontology will reflect the prevailing ‘paradigm’8. From a Popperian perspective, one could argue that when new knowledge does not fit in the ontology, a new one would have to be constructed. In this sense we identify the ontology with the current theory or hypothesis. On the one hand one might expect ‘revolutionary’ moments where ontologies are re-written, but on the other hand Kuhn observed that scientists mostly “patch up” the current paradigm. The reality however is that rarely are there revolutions in ontologies and no one has ever ‘proven’ an ontology to be false. Quine’s view is closer to the actual practice of modern day ontology engineers, especially given the fervent desire to ‘re-use’ knowledge [Fikes and Farquhar, 1999]. New knowledge is added to an existing edifice and the tensions this causes eventually lead to a re-adjustment. We can see this in the evolution of the Gene Ontology [Ashburner et al., 2000] and the manner in which terms are made obsolete or radically changed as knowledge progresses9.

In the knowledge management literature there is an awareness that the codification of knowledge (or at least the attempt) may stultify the dynamic of new knowledge creation [Lang, 2001]. For example, the management of Rolls-Royce have expressed doubts with respect to the use of ontologies and the extent to which they would prevent their engineers from “thinking out of the box”10. This raises the question whether the standardisation which ontology construction leads to is actually beneficial and under what conditions.

TRANSMISSION One of the key roles that ideally ontologies should play for knowledge management and AI is in the transmission of knowledge. If ontologies are adequate representations of knowledge, if they can act as organisational memories, then they should provide a means for knowledge transmission. The transmission in a knowl-

---

8 In this regard consider the conflict between the Linneaen and the cladistic perspectives in biology.
10 Reported by Nigel Shadbolt, p.c.
Figure 2. The taxonomy in Open Directory under the heading food. Numbers indicate the number of web pages to which the user can navigate.

edge management situation would be from one individual or group to another individual or group. In the AI context, it would be from humans to a machine or vice versa. Analysing the corporate context, Hansen et al. [1999] have made the distinction between a ‘personalization strategy’ where most knowledge resides in individuals and computer systems mostly facilitate that interaction, and a ‘codification strategy’ to be found where companies sell standardised products and knowledge is codified in databases and used again and again.

At the heart of knowledge transmission lies a conundrum which was first addressed in Plato’s Meno. The epistemological question raised there was how can one enquire into a subject when one knows nothing about it:

MENO: And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of enquiry? And if you find what you want, how will you ever know that this is the thing which you did not know?¹¹

In the modern version, the question is how do we get someone to acquire knowledge concerning a domain about which they cannot formulate a query. Obviously information retrieval systems cannot help here for lack of relevant keywords. “...the chance is extremely high that an inexperienced user looking for information gets stuck in a dead-end road” [Staab and Maedche, 2001, 3]. This is where ontologies can play a role.

The main way we currently observe ontologies functioning in knowledge transmission between human beings is through the visualisation

¹¹ Meno 80d, [Plato, 1875]
of taxonomies most typically seen in Yahoo (www.yahoo.com) or the Open Directory (www.dmoz.org). In such cases, the ontology functions as a navigational tool to allow the user access to a document, a web page or some other text (cf. “...the principal part of transmission of knowledge consists in the writing of books ...” Bacon 1605). However, the actual structure or format of the taxonomy is not consciously considered by the user. For example, in Figure 2, we can see the taxonomical structure under the term food and we notice the complete lack of consistency in the groupings (cheese, confectionary, Durian, Fast Food, jell-o, ... in one grouping). Only at an unconscious level do users absorb the ‘knowledge’ that one concept is connected or associated to another. Nonetheless one of the main areas where ontologies (even formal ontologies) are promoted is in corporate portals [Staab and Maedche, 2001] and one could argue that navigation is a form of knowledge transmission. Given that ontologies cannot in an active manner convey knowledge, in the context of portals, they can show the user what types of categories there are in the neighbourhood of a given concept. In this sense the knowledge transmitted is a knowledge of what there is ‘out there’. Although most corporate portals are merely search interfaces (a private company-internal search engine), there is an awareness of the importance of using taxonomies and ontologies to augment the knowledge transmission aspects of such technologies [Gilchrist and Kibby, 2000, Reif, 2004]

One of the most interesting examples of the use of ontologies in a portal is the Flamenco project of Marti Hearst (cf. Section 3.4.4)[Hearst et al., 2002, Yee et al., 2003]. Here faceted metadata is used as descriptors for large collections of photographic images. Hearst emphasises how her approach facilitates the user to learn about the image collection and to “navigate laterally” [Yee et al., 2003, :404]. A lot of effort has been expended to make the interface ‘flow’ i.e. to allow “users to move through large information spaces in a flexible manner without feeling lost” [Hearst et al., 2002, :45]. Not getting lost is essential if users are to understand the inter-relation between different items of information i.e. to acquire some form of knowledge.

Both the above examples are cases where the primary intent was for there to be transmission of knowledge from the system to the human user. In contrast, Cyc [Lenat and Guha, 1990] is a project initiated with the express intent to make common-sense knowledge available to computers. The need for the codification of common-sense knowledge in a form suitable for computers was recognised almost since the conception of Artificial Intelligence [McCarthy, 1959]. Under the leadership of Doug Lenat, the Cyc project has undertaken to formalise common sense knowledge so that inference engines could operate over the representations. The project has succeeded in formalising hundreds of thousands of concepts each with associated rules. Knowledge is represented in a version of FOPC and is contextualised by

---

12 This can be considered to be, in effect, a multidimensional ontology, where items can be labelled with concepts from different locations or ontological subtrees. Hearst describes this as using orthogonal sets of categories.
the use of microtheories which consist of sets of assertions, rules and facts about a specific limited sphere of the world. In a sense each microtheory can be seen as a domain specific ontology. Lenat hoped that Cyc could be used for a large variety of AI related applications Lenat and Guha [1991] although the outcome has not been so clear cut. A number of authors have criticised the approach for its excessive dependence on representing knowledge explicitly, for assuming that all knowledge can be represented in one framework and for depending exclusively on deduction [McDermott, 1993, Neches, 1993, Sowa, 1993, Yuret, 1996].

The concept of knowledge transmission is closely related to that of knowledge codification. If knowledge could successfully be codified then face to face interaction could be obviated. However, even in such highly technical and complex domains as nuclear weapons, it has been shown that knowledge transmission cannot occur by impersonal means. MacKenzie and Spinardi [1996] argue that explicit knowledge cannot be lost “barring social catastrophe” but that tacit knowledge (if it is not transmitted from person to person) can be lost, just as skills if not practiced disappear [cf. Ohmagari and Berkes, 1997]. They undertook an extensive study of the process of nuclear weapon design and found that personal face-to-face interaction between researchers was vital because there was a great deal of tacit un-encoded knowledge which could not be encoded in a straightforward manner. In effect, if personal transmission were lost, nuclear weapons would to a certain extent have to be re-invented. Mackenzie and Spinardi are very clear as to the limitations of the assumption that science is both explicit and cumulative in its knowledge.

Flaming [2003] has studied the change from an exclusively oral culture to one based on literacy and formal learning in the training of nurses. Nurses were traditionally trained in an apprenticeship which encouraged contextualised knowledge. As result of transforming nurses into university students, there have been a number of consequences including the “sedimentation of words” i.e. the loss of flexibility as words become acontextualised. He notes also that this has devalued sensorial and experiential knowledge. From differing perspectives, in both these cases (nuclear weapons and nursing), the difficulties of encoding knowledge have significant ramifications for our understanding of the role that an encoding knowledge structure such as an ontology can play in knowledge transmission.

2.3 FORMS OF KNOWLEDGE REPRESENTATION

It is appropriate at this stage to consider the nature and purpose of a knowledge representation, both in general and specifically with regard to ontologies. A simplistic view would say that a knowledge representation is a model of the world which can be used to reason about it. However, there is in fact a whole range of functions, assumptions and aspirations encoded in a given type or instance of knowledge representation [Davis et al., 1993].
A knowledge representation is a surrogate. Intelligent entities go through processes of reasoning about the world, often in order to plan actions. The reasoning involves internal representations but the objects reasoned about are usually external, outside in the world. Consequently, the knowledge representation is a surrogate standing for the objects outside in the world. The correspondence between the surrogate or knowledge representation and the world is the semantics. The ‘fidelity’ of the correspondence depends on what the knowledge representation captures from the real object and what it omits. Perfect fidelity is impossible because “the only completely accurate representation of an object is the thing itself” [Davis et al., 1993, :18]. This is both impossible and absurd, and is reminiscent of the Lagadoans in Gulliver’s Travels who carry “about them such things [the objects themselves] as were necessary to express the particular business they are to discourse on” [Swift, 1726/1967, :230].

Two important consequences are that all forms of knowledge representation must unavoidably “lie” or more precisely “misrepresent” even if it is merely by omission, and that all forms of reasoning about the world will reach false conclusions at some stage. The soundness of the reasoning process cannot avoid this because the representation will always be in some way incorrect. It is essentially a practical engineering decision to find ways minimising the errors given the specific task at hand.

From the perspective of ontologies, this point has important consequences. Firstly one of the major claims for ontologies is that they will facilitate the interchange of knowledge between (for example) agents, or the reuse in different systems. However, if each agent or system has an imperfect model of its universe, knowledge interchange or sharing may increase or compound errors which were not visible in the initial use of an ontology. Secondly, and closely related, ontologies for the same domain will inevitably model different aspects of the external world depending on the focus and assumptions of the ontology’s authors.

A knowledge representation is a set of ontological commitments. The choice of knowledge representation is also a decision about how and what to see in the world. This is both unavoidable because representations are imperfect and useful because it allows the representation to focus on that which the representation’s author considers relevant or interesting. They see these choices as allowing us to cope with the overwhelming complexity and detail of the world. Consequently, it is the content of the representation i.e. the set of concepts chosen and their inter-relation which provides a particular perspective on the world. The choice of notation (logic, LISP, or OWL) is unimportant.

It is interesting that, with respect to ontologies, an immense amount of effort has been expended in developing and defining ontology representation languages and in contrast, relatively little effort has been
made to analyse what ontological commitments particular ontologies make. The only exception to this has been Guarino’s critique of structures such as WordNet for not conforming to a logician’s world view in terms of consistency and logical rigour (Guarino 1998, Gangemi et al. 2001 but cf. Wilks 2002).

An inherent assumption of the all authors in this field (including Davis et al.) is that ‘concepts’ are the key building blocks, and we manipulate concepts with words. All ontologies I have encountered use words to represent the concepts and to mediate or provide a correspondence with the external world. Consequently a large range of items in the world or experiences which do not lend themselves readily to verbal expression cannot be modelled. We could describe this as the ‘Ontological Whorf-Sapir Hypothesis’ i.e. that that which cannot be captured by words cannot be represented in an ontology 13.

A KNOWLEDGE REPRESENTATION IS A FRAGMENTARY THEORY OF INTELLIGENT REASONING. The way a knowledge representation is conceived reflects a particular insight into or understanding of how people reason 14. They believe that selecting any of the currently available representation technologies (such as logic, frames, knowledge bases, or connectionism) commits one to fundamental views on the nature of intelligent reasoning and consequently very different goals and definitions of success [Davis et al., 1993, 23].

Looking at the discussions surrounding the use of ontologies from this perspective is instructive. Currently there are a number of communities which use the term ontology to describe an artifact (as opposed to a branch of philosophy) and they yet derive from very different origins. One such community is that concerning knowledge engineering and knowledge representation (in a sense classic AI). Representatives of this approach include Gruber [1993], Gómez-Pérez [1999], Fensel, van Harmelen, Horrocks, McGuiness, and Patel-Schneider [2001], and historically they are descendants of both the logicist tradition, especially description logics [Brachman, 1977], and the frames tradition of Minsky [1975]. Another community is that which has arisen from the computational linguistics world, typified by Nirenburg’s Mikrokosmos [Mahesh and Nirenburg, 1995, Mahesh, 1996], but also represented by WordNet [Fellbaum, 1998] which has been so widely used in Natural Language Processing research, even though its initial conception was as a tool for psycholinguistics. There are of course other traditions but they have all laid a claim to the currently fashionable Semantic Web, which claims to be all things to all men [Lassila and McGuinness, 2001].

The OWL language, which has been developed by the W3C consortium as a standard language for describing ontologies for the Semantic Web, comes in three flavours (http://www.w3.org/2004/OWL/).

---

13 Limitations of space and time prevent me from investigating the relevance of connectionist and situated theories [Brooks, 1991]
14 The very use of a representation is a commitment to symbolic AI or what Haugeland call “Good Old Fashioned AI” [Haugeland, 1985, 112ff.].
The simplest, *OWL Lite*, is intended merely to provide a classification hierarchy and "provides a quick migration path for thesauri and other taxonomies" [McGuinness and van Harmelen, 2004]. *Owl DL* is an attempt to encode as much of Description Logic as possible without sacrificing computational efficiency. Finally there is *OWL Full* which provides no computational guarantees but allows greater flexibility e.g. treating a class simultaneously as a collection of individuals and an individual itself. The existence of these three flavours reflects the differing traditions that have merged in the current effort to construct a standard. Nonetheless, this does not diminish Davis et al.’s point that by choosing a specific form of knowledge representation, there is a commitment to specific views about the nature of intelligence. After all, Minsky himself in his original paper on frames noted that his work was a “partial theory of thinking” [Minsky, 1975] and so equally there are limits and presumptions in a standard like OWL which have not been fully spelled out, but of which the ‘flavours’ are an indication. Minsky himself has taken a very humble view, stating in an interview, “I want AI researchers to appreciate that there is no one ‘best’ way to represent knowledge. Each kind of problem requires appropriate types of thinking and reasoning – and appropriate kinds of representations” [Minsky and Laske, 1992]. This comment should raise many questions as to the longer term viability of a standardisation approach like that of the Semantic Web community with OWL and the successful use of a limited range of knowledge representations.

A KNOWLEDGE REPRESENTATION IS A MEDIUM FOR EFFICIENT COMPUTATION. In the final analysis a knowledge representation must allow for computational processing, and consequently issues of computational efficiency will inevitably arise. For example, using taxonomic hierarchies both "suggests taxonomic reasoning and facilitates its execution" according to Davis et al. [1993, :27] and goes on to note that in seeking sufficient speed severe restrictions on the reasoning capacity of the representation may have to be made. An apparent example of this is the great disparity in the performance of the different systems which took part in the Halo competition (cf. more details in Section 11.2). The winning system (Ontoprise) had an ontology consisting of 41 concepts while the worst performer (Cycorp) had an ontology of 15000 concepts [Friedland et al., 2004]. Apart from the number of concepts in an ontology, the choice of representational formalism is assumed to affect computational efficiency. Thus the development of the different flavours of OWL mentioned above are a partial recognition of this aspect. However, the fact that *OWL Full* is not guaranteed to be ‘decidable’ unfortunately does not guarantee it to be sufficiently powerful to represent the whole gamut of what we can consider to be knowledge!

A related issue, to which we will return frequently, is the problem of the computational efficiency of the process of arriving at these representations. Most ontologies, just like most forms of knowledge representations in the past, are hand-built which ensures that they
Knowledge acquisition is costly and slow to construct, full of human errors, and out of date upon completion. This provides another layer to the ‘efficiency’ problem.

A knowledge representation is a medium of human expression. All forms of knowledge representation are both mediums of expression for human beings and ways for us to communicate with machines in order to tell them about the world. In general, the focus is largely on how easy it is to express something in a given language of representation, not on whether it is possible: “we must be able to speak it without heroic effort” [Davis et al., 1993, :27]. Thus it may be that certain things are possible to be expressed but the effort involved is too great.

That knowledge representation is a form of human expression is something frequently forgotten in the field. In Nirenburg and Wilks [2001], for example, Nirenburg’s insistence on the possibility of precise unambiguous meaning in a ‘representational language’ ignores this fact. Wilks’ response that the symbols in a representation are fundamentally language-like essentially reflects that a representation language is a means of human communication with all its dynamism, ambiguity and extensibility. This fact is frequently forgotten in the ontology building communities (such as the Semantic Web) who believe that their ontologies will achieve ‘precision’ and ‘exactness’ in the meaning of the terms (classes, concepts, etc.) in their ontologies.

2.4 Knowledge Acquisition

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
T. S. Eliot (The Rock)

In spite of the problems associated with the definition of knowledge and its scope, the fact remains that it is possible to create machine-readable representations of the world. If these capture only some part of what we loosely call ‘knowledge’, a great many potential technological innovations can see the light of day. For example, one of the major challenges for language technologies at present is the poor quality of speech recognition. Signal processing and statistical methodologies may have reached the limits of what they can provide. If that is the case, then it is hard to see how without building in semantic and pragmatic (i.e. real world) knowledge that the accuracy and usability of this technology will improve. Such efforts depend on the construction of knowledge bases of some sort, most typically ontologies.

The construction of ontologies is primarily a task of knowledge acquisition and knowledge maintenance. The knowledge acquisition process proceeds from the assumption that there is a great deal of information out there, some of which can be turned into knowledge,
i.e. some of which is actually useful. The range of potential sources of knowledge is immense. At the most prosaic level, human beings acquire knowledge just by being in an environment and observing it with their senses - vision, smell, hearing, touch, taste. However, computers cannot (as yet) perform the same functions. We cannot apprentice our Apple Mac to Professor Joe Bloggs and expect it to become an aeronautical engineer after some time. Thus, knowledge acquisition is dependent on other methodologies.

Considerable effort has gone into developing "knowledge elicitation" techniques especially for the creation of expert systems. These methods include interviews, process tracing, and conceptual methods (cf. survey in Cooke 1999). Conceptual methods are the closest to the techniques of ontology learning in that they attempt to elicit either through interviews or analysis of documents the relevant concepts and then relatedness judgements are obtained from one or more experts. This data is then reduced and represented (using multidimensional scaling or pathfinder networks, Schvaneveldt 1990) and finally interpreted. These traditional approaches may involve making tacit knowledge explicit e.g. when asking an expert to describe the thought processes they went through in order to come to some decision or design. However, in terms of formalising the outcome this knowledge acquisition process, the human, manual act of interpretation is always present and seemingly unavoidable.

The challenge in automating ontology learning lies in seeing how far this ‘manual act of interpretation’ can be automated. The biggest task for knowledge capture lies in the raw processing of information and turning it into knowledge. While information of specific sorts may be held in databases, such as the customer purchasing patterns in a supermarket mentioned before, the overwhelming majority of data is to be found in texts. Some knowledge may be acquired by obtaining it from different existing sources and integrating it, as for example when companies merge and their respective ontologies or taxonomies are brought together into one knowledge structure. However, the vast majority is present in texts people write because people communicate through the written word. Scientists write papers, reports, emails and textbooks. Engineers write emails, reports, discussion documents and manuals. Managers communicate in writing, in letters, emails, presentations and reports. No matter what the context, there is always some form of written document generated for one reason or another. Both lexicographers and researchers in empirical NLP have long recognised this and sought out to build ‘representative’ corpora (e.g. the BNC), representative of the language in general or of specific linguistic communities. Texts have been recognised more and more as the prime source of information. Short of entering the minds of individual employees, a company’s next best point of access into their thought processes and ‘knowledge’ is to find means to analyse the texts produced as a result of normal activity.

One of the key features of knowledge is that a large proportion of it is dynamic, i.e. it is ever changing, continuously updated, being
added to and also a considerable amount of it is being forgotten. The task of knowledge maintenance is to keep track of this moving target. Some knowledge is like a fleeting fashion, here today, gone tomorrow. Some pieces of knowledge need considerable analysis to identify how, or where, they fit with existing knowledge. Other pieces of knowledge need to be forgotten or just lose their prominence. Yet other pieces of knowledge change the trustworthiness of existing knowledge. All this means that ontologies need to be continuously updated in order to reflect ‘current knowledge’. It is in this context that we have chosen to explore just how far it is possible to acquire ‘knowledge’ from text. There is a great deal which it is clear cannot be captured from texts because text do not or cannot carry that information. In a certain sense, our research is an attempt to determine the limits of what is actually present in texts, i.e. if we take away the world knowledge of the reader, what is left?
"The time has come," the Walrus said,
"To talk of many things:
Of shoes—and ships—and sealing-wax—
Of cabbages—and kings—
And why the sea is boiling hot—
And whether pigs have wings.”

Lewis Carroll
– Through the Looking Glass

The word “ontology” is generally taken to concern the study of being and as such was originally a part of metaphysics - “it is the study of existence, of all the kinds of entities, abstract and concrete, that make up the world” [Sowa, 2000, :51]. The term was a coinage of the late 17th century, a time of great scientific and philosophical ferment when many of the early attempts at constructing what we now call ontologies were made. In this work, I use the term in a general sense to describe structured representations of the world. Ontologies are closely related to thesauri and taxonomies. The former are forms of ‘wordbooks’ [Wilks et al., 1996] where words and terms are organised under headings and sub-headings. Taxonomies in contrast are an attempt to conceptually organise reality (of some sort), most typically in the taxonomies of biology traditionally associated with Linnaeus. The distinctions between these types of data structures (and others) are more a matter of academic background or training than substantive differences in approach and objectives (cf. Section 3.6).

3.1 Historical Background

Collecting items into groups and categorising the world are two basic human instincts which, with regard to human language, eventually became entwined to produce what we now call an ontology. Perhaps the earliest known such activity took place in Mesopotamia during the 3rd millennium BC, where lists were made in Sumerian of legal and administrative terms, of trees, of domestic and wild animals, of occupations, of parts of the body and so on, together with their Akkadian equivalents [Mengham, 1993]. This could hardly be called an ontology in the modern sense and in any case the Sumerian legacy did not survive and thus, for reasons of cultural discontinuity we cannot claim any influence on subsequent dictionaries or thesauri. It is, nonetheless, indicative of the extent to which this urge to categorise the world is innate in the human species. The legacy which did come down to us, however, was that from the Greeks.
3.1.1 Aristotle

The Sumerian efforts were practical solutions to specific needs. In contrast, formal attempts at describing categories go back at least to Aristotle and his work on logic and on categorising animals. The analysis of categories plays a central role in his thought and his views have influenced scientists and philosophers up to the present. There are two aspects which are important here. One concerns the nature of categories and their relation to the world and to language. And the other concerns the use of categories to create taxonomies of the world.

The ‘classical’ view of categories is usually attributed to Aristotle, one where a category has the following characteristics [Taylor, 1989, Lakoff, 1986]:

1. It is defined in terms of a conjunction of necessary and sufficient features. Thus for example, the category MAN may be defined by a set of features [ANIMAL], [TWO-FOOTED] and these are the only features necessary and no further features are needed.

2. The features are binary. This means that either something is [TWO-FOOTED] or it is not. There cannot in the ‘classical’ approach be degrees of two-footedness.

3. The category has clear boundaries. This is a concomitant of 2.

Linguistics has been particularly affected by this view, as can be seen in phonological feature theory and in approaches to semantics such as componential analysis. It is an approach which still dominates NLP and many other scientific domains. However, this view of categories has been severely criticised by a series of writers in the last 40 years [Taylor, 1989, :38ff.] from philosophical [Wittgenstein, 1953, Putnam, 1975], psychological [Rosch, 1973, 1975, Rosch et al., 1976] and linguistic [Lakoff, 1986, Langacker, 1987] perspectives. Essentially, it has been found that the Aristotelian view (or that attributed to him) does not fit with reality and our perception of it. Many categories cannot be defined by a set of necessary and sufficient features; Wittgenstein’s example of “game” is especially famous [Wittgenstein, 1953, :31-3]. Many ‘features’ cannot be considered binary, including [VOICE] in phonology [Ladefoged, 1973], or the feature [TALL]. Categories often do not have clear boundaries, as Rosch has shown for [FURNITURE] and other categories [Rosch, 1973, 1975]. Aristotle began by establishing 10 genera into which all things could be categorised, although it is still debated whether Aristotle intended these categories as a classification of the external world or merely as a classification of words in the language. ‘Aristotelian philosophy sees things in terms of qualitative, as opposed to quantitative, differences. It sees things in terms of kinds, natures and essences ...’ [Slaughter, 1982, :26]. Between the time of Aristotle and the end of the 17th century, science was concerned above all with categorisation, i.e. taxonomizing observable phenomena and experience. An essential assumption of the Aristotelian view was that by categorising an object one was concurrently defining it. Thus
the placement of an object in a given taxonomic hierarchy effectively identified its characteristics. This was what science was about.

3.2 THESAURIS

3.2.1 Wilkins

There was a great movement in the 17th century to establish ‘philosophical languages’, which meant languages that would be universal, which would not have any of the shortcomings of natural human languages, and in which one could perfectly express scientific knowledge. Some of these attempts merely wanted to create a new language by creating word lists; others realised that creating a language involved starting out with concepts. In these efforts “an inventory of non-linguistic items, viz. notions and things, [was] set out in advance; the language itself [would] systematically reproduce this inventory in the linguistic form of the new words” [Slaughter, 1982, :143]. This, it was hoped, would avoid the vagueness and imprecision found in conventional languages due to varying usage and faulty observation. The most famous of these was made by John Wilkins, Bishop of Chester, who was one of the leading intellectuals of the 17th century. Apart from being a founder of the Royal Society and a polymath, in 1668, Wilkins published “An essay towards a real character and philosophical language”. This is the crowning attempt of 17th century efforts to construct universal languages.

Writers such as Bacon and Descartes and whole host of lesser known authors all expressed an interest or made attempts at creating universal languages [Slaughter, 1982, Large, 1985, Knowlson, 1975]. The particular interest of Wilkins is that his attempt is the most worked out and it had great influence on scientific taxonomy. It is based on his understanding that for such a universal language to work it must describe the world and makes a serious attempt to taxonomise the whole of contemporary knowledge about the world. To this end, he obtained the help of a variety of friends and acquaintances (all members of the Royal Society) including Francis Willoughby on animals, Samuel Pepys on nautical terminology and, significantly John Ray on plants. Ray is considered one of the fathers of modern systemics and taxonomy in biology.

The Real Character consisted on the one hand of a complex system of tables describing and taxonomising the known universe and, on the other, a system for pronouncing and writing all the concepts. The tables were an exhaustive attempt to describe all of human knowledge and the relations between its parts. Thus Wilkins based his taxonomy on 40 ‘genera,’ much as Aristotle had proposed ten categories. Each genus was then subdivided according to specific rules (by ‘differences’) and each of these categories was broken down into species so as to classify all the possible items under that general category. Thus STONE VIII was divided into the following:
The importance of Wilkins for the subsequent construction of a thesaurus lies not only in his taxonomy but also in that, under each heading, he placed a number of synonyms and often antonyms. It is interesting to note that Wilkins’ entry was very heterogeneous, especially in the light of subsequent efforts at organising synonyms into coherent sets.

While Wilkins’ efforts appeared very serious and scientific at the time with hindsight we can see that they were merely reflecting the knowledge and prejudices of the time. The example for STONE given above is, for us now, quite silly, and yet it represented the most advanced knowledge of its time. Clearly, stones can be classified in many different ways and there is no definitive means of classification.

3.2.2 Roget

In 1852, Peter Mark Roget published a work which, in edition after edition, has shown itself to be one of the greatest lexicographical success stories of all time. A thesaurus is not a dictionary, as Roget points out, but rather a compendium of words organised into groups which share meaning, sense or are otherwise associated in the minds of an average native speaker:

The purpose of an ordinary dictionary is simply to explain the meaning of the words; and the problem of which it professes to furnish the solution may be stated thus: - The word being given, to find its signification, or the idea it is intended to convey. The object aimed at in the present undertaking is exactly the converse of this: namely, - The idea being given, to find the word, or words, by which that idea may be most fitly and aptly expressed. For this purpose, the words and phrases of the language are here classed, not according to their sound or their orthography, but strictly according to their signification.

— [Roget, 1852, :v]

There is, however, a considerable distance between the lexicographical enterprise which was Roget’s creation and its actual use on an everyday basis. As an innocent user of a thesaurus (if such a person exists), one merely looks up a word in the index and sees where this leads, hoping to find a suitable synonym, phrase or other expression to capture the particular idea in one’s mind. However, Roget also

<table>
<thead>
<tr>
<th>Vulgar and of no price</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle priced</td>
<td>II</td>
</tr>
<tr>
<td>Precious</td>
<td>III</td>
</tr>
</tbody>
</table>

Table 2. Wilkins’ classification of ‘stone’.
provided a taxonomy of language (or rather of ideas) which is ignored by most people, and which is entirely inaccessible in most electronic versions. The taxonomic aspect of the thesaurus is important in that it contains a great deal of latent information concerning the interrelation between different areas of vocabulary.

Roget’s original work of 1852 was organised into a hierarchy of concepts in a manner similar to Wilkins’s taxonomy. Roget had six basic ‘Classes’ under which there were 24 Sections unevenly distributed (as shown in Table 3).

The plan of classification, Roget emphasises, was constructed on the basis of what was going to be most useful to the user, and on his own experience so as to “conduct the enquirer most readily and quickly to the object of his search” [Roget, 1852:10]. Roget had no pretensions, philosophical or scientific, about his system of categories other than their utility. He did not claim it to be perfect, all that he was convinced of was that it would be useful. In this, he differs from Wilkins, who had greater ambitions for his approach.

<table>
<thead>
<tr>
<th>Plan of Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Abstract Relations</td>
</tr>
<tr>
<td>1. Existence</td>
</tr>
<tr>
<td>2. Relation</td>
</tr>
<tr>
<td>3. Quantity</td>
</tr>
<tr>
<td>4. Order</td>
</tr>
<tr>
<td>5. Number</td>
</tr>
<tr>
<td>6. Time</td>
</tr>
<tr>
<td>7. Change</td>
</tr>
<tr>
<td>8. Causation</td>
</tr>
<tr>
<td>IV. Intellect</td>
</tr>
<tr>
<td>1. Formation of Ideas</td>
</tr>
<tr>
<td>2. Communication of ideas</td>
</tr>
<tr>
<td>II. Space</td>
</tr>
<tr>
<td>1. Generally</td>
</tr>
<tr>
<td>2. Dimensions</td>
</tr>
<tr>
<td>3. Form</td>
</tr>
<tr>
<td>4. Motion</td>
</tr>
<tr>
<td>V. Volition</td>
</tr>
<tr>
<td>1. Individual</td>
</tr>
<tr>
<td>2. Intersocial</td>
</tr>
<tr>
<td>III. Matter</td>
</tr>
<tr>
<td>1. Generally</td>
</tr>
<tr>
<td>2. Inorganic</td>
</tr>
<tr>
<td>3. Organic</td>
</tr>
<tr>
<td>VI. Affection</td>
</tr>
<tr>
<td>1. Generally</td>
</tr>
<tr>
<td>2. Personal</td>
</tr>
<tr>
<td>3. Sympathetic</td>
</tr>
<tr>
<td>4. Moral</td>
</tr>
<tr>
<td>5. religious</td>
</tr>
</tbody>
</table>

Table 3. Roget’s Plan of Classification.

Under the 24 sections were arranged approximately 1000 numbered heads which were (and are) the basic organisational units in the thesaurus. Within each head, words were organised first by parts of speech (nouns, verbs, adjectives, adverbs, phrases), and then, within each part of speech, into paragraphs. Apart from organising the vocabulary into heads, and then by part of speech, there existed in the first edition already, in embryonic form, a system of cross-references from one head to another. The paragraphs within heads were later
given keywords so that reference could be made to a head number and keyword within that head.

Roget’s original intention and expectation was that users would familiarise themselves with the structural organisation of the book and use it like that. In fact, most users start by looking at the index and then search under the different headings referred to in the index. This probably reflects laziness and lack of understanding of the organisational structure rather than an inadequacy in its design or conception. In fact, it is apparent from his introduction and from the thesaurus itself that Roget had given considerable thought to the structure and organisation of the book.

Very few changes in structure or layout have been made, neither in further editions in Roget’s lifetime nor since then when edited by Roget’s son, grandson and subsequent editors. The most significant change is in size. While the 1852 edition contained less than 70,000 references, editions since then have grown such that recent editions have over 300,000 references. The increase reflects not only an accumulation of vocabulary by the editors but also a substantial increase in the scope of the work. People expect to be able to find any and every item of vocabulary in a Thesaurus including lists of animals, plants etc.

3.2.3 Thesauri in Natural Language Processing

Thesauri have an honourable history in Natural Language Processing, especially as the categorisation of words was seen as a useful tool in handling word sense disambiguation. The earliest work using thesauri was undertaken by Masterman [1957] who experimented with the use of Roget in machine translation, in effect using “the thesaurus heads as a set of semantic primitives for semantic coding and analysis” [Wilks et al., 1996, :88-89]. Sparck Jones made the first attempts to use a thesaurus together with information retrieval techniques in her thesis (late 1960s, eventually published as Spärck Jones 1986). She focused on a sub-level of a thesaurus called a ‘row’ which consisted of close synonyms substitutable in a given context. These ‘rows’ were taken as sets of features and then clustering techniques were applied to see if they could be classified in accordance with their membership of the thesaurus heads. One of the key problems, as noted by Wilks et al. [1996], was that the clusters did not have labels or names, and this as we shall see below is a crucial difficulty in this area of research.

The thesaurus of Roget was used more recently in work on word sense disambiguation by a number of researchers including [Yarowsky, 1992]. His approach attempted both to adapt a thesaurus to a particular corpus and then use it to perform word sense disambiguation. The results varied, depending on the word being disambiguated, from 38% to 100% accuracy.

---

[1] It is fruitful to compare this with the use of the Internet and the manner of its access, especially the use of search engines vs. taxonomies such as Yahoo!
In Information Retrieval (IR) thesauri have been repeatedly used for query expansion ever since Yu and Salton [1975]. For example, both Qiu and Frei [1993] and Crouch and Yang [1992] present ways to construct a thesaurus automatically from text for the purpose of query expansion, and this has resulted in improved retrieval performance [Baeza-Yates and Ribeiro-Neto, 1993, :138]. Although Voorhees [1998] has shown that WordNet (which is closely related to thesauri in structure) is not useful in improving text retrieval, Kekäläinen [1999] and Sormunen et al. [2001] have argued that domain specific thesauri can significantly improve text retrieval. More recently, Clough and Stevenson [2004] have shown that using EuroWordNet for cross-language information retrieval can be effective.

### 3.3 TAXONOMIES

A thesaurus is essentially a very simple structure whereby ‘heads’ are associated with a set of words. The nature of the relationship between the head and its children is never specified and can in fact range over a whole variety of semantic and pragmatic relations. In contrast, taxonomies are one step towards a more formally specified knowledge structure in that a specific relationship is usually implied between a parent node/concept and the children. In taxonomies, the child is usually expected to have a set-subset relation with its parent. Taxonomies usually have had considerably greater depth than a thesaurus, i.e. more levels between the leaf nodes and the root. Another important contrast with thesauri is that thesauri have been conceived as representations of language, i.e. they are close relatives of dictionaries, while a taxonomy was usually conceived as a representation of the world.

The contrast between a thesaurus and a taxonomy is not always clear as is apparent from our description of Wilkins’ work above. Wilkins was part of a strong 17th century movement which led eventually to the complex biological taxonomies of Linnaeus in the early 18th century. A multitude of biological taxonomies was proposed in that century, but Linnaeus’ system won out due to its simplicity. It consisted of a hierarchy of ever more inclusive categories called ranks and had specific rules for ascribing a species to a specific category. For example, humans are Homo sapiens, part of the genus Homo, the family Hominidae, the order Primates, the class Mammalia, the superclass Tetrapoda, the subphylum Vertebrata, the phylum Chordata, and the kingdom Animalia.

Libraries have long used sophisticated taxonomies in order to clas-

---

2 The current controversy between the Linnaeans and the advocates of the cladistic “PhyloCode” points to significant issues for taxonomies. The PhyloCode advocates object to the old system because it is both an artefact of a creationist view of the world, and also by its principles (e.g. of having type species) it has resulted in the re-naming of many sub-families over time as new discoveries have changed scientists’ views. This objection acknowledges that the Linnaean hierarchy was an arbitrary product of its time, but also falls into the hubris that an alternative will not have to continuously adapt over time and as knowledge changes [Pennisi, 2001].
sify books and other media. The two best known examples are the Library of Congress Classification System (http://lcweb.loc.gov/catdir/cpso/) and the Dewey Decimal System (http://www.oclc.org/dewey/) both of which are products of the 19th century, even if they have been up-dated repeatedly. These systems make the un-stated claim that each topic in the taxonomy is more general than its individual children, i.e. the parent includes the children. The problem with library classification schemes is that they are not trying to classify objects, such as biological species, which appear to be limited in the number of dimensions relevant to classification, but rather they try to classify documents which by their nature often discuss a number of topics. A taxonomy of biological species is trying to ‘place’ each object so as to model the world from a particular scientific perspective. In contrast, a library classification system is trying to facilitate access to objects which may be covering a number of topics, or may be perceived from a number of perspectives. Sinclair and Ball [1996] gives the example of “The memoirs of a retired medical missionary who had an important collection of military paintings, particularly canvases showing the details of early ordnance and regimental uniform; who delighted in the languages and buildings that he had met in his travels, and paid close attention to the level of scientific sophistication in the agriculture of the regions, and the problems of distance from major centres”. Obviously such a document could naturally come under a large number of topic headings.

The problem of where to classify an item is also apparent with one of the most widely used taxonomies of the Internet era, that of Yahoo! (http://www.yahoo.com). In a similar manner to a library classification system, Yahoo! provides access to documents, in this case web pages. Yahoo! has 4 to 8 levels of hierarchy between the main page and the actual content and it has been estimated that it has nearly 70,000 categories. Over 100 people work full time to keep the taxonomy up to date and adding content as and when appropriate. Yahoo! has been incredibly influential in determining the design of internet and intranet portals for the web in content, layout and underlying structure. However, it is an incredibly unwieldy object which attempts to be all things to all men [Morville, 2000].

The significance of these types of taxonomies is that in the commercial world there is far greater demand for taxonomies (such as Yahoo or DMOZ http://www.dmoz.org/) and much less for the formal, logically rigorous ontologies which many in academia are trying to build and promote ([Ellman, 2004]. This is fortunate in certain respects because ontologies are hard to build.

3.4 Ontologies

The word ‘ontology’ has become a buzz word in the past decade for a number of disciplines ranging from the Semantic Web to Knowledge Management. Ontologies hold out the promise of a means to capture, share and re-use knowledge, and are now considered a fundamental
technology because “they interweave formal semantics understandable to a computer with real world semantics understandable to humans” [Ding et al., 2002, :210].

3.4.1 What is an Ontology?

Ontologies are seen in Artificial Intelligence as fulfilling the need to capture knowledge in a manner which would make possible sharing and re-use and because of this they have become a key research topic in Artificial Intelligence in the last decade or so. In many respects, ontologies are building on the work in ‘knowledge representation’ conducted in AI over many years. However, their importance has extended to other domains such as electronic commerce, intelligent information integration, information retrieval and, as we have mentioned, Knowledge Management.

Although derived from philosophy, in AI, the term has come to mean an explicit account of what exists in a specific domain. The most widely quoted definition of an ontology is that provided by Gruber [1993]: “An ontology is a formal, explicit specification of a shared conceptualisation” (cf. also Noy and McGuiness 2001). Because it is a conceptualisation, it is an abstract model of objects in the world. Because it is shared, it is assumed that there is a community of ‘agents’ (i.e. human beings and computers) which accept the definitions and relations specified in the ontology. An ontology is explicit in the sense that the concepts and the relationships between them are fully, explicitly defined and it is formal in the sense that it is machine-readable. The IEEE Standard Upper Ontology working group defines an ontology as “similar to a dictionary or glossary, but with greater detail and structure that enables computers to process its content. An ontology consists of a set of concepts, axioms, and relationships that describe a domain of interest” [IEEE, 2002].

Even if there is a wide consensus that there is a need for ontologies, each of these requirements raises its own set of problems. That an ontology is shared or common to a group of agents is essential for knowledge-based services to function, and yet in the real world it is extremely difficult to reach a consensus in any domain. There are disagreements about what the appropriate terms are, or should be; there are disagreements as to what the terms mean i.e. how the relate to the external world; and there are disagreements as to how to represent these terms (for example, there is a whole range of ontology representation languages). Human beings are notoriously slow at reaching consensus and an ontology represents an ontological commitment to a particular conceptualisation or model of the world. For an ontology to be explicit, every possible logical relation between concepts should in theory be either explicitly specified or logically derivable. This is an old core AI problem, in that, either the representation of the world is so detailed and explicit as to be impossible to construct, or the search space is too extensive for any logical reasoning process to be practical. Since neither of these issues have been solved in AI, all ontologies
to date can only be partial descriptions of a domain and as such are neither complete nor fully explicit. Since ontologies should ideally be formal, they should not have any of the vague and metaphorical properties of natural language which gives it so much of its expressive power and practicality. Gruber notes “For knowledge-based systems, what ‘exists’ is exactly that which can be represented” 1993, :1, a statement which covertly admits that knowledge based systems are limited by what is formally representable.

3.4.2 Knowledge Representation in Ontologies

Ontologies attempt to represent or model human knowledge about the world. However, as Woods [1975] pointed out with respect to semantic networks, in order for this to be achieved a detailed semantics would need to be worked out. Woods particularly focused on the ambiguities present in a simple semantic network (Figure 3) where it is unclear whether the interpretation should be any of the following:

1. a definition of a black telephone
2. “There is a black telephone”
3. “Some or all telephones are black”

The important distinction is one between an instance and a class (i.e. between 2 and 3). When one says, “My old car is a Golf”, then the old banger is an instance of a Golf. But when one says, “A Golf is a Volkswagen”, one is saying that the class “Golf” is a sub-class of “Volkswagen cars”. This is an extremely confusing issue because depending on one’s perspective an apparent instance may be an instance or a class. Thus “Volkswagen” is a generic class for a certain type of car, but it is also an instance of a car manufacturer and a company. A closely related issue is the distinction Woods makes between structural links and assertional links. Structural links are those that define a concept, while assertional ones make statements about the world. If we take Figure 3 to represent the concept of a black telephone, then both links are structural. If we take it to have the meaning in 3. above, which makes a statement about the world, then the ‘SUPERC’ link is structural, while the ‘MOD’ link is assertional. The reality of human cognition is that this distinction is somewhat
difficult to grasp partly because many cognitive links are structural and assertional at the same time. The fact that concept $a$ is a sub-class of concept $b$ is not just a statement about the concepts but also a statement about the world.

For the purposes of this research these issues are side-stepped because we are concerned exclusively with modelling the ‘conceptualisation’ i.e. the inter-relation between concepts. Because of the focus on conceptualisation, the issue of how to represent specific, instantiated facts about individual objects will not be addressed. Our purpose in analysing texts in order to build ontologies is not to ‘understand’ or build a model of the meaning of a specific text. Instead, the intention is to build a model of the concepts which occur in a collection of texts and their inter-relation. Any assertional statements about the world are essentially a side effect, even if we would use a statement to the effect that “Clyde is an elephant” to add ‘Clyde’ as an instance of the class ‘elephants’. In addition, we are assuming that authors represent their ideas accurately in the texts they write. But, in the final analysis, an ontology built from a set of texts is an ontology reflecting that set of texts and makes no epistemological statement about the world.

3.4.3 Types of Ontologies

There are different kinds of, or rather perspectives on, ontologies. Top-level ontologies (otherwise known as foundational ontologies) attempt to provide the general categories at the highest level of an ontology. Aristotle’s ten basic categories and the ‘Plan of Classification’ of Roget (Table 3, p. 35) are both examples. These and more recent efforts (cf. Sowa 2000) are usually the result of a combination of philosophic, linguistic and logical motivations, and are generally top-down impositions of a specific perspective upon the world. Some well-known examples include:

- **Penman Upper Model** (late 1980s): This was developed in the context of a text generation system and was developed in order to establish “a level of linguistically motivated knowledge organization” [Bateman, 1990]. The emphasis on linguistic relevance and practicality is a feature of this work. This work has been extended in the “Generalised Upper Model” [Bateman et al., 1995].

- **Cyc/OpenCyc** (mid 1980s-present): This is an open source version of the Cyc knowledge representation system and consists of “47,000 concepts: an upper ontology whose domain is all of human consensus reality” (http://www.opencyc.com/)(cf. above Section 2.2).

- **Mikrokosmos** (mid 1990s): The Mikrokosmos ontology was developed for a large scale machine translation system. Like the

---

3 This may be a fundamental epistemological difference between the types we are trying to build and the hand-crafted ontologies such as the Gene Ontology which try very hard to reflect the very latest understanding about the world.
Penman Upper Model, it attempted to unify the need for linguistically specific information (i.e. useful for NLP) with the ontological demands of an interlingua [Beale et al., 1995, Mahesh and Nirenburg, 1995].

- EuroWordNet (late 90s): Top-Ontology: This was developed in order to enforce uniformity across the different WordNets, and to allow the linking of EuroWordNet with other ontologies like Mikrokosmos or Cyc [Vossen, 1999, :58-9]4.

- IEE Standard Upper Ontology (2000-2003): This has been developed so as to enable “applications such as data interoperability, information search and retrieval, automated inferencing, and natural language processing” [IEEE, 2002]. It does not include any domain specific concepts but rather provides the structure for such ontologies to be built.

The basic expectation is that such a top-level ontology will enable interoperability between different computer systems and also automated reasoning. It is interesting that considerable effort has been made by the ontology community to build such top-level knowledge bases even though in machine translation the notion of an interlingua has been largely discredited. Furthermore the commitment to relevance to the needs of Natural Language Processing applications apparent in the Penman Upper Model appears to have been lost in more recent efforts.

Domain ontologies are those which are built to cover a specific topic or subject area. Many have been built for specific applications and a great deal are available over the web. Significant collections are held at various sites around the world (cf. resource list in Appendix I). Examples include the “Enterprise Ontology” developed at the University of Edinburgh, the ‘AKT Ontology’ built in the AKT project (www.aktors.org), or the Commerce Ontology (http://www.cs.umd.edu/projects/plus/SHOE/onts/commerce1.0.html).

Navigli et al. [2003], among others, distinguish between ‘core domain ontologies’ and ‘specific domain ontologies’. The core domain ontology may consist of concepts generally characteristic of the domain and may be found in general resources like WordNet, but the specific domain ontology is more specialised. It is this, in their opinion, that is missing as a resource from most application areas, and it is the main challenge in trying to automate the process of ontology building. The problem is acute because many domains of human knowledge are changing so rapidly it is impossible to rely on hand built ontologies.

Some authors have argued for ‘personal ontologies’ [Huhns and Steven, 1999, Chaffee and Gauch, 2000] in the sense that each Web user (or their respective agent) would have their own ontology representing their perspective of the world. There is an inherent contradiction in such a concept given that an ontology is supposed to be a shared conceptualisation. This is further complicated by the fact that different ontologies are constructed.

---

4 The EurowordNet Top Ontology was linked to, but is not identical with, the “Interlingual Index” [Vossen, 1999, :9-10]
individuals have different ideas as to what is shared. People tend to declare that something is a shared view of the world when it is in fact their own personal view imposed on others. On the other hand, if personal agents are to have an internal representation of knowledge, it must be individual to that agent. Kalfoglou et al. [2001] argue that this should be construed more in terms of a ‘personal view’ on an ontology. They do not spell out, however, what this might mean. The current explosion of folksonomies in the contexts of social networking software like flikr, may make the concept of a “personal ontology” more meaningful.

3.4.4 Uses of Ontologies

We attempt here a brief categorisation of different application domains where it is claimed ontologies have a significant role to play.

Knowledge Management: This is a primary area of use for ontologies. Ontologies are considered the appropriate means to facilitate the acquisition, maintenance and accessing of knowledge in organisations of all sizes. More and more of an organisation’s assets are represented in the knowledge in people’s heads and in the texts they produce. The extensive interest companies have in building taxonomies and by extension ontologies is discussed below in Section 4. Within Knowledge Management, there are the following areas of application:

- Searching/Information Retrieval. A number of authors have shown that ontologies can significantly improve search results on document collections. This can be done in a variety of ways. The ontology or thesaurus can be used for domain specific query expansion (e.g. Järvelin and Kekäläinen [2000], McGuinness [1998]). Or else the meta-tags on documents can be used to identify whether the document is in the correct ontological category with respect to the query.

- Information/Knowledge Extraction. Ontologies play a growing role in information extraction both in providing templates for the extraction process and in providing an appropriate knowledge base for extracted information to be organised and thus considered knowledge [Brewster et al., 2001]. Recent work uses an ontology to extract facts from biographical texts on the web and thus populate a knowledge base for subsequent publishing of the biographies on demand [Alani et al., 2003].

- Knowledge Publishing. Ontologies are now being used in the automatic generation of a variety of documents depending on user needs/preferences, for example the MyPlanet system of KMI [Kalfoglou et al., 2001]. This uses the links in the ontology to allow appropriate news stories to be published in accordance with the interest of the user.

5 We can see this in the complexity of the process by which international standards are established e.g. by ISO or IEEE.
• Knowledge and Research. The representation of a domain of knowledge in an ontology allows the user to browse, navigate through topics, discover what they do not know, and provides in the academic environment a powerful research tool for knowledge exploration in general.

Agent Based applications: The growth of the Semantic Web is conditional on the development of a multitude of ontologies [Berners-Lee et al., 2001]. Domain ontologies are crucial, both for the Semantic Web and for web-based applications in general. Two particular areas of development may be singled out:

• Personal assistants: These are software agents usually associated with a particular individual. In general they can gather and filter relevant information and, in the extended scenario proposed by Berners-Lee et al. [2001], they will make appointments, suggest schedules and plans, while communicating with other agents about significant events. These agents will need to have ontologies to provide internal representations, and most probably external ontologies to translate concepts between domains or topic areas. Already in closed or almost closed agent systems ontologies have been used repeatedly [Kumar et al., 2002, Chalupsky et al., 2002].

• Shopbots: A specialisation of the above is the use of agents to undertake a search and retrieval of items corresponding to something one wishes to buy, finding the best deal. This will become easier once both the personal agents and the web-based shops have common ontologies. Currently, wrappers have to be written which both very costly in terms of effort and the resultant information is of poor quality [Fensel et al., 2001].

Commercial Application: Ontologies are important to commerce for a number of reasons. Ontologies provide a means to help standardise knowledge architecture thereby making it easier to implement electronic commercial activities. The main impediment to the development of e-commerce is the inability of systems to share information. This is apparent when large companies try to merge the knowledge resources. If a number of trading partners are trying to interact, a single domain specific standard for content and transactions is vital.

Ontologies or more accurately taxonomies already play an important role in many internet shopping sites. These taxonomies allow the user to browse and navigate through often extremely complex catalogues of products (cf. for example the browse facility on www.amazon.com or the products hierarchy on www.pcworld.co.uk).

• Business to consumer: The shopbot mentioned above is a clear business to consumer situation where ontologies are important, but there are also other areas such as on-line auctions,

• Business to business: The use of ontologies in business to business transactions has already been shown to facilitate processes enormously. Complex business product taxonomies exist which play
an important role for automobile companies to request part manufacturers to bid for a tender. Standard ontologies already exist for many business areas, for example Commerce XML (http://www.cxml.org/), Real Estate Transaction Markup Language (http://xml.coverpages.org/rets.html), RosettaNet (http://www.rosettanet.org).

The above list is merely indicative because in fact there is a plethora of initiatives. Ironically, while all these initiatives at some level are about standardisation, there are too many initiatives to make standardisation a likely outcome.

3.5 Capturing the World

We have already noted above that one dimension along which thesauri, taxonomies and ontologies may be said to differ is in the objects which they describe. Thesauri are concerned with words, taxonomies with real world objects and ontologies with concepts. This distinction is artificial because in all cases the objects are mediated and discussed with words and no one is able to pick up a concept without using words. Furthermore, depending on your perspective, words or concepts relate to real world objects i.e. they “bottom out.” A more significant distinction is one of logical rigour i.e. thesauri are mere loose associations of words, taxonomies impose a certain degree of explicit structure while ontologies (ideally) specify the exact logical relations between all concepts so as to allow ‘Problem Solving Methods’ to apply over these knowledge structure (cf. Figure 4).

There are three major challenges for research in ontologies today. The first, which is the focus of this work, is to try to automate the construction process as much as possible. Given a continuously changing world, where our knowledge and understanding of it in constant flux, it is imperative to find ways to build ontologies with the minimum of human effort (if this is possible). The second challenge concerns the degree of rigour and logical detail. Manual methods can of course be detailed and logically coherent (see for example the instructions in Gómez-Pérez 1999), but automated methods will be rough and ready and very poor output for logical inference and ‘problem solving methods’. Thirdly, if knowledge changes so rapidly and ontologies need to be updated, it is very probable that we will encounter the
Frame Problem [McCarthy and Hayes, 1969].

The Frame problem was originally identified as a problem for a robot, which has to update its internal representation of the environment each time something happens. Events occur that change the state of the world and each time the robot’s or system’s internal representation of the world must change. The problem is that the consequences of an event are unpredictable. Thus, if a box is moved, some object may be on top of it. The movement of the box is not the only event to take into account because the movement of the other object must be accounted for as well. Yet again, one cannot easily predict this by rule (e.g. every moving object takes with it objects on top of it) because an object may be tied to the wall with string or there may be some other factor preventing the normally expected physical consequences. Clearly, a system could check every item in the world to determine if a change had (or could have) occurred or not, but this would be quite unrealistic in even a reasonably large toy world system, let alone the one we live in. As the model of the world or ontology increases in complexity and approaches a more realistic representation of reality, so the problem of calculating the effect of any change in the representation becomes exponentially large.

This research attempts to contribute to the first of these issues (ontology construction), although the other two (logical rigour and the Frame Problem) will be considered where appropriate.

---

Knowledge is and will be produced in order to be sold, it is and will be consumed in order to be valorized in a new production: in both cases, the goal is exchange.

Jean Francois Lyotard [1984]

We will present two perspectives on the current industrial context. On the one hand, there are the facts of current practice in a variety of industries where Knowledge Management is seen as important and where the use and importance of ontologies and thesauri have been recognised. This is perhaps less important from a research perspective, but it does give insights into corporate culture and perceived needs which we believe are significant. On the other hand, there are a number of companies and their offerings in the fields of Knowledge Management, classification software, text mining and information visualisation which are of varying degrees of relevance. We will attempt to delineate some of the more interesting examples.

As Ellman [2004] notes, the overwhelming majority of ontology type structures required by companies are taxonomies. He notes that only two out 22 ontologies built over three years had “inference requirements” – all the rest were subject catalogues, content management systems or text categorisation schemes. In effect, all cases involved using the taxonomy to navigate the user to a required item. There has also been in recent years a huge increase in the appreciation of taxonomies in the business environment. A Delphi Group study showed 50% gain in the proportion of people in business who viewed taxonomies as important to their business strategy. 40% of respondents considered taxonomies very important or imperative to their business strategy [Group, 2004]. Furthermore, in 2002 the US government passed the “E-Government Act” which obliges US government institutions to ensure the accessibility, usability and preservation of government information and specifically obliges the government departments to construct taxonomies. This has made KM and content management companies have an even larger potential market while at the same time ensuring a stable market for their products. Already companies like Convera, Vivisimo and Entrieve are specifically targeting their products to target this market (Convera 2003, Vivisimo 2004 Entrieve 2004). All these are factors which have emphasised the importance of taxonomies in the commercial environment.

4.1 PURPOSE

Excluding the knowledge base potential of ontologies/taxonomies, there are a number of reasons why companies construct and use them
and thus the situation can be very complex. The consultancy TPFL, which specialises in providing companies with advice in this area, says that a taxonomy ‘aspires’ to be the following things:

1. a correlation of the different functional, regional, and (possibly) national languages used by the enterprise.
2. a support mechanism for navigating around, and gaining access to the intellectual capital of an enterprise.
3. a combination of tools such as portal navigation aids, support for search engines, knowledge maps, as well as an authority for tagging documents and other information objects
4. a knowledge base in its own right

The first of these refers to the correlation and standardisation of terminology within a company and across its different subsidiaries. The second item reflects the role of a taxonomy in providing support for the various interfaces mentioned in item 3. most typically its use in the intranet portal (à la Yahoo!) or as part of a search engine. Item 4. which is potentially the most significant from a Knowledge Management perspective is also the least developed in the commercial environment, a fact which is also reflected in the use of the term ‘taxonomy’ rather than ‘ontology’ when describing the knowledge bases most companies have.

4.2 CURRENT PRACTICE

4.2.1 Building the Taxonomy/Ontology

It is clear from a number of sources that the overwhelming majority of industry taxonomies and ontologies are hand built. Consider this extract for example from a current job advertisement on Dice.com:

Duties: The Ontology Manager will hold a key role in meeting customer demands by maintaining the master ontology that organised content for the eBusiness initiatives. This individual will ensure the data is organised to facilitate rapid product selection . . .

Some of these taxonomies (as the figures quoted in Section 3.3 concerning Yahoo show) involve considerable human resource input in the companies concerned. In order to successfully manage a complex knowledge network of experts, the Minneapolis company Teltech has developed an ontology of over 30,000 terms describing its domain of expertise. Teltech employs several full-time ‘knowledge engineers’ to keep this ontology current adding 500-1200 terms per month [Davenport, 1998]. Another typical example is that of Arthur Andersen who have recently constructed a company wide taxonomy entirely
by hand. Their view of the matter is that there is no alternative because the categories used come from the nature of the business rather than the content of the documents. This is paralleled by the attitude of the British Council’s information department who view that the optimum balance between human and computer is 85:15 in favour of humans. Not all companies perceive human input as so sacrosanct; Braun GmbH for example would appreciate a tool for taxonomy creation and automatic keyword identification [Gilchrist and Kibby, 2000, :34]. One of the earliest exponents of Knowledge Management, Price-WaterhouseCoopers consider that “the computer can enable activity, but Knowledge Management is fundamentally to do with people” (ibid.:118).

One manner in which certain companies reduce the manual effort involved is by using ready-made taxonomies provided by others. An example of this is Braun GmbH whose internal taxonomy is based on the (hand-built) resources provided by FIZ Technik (a technical thesaurus, cf. http://www.fiz-technik.de/) and MESH (the medical subject headings provided by the US Library of Medicine, cf. http://www.nlm.nih.gov/mesh/meshhome.html). Nonetheless about 20% of the vocabulary in the taxonomy is generated internally to the company. Another example is the case of GlaxoWellcome who have integrated three legacy taxonomies derived from company mergers using a tool called Thesaurus Manager developed by the company in collaboration with Cycorp, Inc.

4.2.2 Maintaining the Taxonomy/Ontology

Most companies which build a taxonomy manually also maintain it in the same manner. The effort involved varies enormously depending on the size of the company and the ambitions behind the project. At one end of the scale, the British Council and Ernest & Young makes no separate allocation for the task and at the other a company like GlaxoWellcome have a team spread across several countries up-dating the taxonomy. In general, taxonomy maintenance is not very demanding for companies which have relatively stable business activities and who are not involved in the details of rapidly changing technological domains. It is difficult to assess often how much effort is going into changing the taxonomy as opposed to associating content to the categories. But many companies see the taxonomy as “a living reflection of the organisation and the environment in which it lives, both of which are in a process of continuous change” [Gilchrist and Kibby, 2000, :98].

4.2.3 Associating Content to the Taxonomy/Ontology

In companies like Arthur Andersen, the association of tags and metatags to documents is also entirely performed by hand, even though there are tools which could perform this. The reasoning be-
hind this practice is that such tools would not be trusted by the users of the system [Gilchrist and Kibby, 2000, :27]. In contrast the BBC has a system where the taxonomy has been built by hand but the association of a term with a document is performed by a software tool (Smartlogik’s Infosort). Each term is positioned with respect to broader, narrower and related terms in the taxonomy and has associated hand-crafted rules which trigger the association of a document with a certain term. A central team maintain both the taxonomy and the associated rule-base (Ibid.:38). A similar approach is being taken by Microsoft in the development of its company internal taxonomy. It is a commonly held assumption that given the existence of a taxonomy or ontology, authors will be willing to tag their own work in an appropriate manner [e.g. Stutt and Motta, 2000, :218] but the experience of both librarians historically and more recently companies like ICL and Montgomery Watson is that authors tag inadequately or inappropriately their own work.

4.2.4 Confidence and Evaluation

There appears to be an overwhelming dominance of manually developed taxonomies and ontologies. There are two different strands to the companies’ reactions to the possibility of automating the process. Some potential users of such a system would welcome the reduction of time and effort and would particularly appreciate the system proposing new categories and keywords. The other strand is to be found in the lack of confidence people have that such a process can in fact be automated and that the right documents will be associated with the right categories. Furthermore there is considerable emphasis on the way the company’s internal taxonomy reflects their particular perspective on the world, their philosophy and their priorities. No automated system can automatically reflect that and thus the process of manual analysis and coding is inevitable. It would appear that a substantial change in attitude will have to take place for automatic taxonomy/ontology creation tools to be accepted by some companies. This is particularly true of companies which are acutely aware that it is in their accumulated knowledge and experience in which their shareholder value lies. An essential step in this direction is to show the qualitative performance of a system and to make possible the evaluation of the performance of different systems.

Even though there is a great deal of discussion of taxonomies within business, people recognise that the lack of evaluation makes it difficult to convince companies to invest in this type of technology. Rita Knox of Gartner argues that taxonomy projects need to be tied to very specific business goals and notes that lack of a taxonomy can hurt business productivity (cited in Berlind [2004]). However, in general the lack of evaluative mechanisms, whether in terms of return on investment or more precise measures are an abiding problem for all knowledge management technologies.
4.3 CURRENT PRODUCTS

In this section, we attempt to briefly survey current commercial offerings of automated or semi-automated taxonomy/ontology tools. What is written below is based largely on the publicly available documents about a company (web-site material, whitepapers, etc.) occasionally supplemented by comments from journalists or analysts where these have been found.

**DataChannel: the 'XML' solution** ([www.datachannel.com](http://www.datachannel.com)) This company provides a system which is centred on exploiting the advantages of XML and making the process of ontology creation and maintenance as easy as possible. Thus it states that “Using an intuitive folder metaphor, administrators and users alike can quickly create categories in a hierarchical manner that best fit organizational information descriptions and semantics.” Once the categories have been set up by an administrator, then it is possible to associate documents with those categories based on certain parameters. The latter technology is licensed from Excalibur (cf.). DataChannel has been a pioneer in the use of XML but has also been criticised for an over-dependence upon this technology.

**Semio: the 'Computational Semiotics' solution.** ([www.semio.com](http://www.semio.com)) Semio\(^1\) claim to implement a ‘computational semiotic’ approach to taxonomy construction. Their patented approach “extracts all relevant phrases from the text collection. It builds a lexical network of co-occurrences, clustering related phrases and enhancing the most salient features of these groupings. The patterns built by SemioMap reveal the conceptual backbone of the text collection, which is displayed in the concept map.” One of the important emphases of Semio is on phrases of varying length as the basic building blocks for their taxonomy structures. It is unclear from their literature whether their claims of ease of integration with a company’s top level taxonomy implies an inability to actually construct a labelled taxonomy automatically. Semio have a number of existing taxonomies which they can offer clients. Their main claim is that they can automatically cluster terms extracted from documents and then attach them to some taxonomy. The clustering technique is based (according to the patent) on whether two words are more likely to occur together or more likely to occur apart. If they are more likely to occur together they are clustered together. In spite of the claim that Semio is able to construct taxonomies, in reality the top levels and subdivisions are created manually, and it is only at lower levels that directories are generated automatically from textual content (Adams 2001). This may reflect the limits and necessities of the technology, however.

**Autonomy: the ‘Bayesian Networks’ solution.** ([www.Autonomy.com](http://www.Autonomy.com)) Although Autonomy have made immense claims for their prod-

---

\(^1\) The company has now changed its name to Entrieva ([www.entrieva.com](http://www.entrieva.com)).
ucts as a solution to corporate portal needs and all Knowledge Management requirements of companies, observing their software in action, they speak of editorial input for categories and the construction of ‘agents’ which are ‘trained’. In effect, the agents are the categories and the training is what permits a pattern match to occur and thus to identify relevant documents. Consequently Autonomy’s software can only be seen as a combined manual and automatic process where the automatic part lies in the ability to associate documents with a certain category once it has been established and data concerning that category fed into the system. The ‘retrain’ button for agents is indicative of this approach. Clearly, there is no capacity to construct either taxonomies or ontologies of any form. With reference to both Autonomy and Semio, Morville comments that they produce “truly confusing category schemes with tremendous redundancy and mixed granularity” [Morville, 2001].

**Excalibur** ([www.excalibur.com](http://www.excalibur.com) now [www.convera.com](http://www.convera.com)) Excalibur claims to use a ‘unique semantic network’ in their Knowledge Management software; More particularly they state: “The very basis of Excalibur’s Semantic Network, a built-in knowledgebase of word meanings and word relationships, is that it is a framework for reflecting and applying knowledge. Excalibur’s baseline networks of English, French and Spanish, and its specialized subject networks for medicine, the law and defense, reflect linguistic knowledge, derived from published lexical resources, such as dictionaries and thesauri” [Excalibur, 1999]. This implies a pre-built taxonomy of a certain size (and in this regard is probably similar to the approach of Semio). These semantic networks, it is claimed, can be tailored to ‘vertical-markets’. Much like Autonomy, the user can establish ‘agents’ which will automatically associate documents with a certain query. They claim that they achieve superior results to Boolean searches using ‘plain English and a high level of semantic expansion’, but no independent evaluation has been performed to our knowledge.

**Interwoven’s Metatagger** ([www.interwoven.com](http://www.interwoven.com)) is a product whose main focus is on automatically adding XML tags to documents. This ‘metadata’ is added to all documents in the repository and can be quite complex in nature. A detailed schema has to be written to determine what metadata to capture for the content. The system uses extensive named entity recognition as part of its tagging process, and then uses pre-classified text to train a text classifier. Existing taxonomies or word lists can be imported into Metatagger in XML format. The company claims the software is able to generate taxonomies automatically from the content of a document collection and then permit the user to review this using visual tools. However, as is usual, no information is available concerning the techniques used to achieve this. Perhaps the most interesting aspect of this company is that it stresses the need for the user to remain in the loop because “no completely automated content intelligence product can consistently deliver highly
accurate metadata. Computers and software simply do not have the human judgement and knowledge to consistently deliver the content relevance users want and expect” [Interwoven, 2002].

Mohomine (www.mohomine.com) provides a classic text classification solution which essentially assumes a manually built taxonomy and 15-20 texts which are already associated with each category. From this data, using machine learning it is able to classify text with high degrees of accuracy (they claim over 98%). No attempt is made to construct a taxonomy or maintain an existing one. They lay great stress on the use of XML in order to make the output of their system uniform, irrespective of the format of input text files.

The major issue at present with commercial offerings is that there exist no means to assess the validity of competing claims, no means to evaluate a product with respect to a specific set of user needs. This evaluation problem is compounded by the fact that often very little is known about the internal workings of the software product on offer.

There is common consent among KM experts that one of the key flaws of all such methods, actual or potential, is that they are ‘content-centric’ i.e. they are able (or trying) to organise documents in accordance with the data present in the texts. Some consulting firms have the category ‘business opportunity’ in their taxonomy, and Reuters have the news category ‘Oddly Enough’. In neither case could these categories be extracted from the text associated with this node, nor could a document classifier be easily trained to recognise such documents. “If you don’t understand the goals and strategy of the business, how can you organise content to further those business objectives?” [Morville, 2000]. Thus for the foreseeable future, commercial software will have to integrate both manual and automated input [Adams, 2001]. The only software developer who fully acknowledges the role of the user appears to be Interwoven.
Knowledge, a rude unprofitable mass,
The mere materials with which wisdom builds,
Till smoothed and squared and fitted to its place,
Does but encumber whom it seems to enrich.
Knowledge is proud that he has learned so much;
Wisdom is humble that he knows no more.

William Cowper - The Task

Current methods for building an ontology or taxonomy may be considered closer to a craft than a science. A great many decisions are based on experience previously accrued, and on ‘instinctive’ ‘intangible’ knowledge concerning the appropriateness of the inclusion of a concept, its particular encoding, whether it is a class or an instance etc. In this as in so many other ways, building an ontology resembles lexicography. Noy and McGuiness [2001] describe the following steps in building an ontology by hand:

1. Determine the domain and scope. This step involves identifying the domain covered by the ontology and what it will be used for. It is useful to develop a set of questions which the ontology should be able to answer (e.g. “Does Cabernet Sauvignon go well with seafood?”). They also note that if the ontology will be used in as an NLP component then the inclusion of synonyms for concepts and other linguistic knowledge has to be decided on.

2. Consider reusing existing ontologies. They argue strongly that it is beneficial to re-use an existing ontology for a domain if one is available, and just adapt it to the specific needs one has. There are several repositories for publicly available ontologies (cf. Appendix I).

3. Enumerate important terms in the ontology. This involves making a list of all significant terms in the chosen domain, together with their properties. It is not important at this stage if terms refer to overlapping concepts.

4. Define the classes and class hierarchy. The class hierarchy can be developed either top-down or bottom-up, or even a combination of both. The latter approach appears to be the most natural as middle level concepts are the most salient. The hierarchy is assumed to be an ‘is-a’ hierarchy where each child is a sub-class of its parent.

5. Define the properties of the classes - slots. Slots are properties associated with a particular class, such as colour, body, flavour with respect
to the class/concept wine. All subclasses inherit the slots of its super-class, which means the property/slot has to be positioned on the highest class which takes that property.

6. Define the facets of the slots. The facets of a slot are the different values applicable to that slot/property. Thus for each slot, the cardinality (number of values), the value type (e.g. string, number, instance, etc.), and the range (allowed classes) must all be determined.

7. Create Instances. This involves taking a particular class and filling in the slot values so as to create a particular exemplar.

Noy and McGuinness stress that the whole process, but especially steps 4 and 5, is iterative and the ontology changes as it is developed and built up. The interesting issue is to determine how much of this process can be automated. Step 1 is clearly a human decision. Step 2 could be partially automated if Step 1 first identified a specific corpus which accompanied the domain, and if Step 3 were to be automated to identify the key terms in the given corpus/domain.

Step 3 is a process where term recognition is the key technology. A term is the “designation of a defined concept in a special language by a linguistic expression” [ISO 704, 1987]. Terms are single or multi-word units (such as digital printing technology) differing from ordinary words in having a specific definition (often explicitly stated) within a given domain. Identifying the terms characteristic of a domain is a vital step to identifying the “shared conceptualisation” of the domain. Most ontology-building systems have an explicit term extraction phase [Morin, 1999b, Agirre et al., 2000, Navigli et al., 2003].

The main problem with term recognition is that the majority of the approaches which are most successful are dependent on extensive tools such as taggers and parsers which in themselves need to be adapted to the specific characteristics of the domain texts. It is for this reason that certain authors choose to use term recognition tools which are ‘knowledge-lite’ [Merkel and Andersson, 2000, Ahmad, 2000].

Steps 4-6 above are the heart of the process and it is here that we focus our attention. In using NLP techniques, it is not yet clear how to differentiate classes and properties of classes. Noy and McGuinness provide guidelines for deciding when to introduce a new class or just to use properties on existing classes: “Subclasses of a class usually (1) have additional properties that the superclass does not have, or (2) restrictions different from those of the superclass, or (3) participate in different relationships than the superclass” (2001:16). However, making such choices or evaluations is clearly an editorial interpretative call which cannot be automated as yet. In the following sections, we discuss these stages from a language technology perspective.
5.1 CORPUS SELECTION

The task corresponding to Step 1 (above) is that of selecting a suitable corpus. Once a domain or topic area is chosen, the next step in order to permit the automation of the process lies in selecting a suitable coherent and constrained corpus. The more carefully selected it is the better the overall results will be. Many NLP techniques use machine learning where there must be absolute congruence between the training texts and the texts where the application is applied, otherwise results will be extremely poor. Care must be taken because texts can appear to belong to a similar domain and yet this may not be the case. For example, Ciravegna\(^1\) cites the example of training a system on US CVs and then using it on European CVs and finding that the system was unable to perform adequately. A closely related issue is the selection of a reference corpus, as certain methods depend on the comparison of the behaviour of terms in two different corpora. The usual default choice is the BNC (http://www.hcu.ox.ac.uk/BNC/) or some such similar large scale, general purpose corpus which can claim to be representative of the whole language. However, in certain cases, because terminology is so specialised and inevitably absent from the general purpose corpus, it may be more appropriate to use a corpus of intermediate degree of specialisation.

5.2 ASSOCIATING TERMS

We can imagine an abstract segment of an ontology/taxonomy as depicted in Figure 5. We assume that we have applied some form of term recognition so as to identify the terms to be used as candidate classes/concepts (here a,b,c,d,e). From the collection of terms identified it would not be possible to distinguish terms which must be associated with classes i.e. class labels, and those which refer to properties or slots. At this stage those terms which most strongly associate together must be identified. There are a number of methods available for this (cf. below Section 6.1) but these in most cases provide no information as to whether a term is a subclass or superclass of another.

Figure 5. A fragment of an abstract ontology/taxonomy

--
\(^1\) Personal communication.
5.3 BUILDING A HIERARCHY

Given information about which terms are associated with each other, the next step is to organise the terms into some form of hierarchy. This will be, in the majority of cases, an is-a hierarchy where the primary relation between terms is set inclusion or hyponymy (depending on your perspective). The important matter is to provide a certain degree of depth to the organisation of the terms, otherwise the taxonomy will resemble a dictionary or other such shallow word list. In such a case, a considerable amount of information will either have been lost or not captured. While a handful of methods exist to construct hierarchies of terms, the majority are unable to suggest candidate terms for the nodes (e.g. a in Figure 5).

5.4 LABELLING RELATIONS

The biggest challenge lies in establishing the relationship between terms in such a knowledge base. These have been described as the “non-taxonomic conceptual relationships” by Maedche and Staab [2000a]. The default can be taken to be hyponymy but limiting a system to this would a misrepresentation of the knowledge concerning a domain. The classic ‘tennis problem’ with respect to WordNet is a case in point [Stevenson, 2002]. Because terms are placed strictly by hyponymy in WordNet, the terms concerning tennis are in disparate parts of the knowledge base. Thus it is not possible from WordNet to answer questions such “What do I need to play tennis?” or “What does one hit a tennis ball with?”. Another way of looking at this is to observe that WordNet does not have conceptual class clusters like Roget does. A large number of taxonomies contain overtly or covertly other relations such as meronymy, or other quite arbitrary ones specific to the domain. In many cases, such relations might be modelled as slots on a class rather than subclasses exhibiting peculiar relations with their super-class. This is dependent on how logically rigorous the knowledge base being created is and how sophisticated are the means for encoding. However, it is our opinion that there should be variety of types of relations between terms to enable the representation of a full range of knowledge relationships. For this to be possible, the challenge lies in finding ways to identify the type of relationship between two terms and to label that relation. This is the most difficult task and the one needing the maximum of human intervention. The reason for this is that so much is mediated by world knowledge that is not encoded in texts.

5.5 METHODOLOGICAL CRITERIA

In this section, we consider a number of criteria to be used when choosing methods with which to process texts and produce taxonomies or components of taxonomies as their output. Our purpose here is
twofold. First, we wish to create a set of criteria in order to help guide the choice of appropriate tools to use in the automatic construction of taxonomies. While there are a large number of methods which might conceivably produce appropriate output, in fact only a subset will actually fulfil these criteria. Secondly, we hope thereby to contribute to means by which different approaches to constructing taxonomies can be evaluated, as there is a complete dearth of evaluative measures in this field. Writers on ontology evaluation concentrate on a limited number of criteria which are only appropriate to hand-crafted logical objects [Gómez-Pérez, 1999, Guarino and Welty, 2000].

5.5.1 Coherence

A basic criterion is one of coherence, i.e. that the taxonomy generated appears to the user to be a coherent, common sense organisation of concepts or terms. There are, however, many ways in which terms or concepts are associated with one another. The term ‘grandmother’ is associated in each person’s mind with specific images, ideas, concepts and experiences. But these specific cases are not universal even for a subgroup and thus would not make sense to a third party. Coherence is dependent on the terms associated in an ontology and the nature of their association being part of the ‘shared conceptualisation’ Gruber described.

Here it is important to distinguish linguistic from encyclopaedic coherence: in a thesaurus such as Roget [1852] under a specific category (e.g. smoothness) we encounter a collection of synonyms of varying degree of closeness. Here we encounter linguistic coherence in the sense that the grouping ‘makes sense’ given linguistic criteria. A good example of this Wordnet, which organises a large vocabulary according to a linguistically principled hierarchy. However, it does not provide a useful organisational principle for information retrieval, reasoning or Knowledge Management in general. It is a linguistic resource much like a dictionary is. But in a taxonomy or browsable hierarchy, we find concepts or terms are organised for the purpose of finding firstly relevant subcategories, and secondly specific web sites. Thus in Yahoo under Education → Higher Education → Universities we find a list of universities not a list of synonyms for the concept university. Linguistic coherence can be expected to be much more stable over time than encyclopaedic coherence, partly because language changes relatively slowly, and partly because our knowledge or understanding of the world tends to be revised rather dramatically in the light of social and cultural influences.

The ‘Tennis Problem’ mentioned above (Section 5.4) is the reflection of a particular set of expectations, a particular viewpoint on how terms should be organised. It would be reasonable to argue that it is an encyclopaedic fact about the universe which brings that set of terms together i.e. reality determines that cluster of terms. However, there are many ways of cutting reality and the particular perspective is specific to a user or group of users. It follows therefore that the notion
of coherence must in effect be a specification of user requirements in terms of their unique perspective on the knowledge represented in the ontology.

Given these observations, the notion of coherence must be understood as being application specific in the sense that it is in designing a specific context of usage for an ontology that we can have a clearer idea of the users and their particular perspective or model of the world. For our purposes in constructing taxonomies for Knowledge Management, in general, the notion of encyclopaedic coherence is primary while linguistic coherence can only play a secondary role depending on the needs of an application and on the extent to which (for example) a specific term is referred to by a number of other synonymous ones. The hierarchical structures generated must maximally be sensible, useful and representative of the associations and juxtapositions of knowledge which human users actually need and make.

Having made this seemingly uncontroversial proposal, it is in fact very difficult to evaluate a taxonomy or hierarchy from this perspective. Given a method, given a specific input and output, there are no widely established criteria for deciding that a particular taxonomy is correct or incorrect, or that one is better than another. While, in fields like information retrieval, we can speak of precision and recall, there are no equivalent measures for an ontology or taxonomy. This is because knowledge is not really a quantitative entity, it is not something that anyone has come up with easy ways to measure (witness the controversies surrounding exams in education). Coherence as conceived here is a qualitative parameter which as yet merely begs the question for its evaluation.

5.5.2 Multiplicity/ Multiple Inheritance

By multiplicity, we mean the placement of a term in multiple positions in the taxonomy. The criterion of multiplicity needs to be distinguished from semantic ambiguity. There are clearly a large number of terms which are ambiguous in that the have a number of separate definitions. Obvious examples include terms like class, which has an entirely different meaning in the domain of railways, sociology and computer programming. A more extreme example is the word crane referring to a bird and a mechanical object. These might be distinguished as in some dictionaries by means of a subscript: word₁, word₂, word₃, etc. On the other hand, there is often a multiplicity of facets for one single term which justify its multiple placement in a taxonomy or ontology depending on the particular focus of that sub-structure. A cat is both a pet and a mammal. A research assistant is both an employee and a researcher, and often a student as well. This is a classic problem in librarianship where a book is often concerned with a multiplicity of topics and the necessity in traditional library classification schemes (Dewey, Library of Congress) to place a book under one class mark (which determines where it will be physically placed) caused much controversy and anxiety. Similarly, many concepts can be placed in
different positions in a taxonomy depending on the particular facet of the concept one is interested in or emphasising. Thus, to take a simple example, the term *cat* clearly has its place in a taxonomy of animals from a purely zoological perspective. It is also obviously a pet but the category of pets does not in any way fit in with the classification of animals. Many other obvious examples can be given of this.

As a consequence, the methods of processing texts that has to be used must allow *cat* to occur both in association with the term animal or mammal and also in association with pet. They must take into account that support can occur in different senses in the same context. At a minimum, methodologies which force a unique placement for any given term should be avoided. Better still, we need to identify methods which take into account the different senses of a term. Noy and McGuinness term this phenomenon ‘multiple inheritance’ and give it a recognised place in their methodology. The problem really arises due to the bluntness of automated methods.

5.5.3 Ease of Computation

One of the major issues in Knowledge Management is the maintenance of the knowledge bases constructed. As has already been mentioned, an ontology or taxonomy tends to be out of date as soon as it is published or made available to its intended audience. Furthermore, from the developer’s and editor’s perspective it is important to have output from the system as quickly as possible in order to evaluate and validate the results. In many contexts, there is a continuous stream of data which must be analysed and where each day or week represents an extraordinary large amount of data whose effects upon the overall ontology cannot be determined a priori.

Thus it appears to be very important that the methods chosen to automatically generate ontologies (or candidate ontologies) do not have great complexity and therefore excessive computational cost. This may appear to be an unimportant issue in this time of immense and cheap computational power but, when one realises that some algorithms have a complexity of $O(V^2)$ where $V$ is the size of the vocabulary in the text collection, then it can be seen that this is not an insignificant factor in the selection of appropriate methods. The practical significance of this is that in some application contexts computational complexity needs to be seriously considered. There are of course other contexts where it is much less of a concern (where the quantity of data is limited).

5.5.4 Single labels

Another criterion for the selection of methods is that all nodes in a taxonomy or hierarchy need to have single labels. Sanderson and Croft [1999] discuss the difference between polythetic clusters (where members of a cluster have some but not necessarily all of a set of
features) and monothetic clusters (where all members of the class are guaranteed to have that one feature). They argue that clusters characterised by one feature are much more easily understood by the user. For example, a well-known approach developed at the Xerox Palo Alto Research Center was called Scatter/Gather [Cutting et al., 1992] where documents would be organised into hierarchies and a set of terms would be extracted from the documents to characterise each cluster. A group of documents might be characterised by the following set of terms:

{battery California technology mile state
recharge impact official cost hour government}

While this is comprehensible, it is not very easy to use and it is discouraging for most users [Sanderson and Croft, 1999, :1]. If Yahoo! at every level of its taxonomy were to label a node by a large collection of terms associated with the topic considerable confusion would be caused. Thus, in order to be easy to use, nodes in a taxonomy need single labels even if this is a term composed of more than one word. This does not mean that synonyms are not to be included, but this is different from using a set of disparate terms to characterise a subject area. Synonyms can act as alternative labels for a particular node.

Methodologies which produce single labels for a node are to be preferred to ones (such as Scatter/Gather) which produce multiple labels for a node.

5.5.5 Data Source

The data used by a specific method needs to be of two sorts. First, documents must be used as the primary data source for the reasons mentioned above (Section 2.4). Secondly, it should permit the inclusion of an existing taxonomy (a ‘seed’) as a data structure to either revise or build upon as required.

Ontologies and taxonomies are often legacy artefacts in an institution that they may be the result of years of work and people are loath to abandon them. As mentioned above (Section 4.2), companies often merge and as a result two different company’s taxonomies need to be merged. These existing data structures need to be maintained. Furthermore, many institutions view their taxonomy as reflecting their own world-view and wish to impose this particular perspective at the ‘top-level’.

Given these constraints, methods need to be used which take as input documents, but which also have the possibility of using an existing taxonomy or ontology as part of their input and to use the documents to propose additions or alterations to the existing taxonomy. This is essential, of course, from the perspective of maintaining a taxonomy. From a practical perspective, given the existence of many taxonomies for one purpose or another, the use of ‘seed’ taxonomies will be predominant.
EXISTING APPROACHES

Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it.

Samuel Johnson [Boswell, 1791, Entry for April 18th, 1775]

In this section, we review a number of methods existing in the literature which we believe to be relevant to the automation or semi-automation of the process of ontology construction. Often these are methods developed for reasons or applications far removed from our current objectives, but finding new uses for existing methods is a common method of progress in science.

6.1 ASSOCIATING TERMS

Assuming that a suitable means of term recognition has been used and all the relevant terms for the domain have been isolated in a corpus, the next significant step is to discover for each term those with which it strongly associates. At this stage, the association of terms makes no claims as to hierarchical structure, nor whether the terms will finally be coded as concepts in the ontology or as slots.

6.1.1 Simple Word Association

The role simple word association plays in building an ontology or taxonomy is that associating one word with another is the weakest step in building an ontology. When more sophisticated methods of labelling the relation between two terms fail, it can be useful to know that there is a greater ‘stickiness’ present. This can be used for example in semi-automatic situations where candidates are presented to the user for validation.

Associated Words

Scott [1998, 1997] has proposed that it is possible to derive a set of associated words for a given word using a variation of the tf.idf measures first proposed by Salton and McGill [1983]. Thus, by comparing the frequency of a word in a given document with its frequency in a general reference corpus, it is possible to determine whether this is a key word or not, i.e. a word with a proportionately higher frequency. It is possible to construct thereby a set of key words for any given document. By extension, we can analyse all documents where a given word is a key word and identify the set of key words which are key in
all these documents. These we call the key-key words. The associates of any given key-key word are those key words which co-occur in a number of texts. The method in detail is as follows:

1. Initial input: Two corpora consisting of the the main corpus to be analysed, and a reference corpus usually of considerable size and ideally representing the language as a whole. Often the BNC is used as a reference corpus.

2. Frequency List: This is a list of words or phrases which are obtained from the raw text and are ordered by the frequency of occurrence. Such a list is created for the whole corpus/collection of texts and individually for each file.

3. Key Word List: A key word list is obtained by comparing the frequency list of the corpus under analysis with the frequency list of the reference corpus. “To compute the “key-ness” of an item, the program therefore computes:

- its frequency in the small wordlist
- the number of running words in the small wordlist
- its frequency in the reference corpus
- the number of running words in the reference corpus and cross-tabulates these. Statistical tests include:
  - the classic chi-square test of significance with Yates correction for a 2 X 2 table
  - Ted Dunning’s Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus.

A word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist.”

4. Key Key words: A key key word is obtained by identifying words which are key in more than one text. The more texts they are key the more ‘keyness’ they acquire.

5. (Final Output) Associated Words List: An “associate” of key-word X is another key-word (Y) which co-occurs with X in a number of texts. It may or may not co-occur in proximity to key-word X. (A collocate would have to occur within a given distance of it, whereas an associate is “associated” by being key in the same text.)

Results using this methodology vary considerably depending on the amount of ‘noise’ in the texts used, both the actual documents and the reference corpus. For example, using 50m words of The Times (1995) as a reference corpus, and a collection of articles from the Britanica.com website which were returned when looking for the word cheese, it is possible to derive the set of associated words shown in Table 4.

However, when using as our document collection a random set of web texts which included the word cheese, we obtain the results shown in Table 5. This shows that the quality of the document collection greatly affects the results with the methodology as it stands. In essence,
Table 4. Associated Words for cheese using Encyclopaedia Britannica texts

<table>
<thead>
<tr>
<th>N</th>
<th>WORD</th>
<th>No. of Files</th>
<th>AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CHEESE</td>
<td>12</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>BRITANNICA</td>
<td>11</td>
<td>91.7</td>
</tr>
<tr>
<td>3</td>
<td>COM</td>
<td>10</td>
<td>83.33</td>
</tr>
<tr>
<td>4</td>
<td>MILK</td>
<td>5</td>
<td>41.67</td>
</tr>
<tr>
<td>5</td>
<td>WHEY</td>
<td>3</td>
<td>25.00</td>
</tr>
<tr>
<td>6</td>
<td>CHEESES</td>
<td>3</td>
<td>25.00</td>
</tr>
<tr>
<td>7</td>
<td>ACID</td>
<td>2</td>
<td>25.00</td>
</tr>
<tr>
<td>8</td>
<td>RIPENING</td>
<td>2</td>
<td>16.67</td>
</tr>
<tr>
<td>9</td>
<td>ROQUEFORT</td>
<td>2</td>
<td>16.67</td>
</tr>
<tr>
<td>10</td>
<td>CURD</td>
<td>2</td>
<td>16.67</td>
</tr>
<tr>
<td>11</td>
<td>EMMENTALER</td>
<td>2</td>
<td>16.67</td>
</tr>
</tbody>
</table>

Table 4. Associated Words for cheese using Encyclopaedia Britannica texts

This approach depends on identifying the key words in each text because their relative frequency differs from some reference corpus. Thus, if the reference corpus changes, so too will the key words. This means that a certain amount of effort must be put into choosing appropriate corpora to act as reference corpora in order for the results to be maximally useful.

Table 5. Associated Words for cheese using random web texts

<table>
<thead>
<tr>
<th>N</th>
<th>WORD</th>
<th>NO. OF FILES</th>
<th>AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CHEESE</td>
<td>69</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>MILK</td>
<td>28</td>
<td>40.58</td>
</tr>
<tr>
<td>3</td>
<td>CHEESES</td>
<td>24</td>
<td>34.78</td>
</tr>
<tr>
<td>4</td>
<td>IT'S</td>
<td>17</td>
<td>24.64</td>
</tr>
<tr>
<td>5</td>
<td>CHEDDAR</td>
<td>16</td>
<td>23.19</td>
</tr>
<tr>
<td>6</td>
<td>DON'T</td>
<td>13</td>
<td>18.84</td>
</tr>
<tr>
<td>7</td>
<td>I'M</td>
<td>10</td>
<td>14.49</td>
</tr>
<tr>
<td>8</td>
<td>FLAVOR</td>
<td>9</td>
<td>13.04</td>
</tr>
<tr>
<td>9</td>
<td>COLOR</td>
<td>8</td>
<td>11.59</td>
</tr>
<tr>
<td>10</td>
<td>CAMEMBERT</td>
<td>7</td>
<td>10.14</td>
</tr>
<tr>
<td>11</td>
<td>GOAT</td>
<td>7</td>
<td>10.14</td>
</tr>
<tr>
<td>12</td>
<td>BREAD</td>
<td>7</td>
<td>10.14</td>
</tr>
<tr>
<td>13</td>
<td>TEXTURE</td>
<td>7</td>
<td>10.14</td>
</tr>
<tr>
<td>14</td>
<td>COM</td>
<td>6</td>
<td>8.70</td>
</tr>
<tr>
<td>15</td>
<td>PASTA</td>
<td>6</td>
<td>8.70</td>
</tr>
</tbody>
</table>

Table 5. Associated Words for cheese using random web texts

In spite of these mixed results, the methodology shows promise in that we are able to take the original key word as a hypothetical node in an ontology and determine a set of potentially associated words. Similar work has also been undertaken by Eduard Hovy and his colleagues
where they built ‘topic signatures’ (i.e. lists of associated words) for concepts found in WordNet.

Collocational Similarity

Hays [1997] in an outstanding thesis on the interrelation of meanings between different forms in a lemma, proposes a method of grouping citation lines together using a measure of “collocational similarity”, which can be extended to be applied to the identification of synonyms. Collocational similarity is a measure of how similar two arbitrary citation lines, which have a common node, are. This approach has several advantages over other methods discussed in this chapter, perhaps the most important of which is the ability to handle forms with very low frequency or even hapax. This is because the method is not based on the frequency of occurrence of one form with another but on the comparison of one citation line (i.e. the immediate environment around a node word) with another such citation line. The method in its bare bones may be described as follows:

1. Citations for one type are extracted from a corpus.
2. Each citation is compared with each other citation and two measures are counted a. the number of common collocates ($C$) and b. the number of positionally fixed common collocates ($E$). These two measures are combined into one measure as explained below.
3. This combined measure of similarity is calculated for each pair of citations in a file of concordance lines from the corpus. A two dimensional matrix is constructed with the similarity values for each pair of citations.
4. Using this matrix a hierarchical clustering technique (unweighted pair-group average method [Kaufman and Rousseeuw, 1990] is applied which groups (fuses) specific citations with each other. In this manner the citations are clustered together.

E and C and collocational similarity are calculated as follows:

$$S(\alpha, \beta) = kE + l(C - E)$$
when \( k \) and \( l \) are weighting factors. These weighting factors are used to indicate the relative importance of collocates to positionally fixed co-selections in determining the meaning of a word.”


Hays presents two important measures of the success of his approach. One is called MONOSEMY which measures the extent to which the citations clustered do indeed reflect similar meaning. The figures for MONOSEMY are in excess of 95% and often over 99% for nearly all the citation files he discusses. The other measure, he calls SALIENCY, which is a measure of how many of the citations given are accurately fused i.e. the percentage of all citations which are significantly similar to another. Here Hays presents figures usually around 30-35%, which appear to be low but Hays argues that this is in fact a measure of the number of citations where “lexical content alone is a clear realisation of meaning” [Hays, 1997, :119]. He furthermore emphasises that measures calculated across the whole of a corpus are not important for his purposes, but rather the similarity between citations.

As part of the tests run on the system, Hays tested the success of the system in clustering citations from groups of synonyms (such as enormous, huge, immense, large). The assumption which Hays made, and which is central to the present work, is that “if synonyms are used to realise similar meaning units, then their usage should be similar. If the contexts of usage are sufficiently similar, then that would lead to significant collocational similarity between citations” [Hays, 1997, :137]. While the “saliency” of the results is significantly lower (22%) than for citations for one type, the “monosemy” of the clusters is extremely high. In the case of the lexical items huge/enormous/im- mense/large “monosemy” was 99%, while for the set feed/food/fodder the result was 94%. Hays concludes that his method does successfully cluster different types (i.e. lexical items) on the basis of their collocational similarity and the grouping do reflect similarity of meaning.

Hays is in fact quite strict in his criteria for similarity of meaning and certainly far stricter than most modern thesauri in a printed form. However, the method shows considerable potential for providing a method of deriving synonyms.

**SEXTANT - Grefenstette**

The most substantial effort at automating the process of creating sets of similar words which denote the same concept is that undertaken by Gregory Grefenstette in his *Explorations in Automatic Thesaurus Discovery* [1994]. This work rejects a variety of other computational methods such as those of Brown et al. [1992]. Grefenstette’s approach is based on part-of-speech parsing and a similarity metric.

Grefenstette aim is to extract semantic information by computers from unrestricted, large corpora. He stresses the importance of this
in the face of the ever growing quantity of electronic texts and the limitations of current approaches in processing these texts from a semantic perspective. He believes the following assumptions must hold in order to make a domain independent robust system:

- No hand-built domain-dependent knowledge structures
- No interest in extracting information from non-renewable sources such as human-oriented dictionaries, encyclopaedias or thesauri.
- No manually-assigned semantic tags
- No word specific information in the lexicon (such as linking constraints, valences, transitivity)
- No interest in prescriptive grammars. [Grefenstette, 1994, :34]

The method that is employed by Grefenstette in his system, which he calls SEXTANT, is based on identifying the syntactic context for each word of interest, and then using each word in the context which bears a particular syntactic relation to the node as a semantic clue. Words which share such ‘clues’ are held to be alike. Node words are compared on the basis of the number of attributes (‘clues’) the two words share and the relative importance of the attributes. The steps may be outlined as follows:

1. Raw text is divided into words and, after excluding proper names, the words are assigned a limited number of syntactic categories of the type SINGULAR-NOUN, ADJECTIVE, PLURAL-VERB-ACTIVE. This is done by simple look up in a 100,000-word dictionary.
2. Words with more than one grammatical category are disambiguated using De Marken’s [1990] stochastic grammar based on the frequency of word-category pairs and the frequency of category sequences.
3. The sentences of the text are then bracketed into NPs and VPs. This is based on an algorithm presented in Debl [1982] and Grefenstette [1983], and is based on a look up of possible category sequences. Thus a database has been constructed which details which pairs of categories are permissible and also which categories can begin and end noun phrases and verb phrases.
4. The syntactic context is then extracted for all NPs and VPs using a sequence of passes, analysing NPs left-to-right, NPs right-to-left, VPs right-to-left, VPs left-to-right, and progressive participles. In each case, relations are identified between a word and those in its immediate environment, and thus a list of relations is established for the word, for example, as for the following text:
"It was concluded that the carcinoembryonic antigens represent cellular constituents which are repressed during the course of differentiation of the normal digestive system epithelium and reappear in the corresponding malignant cells by a process of depressive dedifferentiation."

antigen carcinoembryonic epidemic digestive
antigen repress-DOBJ epitheleum system
antigen represent-SUBJ differentiation epithelium
cellular constituent cell correspond-SUBJ
cellular constituent represent-DOBJ cell correspond-SUBJ
course repress-IOBJ cell correspond-SUBJ
course repress-IOBJ cell malignant
course differentiation cell reappear-IOBJ
digestive normal cell process
epithelium normal dedifferentiation derepressive
system digestive process dedifferentiation

These relations are treated as ‘attributes’ of each word. Grefenstette is aware of the limitations of his parsing methods but argues that his approach is very fast and that errors (such as the failure to recognize constituents as the subject of reappear in the above text) only means that the system defaults to the textual windowing techniques of statistical approaches.

5. The similarity of a node with every other word is calculated using the Jaccard similarity measure (Romesburg 1990, Tanimoto 1958). This measure calculates “the number of shared attributes [as exemplified above] divided by the number of attributes in the unique union of the set of attributes for each object” (Grefenstette 1994:47):

\[
\frac{\text{Count}(\text{Attributes share by Object}_m \text{ and Object}_n)}{\text{Count}(\text{Unique Attributes possessed by Object}_m \text{ or Object}_n)}
\]

In fact Grefenstette uses a weighted version of this formula, weighting each attribute according to how many different objects it associates with, and how frequent it is in order to ‘temper the effect of frequently appearing modifiers in the Jaccard calculation’ (Grefenstette 1994:50).

6. As a result of these calculations, the similarity of each word pair in a corpus is calculated. Then, for each word, a list is produced of the most similar words as shown in Table 6. For each word the number of contexts (attributes) is indicated, and the list of similar words divided into groups whose similarity is within 0.001 of each other.

Of as much importance in these calculations as the shared attributes are the attributes which are not shared. Thus Grefenstette shows how a word may share more attributes with the node word than another but be classified as less similar because it has many more attributes not shared with the node word (1994:52).
One of the interesting applications of Grefenstette’s approach is the creation of corpus specific ‘thesauri’. Thesauri can be created by running the system over different domain specific corpora and thus different thesauri are created where for the same word different sets of similar words are identified (Table 7).

In Table 7, the M represents a medical abstract corpus, and the I represents a corpus about libraries.

With both Scott’s and Grefenstette’s methods, words may be associated which have a semantic opposition (good, bad). This occurs because the criterion of association in both cases is one of distributional similarity, at a document level in Scott, and at a sentential level in Grefenstette.

**Using Mutual Information**

A method developed by Brown et al. [1992] (not to be confused with their hierarchical clustering method presented in the same paper and described below) associates words using the Mutual Information metric. Instead of using MI to identify collocations within a window of two words around the node word [Clear, 1993], they use mutual information to analyse a window of 1001 words excluding a window of 5 words centred on the node word. They call two words semantically sticky if $Pr_{near}(w_1w_2)$ is much larger than $Pr(w_1)Pr(w_2)$. This also identifies extremely important cognitively associated words such as those shown in Table 8. Significantly from our perspective, this method

<table>
<thead>
<tr>
<th>word [Contexts]</th>
<th>Groups of most similar words</th>
</tr>
</thead>
<tbody>
<tr>
<td>tissue [350]</td>
<td>cell</td>
</tr>
<tr>
<td>treatment [341]</td>
<td>therapy</td>
</tr>
<tr>
<td>concentration [339]</td>
<td>level content</td>
</tr>
<tr>
<td>defect [338]</td>
<td>disturbance case malformation</td>
</tr>
<tr>
<td>rat [331]</td>
<td>animal mouse</td>
</tr>
<tr>
<td>method [298]</td>
<td>technique</td>
</tr>
</tbody>
</table>

Table 6. From Grefenstette 1994: 51

| M | administration [114] | injection | treatment therapy | infusion |
| I | administration [32]  | graduate office campus | education |
| M | amount [148]         | excretion | concentration level activity |
| I | amount [82]          | quantity | cost body value | set |
| M | approach [36]        | management | intervention | member error |
| I | approach [200]       | method | technique model aspect procedure |
| M | aspect [75]          | history data symptom management problem |
| I | aspect [177]         | approach | model structure | theory |

Table 7. From Grefenstette 1994: 67
we our us ourselves ours
morning noon evening night nights midnight
bed school classroom teaching grade math
sell buy selling buying sold

Table 8. “Semantically sticky” clusters (Brown et al. 1992:478)

while producing in many cases coherent and interesting sets of words, does not create a hierarchical organisation for these clusters.

All the above methods allow one to group some words together but do not provide for obtaining hierarchical information. Consequently, from our perspective their only utility is in providing candidates for semi-automatic procedures.

6.1.2 Quasi-thesauri

The computational work which bears a clear relation to the construction of a thesaurus by automatic means is the result of researchers attempting to improve statistical language models. Statistical language models are derived from the concepts developed in information theory and are concerned chiefly with developing methods to predict sequences of types at various levels of linguistic analysis. Thus researchers may wish to analyse and predict sequences of phones or phonemes (for speech recognition), sequences of letters (for spell checkers), sequences of syntactic categories (for POS tagging), or words. Such models have been proposed for speech recognition [Bahl et al., 1983], for machine translation [Brown et al., 1990], and for spelling correction [Mays et al., 1991].

One of the major problems in the efforts to create language models along these lines is what has been termed the sparse data problem. This concerns the fact that there are far fewer multi-word sequences actually encountered in a corpus than the theoretical potential number\(^1\), and this means there are insufficient exemplars to calculate appropriate probabilities. Brown et al. present the following table (9) and note that “We can be confident that any 3-gram that does not appear in our sample is, in fact, rare, but there are so many of them that their aggregate probability is substantial” [1992, 470].

This problem has been the main motivation for the development of methodologies which cluster words. By assigning words to classes it is possible to make better predictions of word sequences which have not arisen in the training corpora.

---

1 In inflected languages like Modern Greek, the same is true of the actual number of morphological forms encountered vs. the theoretical potential number of forms.

2 It is unclear how Brown et al. account for unigrams occurring zero times, as Yorick Wilks (p.c.) has pointed out. Possibly the unigrams refer to a vocabulary specified externally to the corpus.
Several researchers have investigated the problem, using a variety of techniques to cluster words into significant sets. The basic steps involved are as follows [Li and Abe, 1996]:

1. extract co-occurrence data (adjacency, syntactical, etc.)
2. assuming all words in vocabulary form one class (or each word a separate class), divide (or merge) classes based on the co-occurrence data, using some kind of criterion of similarity or distance.
3. Repeat step 2 until a stopping criterion is met.

The result of this process is a hierarchical tree of word clusters which bears a striking resemblance, in many cases, to categories and structures found in Roget’s Thesaurus. The purpose, for their research, of using such an approach, rather than Roget itself, was to construct language models specific to a corpus, and also Roget as a resource is rarely up to date.

Brown et. al.

In the same paper as that mentioned above, Brown et al. [1992] present a method for constructing thesauri. They note the importance of language models for a range of applications, and state their general assumption as follows: “... some words are similar to other words in their meaning and syntactic function. We would not be surprised to learn that the probability distribution of words in the vicinity of Thursday is very much like that for words in the vicinity of Friday” [1992, :470].

They developed methods based on using the mutual information metric [Church and Hanks, 1990, Cover and Thomas, 1991] as the similarity metric. They go through steps outlined above as follows:

1. Assigned all types in a large corpus (over 300 million tokens) to separate classes.
2. Merge the classes which have the minimum loss of mutual information.

<table>
<thead>
<tr>
<th>Count</th>
<th>1-grams</th>
<th>2-grams</th>
<th>3-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36,789</td>
<td>8,045,024</td>
<td>53,737,350</td>
</tr>
<tr>
<td>2</td>
<td>20,269</td>
<td>2,065,469</td>
<td>9,229,958</td>
</tr>
<tr>
<td>3</td>
<td>13,123</td>
<td>970,434</td>
<td>3,653,791</td>
</tr>
<tr>
<td>&gt; 3</td>
<td>135,335</td>
<td>3,413,290</td>
<td>8,728,789</td>
</tr>
<tr>
<td>&gt; 0</td>
<td>205,516</td>
<td>14,494,217</td>
<td>75,349,888</td>
</tr>
<tr>
<td>≥ 0</td>
<td>260,741</td>
<td>6.799 × 10^{10}</td>
<td>1.773 × 10^{16}</td>
</tr>
</tbody>
</table>

Table 9. Number of n-grams with various frequencies in 365,893,263 words of running text. Numbers absolute counts. (Brown et al. 1992:470)
3. Cycling through the whole vocabulary, move some words from one class to another so as to maximise average mutual information.

4. Repeat 2 and 3 for $V - 1$ times, where $V$ is the vocabulary of the corpus.

This results in a single cluster for the whole vocabulary but the order in which the clusters have been merged$^3$ corresponds to a binary tree with intermediate nodes corresponding to groupings of words. One such subtree is shown in Figure 6 and some classes are given in Table 10:

<table>
<thead>
<tr>
<th>Friday</th>
<th>Monday</th>
<th>Thursday</th>
<th>Wednesday</th>
<th>Tuesday</th>
<th>Saturday</th>
<th>Sunday</th>
<th>weekends</th>
<th>Sundays</th>
<th>Saturdays</th>
</tr>
</thead>
<tbody>
<tr>
<td>mother</td>
<td>wife</td>
<td>father</td>
<td>son</td>
<td>husband</td>
<td>brother</td>
<td>daughter</td>
<td>sister</td>
<td>boss</td>
<td>uncle</td>
</tr>
<tr>
<td>director</td>
<td>chief</td>
<td>professor</td>
<td>commissioner</td>
<td>commander</td>
<td>treasurer</td>
<td>founder</td>
<td>superintendent</td>
<td>dean</td>
<td></td>
</tr>
<tr>
<td>asking</td>
<td>telling</td>
<td>wondering</td>
<td>instructing</td>
<td>informing</td>
<td>kidding</td>
<td>reminding</td>
<td>bothering</td>
<td>thanking</td>
<td>deposing</td>
</tr>
</tbody>
</table>

Table 10. Classes from a 260,741-word vocabulary (from Brown et al. 1992:475)

While the examples cited here are striking, as are others given by Brown et al., they also given a random sample of word clusters their system obtained which are not so impressive from a linguistic point of view (cf. Table 11$^4$):

If Brown et al.’s initial assumption concerning the probability distribution of words is correct, the results shown in Table 10 are anomalous and they provide no explanation for them. The interesting question

![Figure 6. subtree from Brown et al. (1992:474)](image)

$^3$ This is calculated using the commonly accepted method i.e. the frequency of the words adjacent to each other (not using a window excluding central words of the type described above in section 6.1.1 above).

$^4$ Note that this is a different methodology from that used for Table 8.
arises as to what it is that these anomalous clusters of words share so as be generated by the system\textsuperscript{5}.

This work is based on a corpus two orders of magnitude larger than that used by other researchers such as Brown et al. [1990] and even their vocabulary is an order of magnitude larger. This is important in making contexts more balanced and more statistically significant [McMahon and Smith, 1996]. However, the computational power required to undertake these calculations is extremely large, although Brown et al. do not give any details of how long they actually took.

\textit{McMahon and Smith}

The work of McMahon and Smith takes this approach substantially further both in the quality of the results and their awareness of the limitations of \textit{n}-gram models. They note that such models consider “all previous word contexts to be identical if and only if they share the same final \(n - 1\) words” [1996, :218], and emphasise that these models are not suited to modelling long-distance phenomena\textsuperscript{6}. Their objective is to construct an algorithm which would structure the vocabulary as layers of similarity, i.e. for a given word at one level, it should be distinguished from other parts of speech (e.g. verbs), at another it should be within a layer containing all plural nouns and at deeper layers showing significant semantic differences. This is no different from any other hierarchical model apart from the desire on their part that certain levels of the hierarchy bear a certain specific significance. The approach they espouse is also based on MI, and uses a similar formula to Brown et al. where \(c_i\) and \(c_j\) are word classes and \(M_s(t)\) is the average class mutual information\textsuperscript{7}:

\[
M_s(t) = \sum_{c_i, c_j} P(c_i, c_j) \times \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)}
\]

The implementation differs in several significant details. Their procedure is as follows:

\textsuperscript{5} cf. also the more extended description of their methodology in Manning and Schütze [1999, :509-512].

\textsuperscript{6} It should be noted however, that the ‘sticky’ methodology of Brown et al. should handle long distance phenomena

\textsuperscript{7} More exactly, \(M_s(t)\) is the “average class mutual information for structural tag classification \(t\) at bit depth \(s\)” - for a full explanation please consult the paper.
1. Input is an untagged corpus (e.g. the LOB corpus Johansson et al. 1986)
2. Words are assigned to an initial hierarchical structure using the computer's random number generator.
3. The M(t) (average mutual information) value is calculated.
4. Then a single word is moved to another position in the classification space.
5. Its M(t') is calculated.
6. After iterating through the whole vocabulary, the M(t') with the highest value is selected.
7. Repeat 1-5 at the next lower level of classification.
8. Output is a clustering hierarchy.

In order to minimise the disadvantages of the sparse data problem, they start by evaluating the root binary classifications and then calculate sub-classifications on subsequent iterations. Thus they first divide the vocabulary into two classes so as to maximise the Mutual Information score, and from then on the position of the vocabulary at that level in those two classes is fixed no matter what changes may take places lower down in the hierarchy. For example if word x is in class 1 at the top binary bifurcation it will always remain so. They term this ‘top-down clustering’ and it is used in their system only to classify the most frequent words (the top 569 in a vocabulary of 33360). The classification of the next 15000 words is assumed not to affect the first set. The latter are added using an auxiliary algorithm which goes through each level deciding which binary class it should belong to (The whole system has 16 levels of classification i.e. 216 classes.). McMahon and Smith justify this in view of Zipf’s law [Zipf, 1949]. This has considerable computational advantages, especially when it must be considered that the processing of the first set (using the first algorithm) takes several weeks on a Sparc-IPC and of the second (using the auxiliary algorithm) only a few days. Their approach is very successful both in making major word class categorisations and also in identifying semantically significant groupings at finer levels of analysis.

The results of this approach are quite impressive. Although, the authors applied the algorithm to a variety of material (phonemes, letters, etc.), we will concentrate on the results for word classification. The first point to note is that their algorithm produces extremely coherent classes at the upper levels with a clear syntactic classification. Thus nouns are clearly differentiated from verbs, most prepositions occur as a separate class, as do pronouns. At lower levels of the binary hierarchy, semantic differences are made with classes such as those shown in Figure 5. Note in particular clusters such as boy, girl child, etc. and eye, face, feet, hand etc. In addition there are many anomalous groupings, which do not make semantic sense, and are consequently
quite useless for the purposes of our discussion. Examples included in Figure 5 include reason, thing, place, way etc. and cases, days etc. However, it is important to consider why these arise. McMahon and Smith are aware of the problem and suggest that this is due to the fact that words are polysemic. In this the authors do not depart from other researchers in the field by allowing only one position for each word in the classification structure. This induces considerable ‘noise’.

The clustering criterion only exploits bi-gram information, thus it cannot take into account semantically significant relations two or more words distant from a particular word. For computational reasons, their algorithm like others, searches for locally optimal classifications while at a global level the particular classification may not be entirely correct.

As part of an attempt to provide an objective evaluation of their particular approach, the authors undertake an interesting comparison between the their ‘top-down’ clustering approach and the ‘bottom-up’ merging approach taken by Brown et al. [1990] and other authors.
(e.g. Brill and Marcus [1992]). Using the VODIS corpus [Cookson, 1988], they use it as input for the first Brown et al. algorithm and their own system. Only words with a frequency of 30 or more were used, thus limiting the vocabulary to the 256 most frequent words. The results are very interesting, while difficult to interpret. For example, the merge algorithm results in a primary division of number words vs. all the rest, while in the top-down algorithm, they are grouped together at the fourth hierarchical level. The merge approach mixes times and days with words relating to trains and tickets, and separates the days of the week, while the other approach has more homogenous groupings. They note that their method performs better “with respect to overall classification topology” which we take to refer to the top level class divisions, but it “loses its quality at lower levels . . . trees (as opposed to directed graphs) are inherently vulnerable to unnecessary data fragmentation”, and go on to suggest that “The inaccuracies introduced by the first of these characteristics may be controlled, to a limited extent only, by using a hybrid top-down and bottom-up approach.” [McMahon and Smith, 1996, :236]. In spite of these comments, in reality it is extremely difficult to assess the comparative merits of the systems without having detailed access to the data and without tracing the clustering process so as analyse the influences which determine specific clusterings.

It should be stressed that the approaches which Brown et al. and McMahon and Smith represent result in major divisions of the vocabulary according to syntactic categories (or quasi-syntactic categories) and at lower levels provide semantically significant groupings. In this, the hierarchies developed bear no relation to the taxonomy developed by Wilkins, Roget or even Glazier. Furthermore, these approaches have been dominated by Mutual Information as the clustering criterion. There appears to have been no attempt on these authors parts to justify this particular criterion, or to consider its merits in comparison with other potential clustering criteria.

Other Research

Another group of researchers, while following the same essential methodology, have had the construction of a thesaurus (in the sense of hierarchically clustering words) as their prime objective. For example, Li and Abe [1996] approach the problem with the stated intent to construct thesauri rather than improve statistical language models. The algorithm they employ uses simulated annealing with an energy function based on the Minimum Description Length Principle. They use the automatically constructed thesauri to solve syntactic disambiguation. While their approach is interesting they do not give sufficient experimental results for one to assess its potential, without replicating their algorithm.
6.2 Building a Hierarchy

Even though it is a fundamental aspect of ontology learning, the automatic construction of hierarchical organisation of terms has not been addressed that often or that clearly. In the following we present some attempts which we believe could usefully contribute to ontology learning, even if their original intent may have been quite other.

6.2.1 Pathfinder Networks

Pathfinder networks are associated with Don Dearholt and Roger Schvaneveldt. They are a way of representing proximity data derived from multiple sources. Classically students take psychometric tests asking them to identify the similarity between a series of terms and the average similarity is used as the proximity data to create a graph representation. Proximity data can be obtained from many different sources including similarities, correlations, distances and conditional probabilities, all derived from many different types of raw data. The entities (words, concepts, etc.) form the nodes and the links are created by the pattern of proximities. This approach has been applied to texts such as the Longman Dictionary of Contemporary English (LDOCE)\(^8\) [McDonald et al., 1990] and as an interface for an IR system [Fowler and Dearholt, 1990].

In the former paper, the authors use the definitions (glosses) from LDOCE as raw data and assume words which co-occur in the same sentence are semantically related\(^9\). They consider a number of different relatedness functions i.e. different ways to measure relatedness between terms, about which they make interesting observations (Table 12).

The functions which provide the best results are \(d_{cp, min}\) and \(iou\) ‘on intuitive grounds’ (these are shown in Table 13) and the authors went on to test these relatedness values by comparing them with results from human subjects. The correlation with human subjects ranged from 0.60 - 0.69. However, the functions which correlated best were not the ones which provided the best intuitive results. The authors dealt with the obvious problem of ambiguity present in the list of related words shown in the table by sense tagging each occurrence of an ambiguous word. This was done by identifying the maximum overlap between the context of a given word and the different definitions of that word in the LDOCE.

This processing of the raw LDOCE data allowed the generation of Pathfinder Networks (PFNETs) such as the one shown in Figure 8. This is clearly a long way still from being an ontology, although Schvaneveldt saw Pathfinder Networks as a form of knowledge representation. However, he was more interested in such structures in terms of what they revealed about psychology and the cognitive organisation.

---

8 The LDOCE was extensively used in NLP research in the 1980s and early 1990s as it was one of the first dictionaries available for researchers [Wilks et al., 1996]

9 This is the distributional hypothesis cf. Section 6.5
Table 12. Relatedness functions from McDonald et al. [1990]

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( cp(x, y) )</td>
<td>[ \frac{f_{xy}}{f_y} = \Pr(x</td>
<td>y) ]</td>
</tr>
<tr>
<td>( dcp(x, y) )</td>
<td>( \Pr(x</td>
<td>y) - \Pr(x) )</td>
</tr>
<tr>
<td>( dcp_{\text{min}}(x, y) )</td>
<td>( \min(\text{dcp}(x, y), \text{dcp}(y, x)) )</td>
<td>Minimum of ( dcp ) in both directions. Symmetric. Sensitive if ( f_x ) and ( f_y ) are similar, but results in zero if they are considered different.</td>
</tr>
<tr>
<td>( iou(x, y) )</td>
<td>( \Pr(x \text{ and } y</td>
<td>x \text{ or } y) )</td>
</tr>
<tr>
<td>( dex(x, y) )</td>
<td>[ \frac{f_{xy} - f_x f_y}{\min(f_x, f_y) - f_x - f_y} ]</td>
<td>(dependency extraction) Normalises ( f_{xy} ) by mapping it to ([0,1]) according to its scaled position between its minimum and maximum possible values. Symmetric. Fully sensitive for all ( f_x ) and ( f_y ).</td>
</tr>
</tbody>
</table>

Table 13. Words most strongly related to \textit{bank} from McDonald et al. [1990]

<table>
<thead>
<tr>
<th>Function</th>
<th>Related Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>( dcp_{\text{min}} )</td>
<td>account cheque criminal earn flood flow lake lend money pay prevent promise rate river . . .</td>
</tr>
<tr>
<td>( iou )</td>
<td>account busy cheque criminal earn flood flow interest lake lend money overflow pay river rob . . .</td>
</tr>
</tbody>
</table>

of knowledge than their application to AI problems.

PFNETs have been used extensively to model a great variety of topics using very different data. For example, recently White et al. [2004] used PFNETs to analyse the proximity of terms in the Proceedings of the National Academy of Science (PNAS) database using the Medical Subject Headings (MeSH) standardised terms. Given an input term, the system would represent visually the 25 most highly co-occurring terms in the bibliographic database as shown in Figure 9.

From our perspective, the structures could clearly be used as starting points for ontology construction. Many ontology learning systems (e.g. KAON) provide user interfaces which appear very similar to a PFNET. However, the main problem is that PFNETs do not bring out taxonomic relations. On the contrary, what they appear to be good at is modelling non-taxonomic relations of the ‘tennis problem’ type (cf. Section 5.4). In this sense they could be useful in complement to other techniques.
Figure 8. From McDonald et al. [1990]
Figure 9. PFNET derived from term co-occurrence in PNAS, from White et al. [2004]

such as the one described in the next section.
6.2.2 Document Subsumption

Sanderson and Croft [1999] have presented an approach from the perspective of wanting to create a concept hierarchy in order to organise documents retrieved in a query. They adopted the following design principles:

- Terms for the hierarchy were to be extracted from the documents;
- parent terms would refer to a more general concept than its child, in other words, the parent’s concept subsumes the child’s;
- the child would cover a related subtopic to the parent;
- a strict hierarchy, where every child had only one parent, was not considered important;
- ambiguous terms would be expected to have separate entries in the hierarchy

Their approach is based on the document frequency of a term. They construct hierarchy using the notion of subsumption which is defined as follows:

“x subsumes y if the documents which y occurs in are a subset of the documents which x occurs in.”

More formally this was defined as follows, where given that x and y are two terms, x is said to subsume y if the following two conditions hold:

\[
P(x|y) = 1, P(y|x) < 1
\]

In practice, they found they had to set the value of the first condition to 0.8 because just a few y terms were not co-occurring with x. They avoided the issue of ambiguity by a set of documents “where ambiguous terms were only used in one sense”. This, they claimed, was possible by using the top ranking documents retrieved from a query.

1. Given a query a set of documents was retrieved.
2. Local Contextual Analysis [Xu and Croft, 1996] was applied to expand the query terms.
3. The set of terms was further expanded by comparing the frequency of occurrence of the terms in the retrieved documents (\(x_r\)) with the frequency in the whole collection (\(x_c\)). Those above this threshold were included: \(x_r/x_c = 0.1\)
4. 500 top ranked documents were selected yielding 2, 420 terms.
5. Every term was compared to every other term for the subsumption relationship resulting in about 200 subsumption pairs.
6. These were organised into a hierarchy after removing infrequent pairs.
For the TREC topic “Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?” the hierarchy shown in Figure 10 was produced.

![Figure 10. From Sanderson and Croft (1999:4)](image)

The authors also performed a user evaluation in which they concluded that their method was appropriate for the construction of concept hierarchies in accordance with their design principles. However, it is important to observe that the links between terms in the above structure are extremely heterogeneous. No two links have the same semantic interpretation (Wilks p.c.). This is significant in terms of evaluating the applicability of this method to ontology construction in general.

6.3 LABELLING RELATIONS

It is characteristic of the approaches discussed so far that they are unable to establish what kind of relationship exists between any two words or concepts. While hand-crafted concept hierarchies/ontologies like WordNet or Mikrokosmos are very precise in the nature of the semantic relations between any two synsets or concepts, a hierarchy such as that presented in Section 6.2.2 does not and cannot. In part, this is inevitable, since the determination of a semantic relation is an act of interpretation, of categorisation which the human intellect must consciously impose on linguistic phenomena. In spite of this, some attempts have been made to characterise certain kinds of semantic relations in order to be able automatically to identify the lexical items involved.

6.3.1 Synonymy or Substitutability

Church et al. [1994] integrates the use of statistical tests and the exploitation of syntactic parsing to concentrate exclusively on creating
sets of semantically related words. Their work may be considered a unification of Grefenstette’s syntactic focus with the exploitation of Mutual Information discussed above (Section 6.1.1). They develop a ‘sub-test’ i.e. a test of how substitutable one word is for another in order to make the notion of synonymy more precise. Rather than working with MI figures for a window of words around the node, in their work they first parse their corpus so as to be able to identify Verb-Object sequences (for example). This permits a comparison to be made of the objects which follow two verbs. Thus, they analyse the verbs request and ask for and note that in their corpus of 44 million words the former has 59 significant objects and the latter 85, while there is an overlap of 28 significant objects:

They construct their ‘test’ so as to answer the question “if we substitute the word x for the word y in some context, what is the chance that we would end up with a coherent sentence of English”? They use the t-score to analyse the overlap (shown above) and determine whether this is chance and if not how interesting it is. Thus they are able to construct a table of possible substitutes for a given verb as shown in Table 15.

<table>
<thead>
<tr>
<th>xy</th>
<th>yz</th>
<th>xy &amp; yz</th>
<th>expected</th>
<th>t</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>85</td>
<td>28</td>
<td>2.30</td>
<td>6.25</td>
<td>request</td>
<td>ask_for</td>
</tr>
<tr>
<td>208</td>
<td>85</td>
<td>46</td>
<td>8.12</td>
<td>5.74</td>
<td>seek</td>
<td>ask_for</td>
</tr>
<tr>
<td>88</td>
<td>85</td>
<td>46</td>
<td>3.43</td>
<td>4.17</td>
<td>grant</td>
<td>ask_for</td>
</tr>
<tr>
<td>92</td>
<td>85</td>
<td>18</td>
<td>3.59</td>
<td>3.40</td>
<td>obtain</td>
<td>ask_for</td>
</tr>
<tr>
<td>102</td>
<td>85</td>
<td>17</td>
<td>3.98</td>
<td>3.07</td>
<td>demand</td>
<td>ask_for</td>
</tr>
</tbody>
</table>

Table 15. Extract from a table of possible substitutes [Church et al., 1994]

The authors note that it is often unclear what the semantic relationship is between ‘substitutable’ words and that it is “not always one that fits nicely into a familiar category such as synonymy, antonymy, and hyponymy” [Church et al., 1994, :13]. Also there is a difference between grammatical classes. Thus while nominate, impeach and assassinate all are highly substitutable for elect, the same is not true of the corresponding nouns: nomination, impeachment and assassination are not highly substitutable for election.
Their approach shows great promise especially since there are many imperfections due to the nature of the data input (tagging errors, parsing errors etc.) which could be reduced to improve performance. However, it does not make explicit how to evaluate whether a sentence resulting from a substitution is coherent or not. This remains a human judgement. Also, it does not concern itself with the respective relation between groups of synonyms.

6.3.2 Antonymy

Justeson and Katz [1992] have analysed the textual patterns of distribution of adjectival antonyms. After considering particularly the Deese [1964] antonyms in a tagged version of the Brown corpus, they concluded that “antonyms are those semantically opposed words that are conjoined and often opposed to one another at relatively high rates in sentences by substitution for one another in otherwise essentially identical (or parallel) phrases” [Justeson and Katz, 1992, 181]. They provide several examples of these ‘essentially identical phrases’:

- They indicated that no new errors were being made and that all old errors would be corrected ...
- ...one is hard to understand and the other is easy to misunderstand.
- ...things would get worse before they got better

The pattern, according to them, is so pervasive that there is “no chance for a genuine antonym pair to fail to show up” (ibid.). However, they note that co-occurrence by phrasal substitution is not restricted only to adjectives which are traditionally considered antonyms but also characterises such pairs as social-political or economic-political.

This work is entirely descriptive i.e. no attempt appears to have been made to use this method to identify candidate antonym pairs in a fresh corpus. However, as it is very similar in spirit to the Hearst/Morin work on hyponyms, it would appear to have potential as a discovery procedure. The question remains how much human intervention is necessary in such an approach.

6.3.3 Hyponymy

Another classic paper concerning word associations is that of Marti Hearst (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. In this paper, she proposes to “identify a set of lexico-syntactic patterns that are easily recognisable, that occur frequently and across text genre boundaries, and that indisputably indicate the lexical relation of interest” (ibid.). Hearst argues that pattern matching is more successful than parsing in identifying patterns in text which reveal hyponymy relations between words. She presents five such possible patterns:
The method by which such patterns are found involves the following steps:

1. Identifying a potential lexical relation of interest e.g. group/member
2. Collect a list of exemplars of this relation e.g. England/country using MRDs, KBs, etc.
3. Collect a list of citations where these lexico-syntactic expressions occur.
4. Identify the commonalities in the syntactic/lexical environment in order to construct a pattern.
5. Use the pattern to collect instances of the target relation.

Because step 4 is undetermined, Hearst did not implement this procedure. For one relation (such as), out of 8.6M words of encyclopaedia text, she found 7067 sentences which contained the phrase, and of these 152 fitted the pattern. With a slight easing of restrictions, 330 exemplars were found. When comparing with WordNet, she found that 61 out of 106 possible relations already existed in WordNet. A number of difficulties are noted such that the exemplars are not properly in a hyponym relation (e.g. king & institution, device & plot) or reflected a very restricted point of view (e.g. Washington & nationalist, aircraft & target). Furthermore, she does not present enough figures to properly quantify the success of her method which she describes as ‘encouraging’.

The main problem with her approach is that it is not at all clear that one can find a sufficient number of patterns in order to properly construct a hyponym tree in all the complexity we observe in Wordnet. Emmanuel Morin in a series of papers has built on Hearst’s work [Morin, 1999a,b, Finkelstein-Landau and Morin, 1999, Morin and Jacquemin, 1999]. Morin has developed a system which he calls
PROMÉTHÉE and which in effect is an implementation of the procedure outlined by Hearst in her original paper. In Morin’s system, there are seven steps [Finkelstein-Landau and Morin, 1999, 2-3]:

1. Select manually a representative conceptual relation, for instance the hyponym relation.
2. Collect a list of pairs of terms linked by this relation. The list of pairs can be extracted from a thesaurus, a knowledge base or can be specified manually.
3. Find sentences in which conceptually related terms occur. These sentences are lemmatised, and noun phrases are identified. Therefore, sentences are represented as lexico-syntactic expressions.
4. Find a common environment that generalises the lexico-syntactic expression extracted at the third step. This environment is calculated with the help of a measure of similarity and a procedure of generalisation that produce candidate lexico-syntactic patterns.
5. Validate candidate lexico-syntactic patterns by an expert.
6. Use new patterns to extract more pairs of candidate terms.
7. Validate candidate terms by an expert, and go to step 3.

Crucially, Morin has not been able to avoid the intervention of an expert in at least two steps of his process. He states that it “can find only a small portion of related terms due to the variety of sentence styles and the inability to find a common environment to all those sentences” [Finkelstein-Landau and Morin, 1999, 6]. In the later paper, he and his co-author argue in fact for an integration between the supervised method of PROMÉTHÉE and unsupervised methods based on mutual information and log-likelihood.

The approach proposed by Hearst has come to dominate ontology learning. More than any other idea, the idea that, for a given ontological relationship, there exists a set of lexico-syntactic patterns, has been exploited by a large number of researchers attempting to build ontologies. We will return to Hearst’s lexico-syntactic patterns repeatedly below.

6.4 OTHER SYSTEMS AND APPROACHES

6.4.1 CAIULA

The approach that Grefenstette espouses was used chiefly to group nouns and adjectives, and we now turn to another effort in this field whose main objective is to classify verbs. Basili et al. [1996b] propose a system, also consisting of a series of various NLP components, whose objective is to construct a hierarchy of verbs where the clusters have semantic descriptions. Thus for example, their system clusters the following (Italian) verbs: *applicare*, *corrispondere*, *distribuire*, *effettuare*, *derivare*, *dichiarare*, *eseguire*. Their system may be described as follows:
Class 2187: archive, publish, develop, extract, contaminate, derive, base, contain, use, do, provide, apply, operate
Class 2124: select, filter, obtain, project, use, simulate, describe

Table 16. Example classes from the Basili et al. approach

1. A corpus is tagged and parsed using a shallow parser [Basili et al., 1993a].

2. Syntactic pairs and triples are extracted together with high level semantic tags (derived from WordNet for the English texts, hand tagged for the Italian).

3. Clustered association data of the form V_prep N(ACT, to, HUMAN/ENTITY) is derived from merging collocational data of the form V_prep N(sell, to, shareholder) and V_prep N(assign, to, taxpayer). Note: the system concerns itself exclusively with verbs and uses only data concerning the complements and adjuncts.

4. This data is presented to a linguist to be replaced by a conceptual relation, e.g. [ACT]1(BENEFICIARY)1[HUMAN ENTITY].

5. For each content word, all collocations in which it participates are collected. From the data concerning the word measurement i.e. N_prep N(measurement, from, satellite), N N(radar, measurement), N_prep N(measurement, from, antenna) the system then derives a selectional restriction [INSTRUMENTALITY] <= (INSTRUMENT) <= [MEASUREMENT]. Note: this step in effect generalises over verb complements in order to determine the correct selectional restriction.

6. From the resulting database, it is possible to extract, for each verb, data of the type: produce / (AGENT: HUMAN ENTITY, OBJECT: GOODS, INSTRUMENT: INSTRUMENTALITY).

7. On the basis that semantic similarity exists where words of the same conceptual type play the same role, the verbs are clustered using a system of incremental clustering called COBWEB [Fisher, 1987]. Basili et al. introduce the notion of category utility which is given a formal definition, and the COBWEB system then implements a hill-climbing algorithm which decides on cluster structure which maximises infra-class similarity and intra-class dissimilarity.

The authors provide no formal evaluation of their methodology but rather state that “the resulting classifications were judged in general quite expressive and semantically biased” (ibid.: 135). They note that, by changing various parameters in their algorithm, they can obtain either very small clusters (2-3 verbs) which have high similarity or larger clusters of more heterogeneous verbs. This is because the “thematic structure of verbs is variegated and poorly overlapping” (ibid.: 138). Example output is shown in Table 16.

From a linguistic perspective, these do not seem particularly ‘se-
mantly biased’ but their approach does have the advantage that it specifies a ‘degree of membership’ scale for each item.

6.4.2 **ASIUM**

Faure [Faure and Nédellec, 1998, Faure and Nédellec, 1999] has developed a system which he calls ASIUM and which is similar in certain ways to the work of Basili et al. although not so algorithmically complex and not focused on verb hierarchies. There is again the intervention of an expert i.e. it is not an entirely automatic system. The method can be described as follows:

1. Input to the system is tagged and parsed text (using the SYLEX system of Constant 1995). Stop words and adjectives are removed - only nouns, verbs and prepositions are used.
2. Thematic information (subject, object) is obtained for each verb and an ‘instantiated’ subcategorisation frame is extracted of the form <to travel> <subject: father > <by: car>. The latter would be derived from a sentence such as “My father travels by car.”
3. Factorisation: All headwords (i.e. nouns) which occur in the same context - with the same verb and preposition are gathered together into ‘basic classes’.
4. Clustering: The basic classes are aggregated into clusters. This clustering process depends on a distance metric between clusters which is calculated as follow:
\[
\text{dist}(C_1, C_2) = 1 - \frac{\sum FC_1 \times N\text{comm}\{C_1\} + \sum FC_2 \times N\text{comm}\{C_2\}}{\sum \text{card}(C_1) f(\text{word}_i\{C_1\}) + \sum \text{card}(C_2) f(\text{word}_i\{C_2\})}
\] (6.1)

where \( C_1 \) and \( C_2 \) are the two clusters, \( \text{card}(C) \) represents the number of different headwords there are in a given cluster, \( N\text{comm} \) the number of common headwords, \( \sum FC_i \) the sum of frequency of the head words in a cluster, and \( f(\text{word}_i\{C\}) \) the frequency of the \( i \) th word in cluster.
5. Expert validation: each candidate cluster is presented to an expert to a) approve or reject, and b) to name the new cluster.
6. As a result of the aggregation of classes into clusters, there is a process of subcategorisation frame learning whereby if two classes are combined then the subcategorisation frames are generalised.

Faure et al. argue that by incorporating user validation into the system there are several advantages. Firstly, by labelling the clusters themselves, the user is able to make the ontology ‘more comprehensible’ i.e. express it in terms relevant to their needs. Faure rejects an approach which labels the clusters with the most frequent word. Secondly, as the ontology is validated at each level, and does not proceed before it is fully validated, each stage of construction has a ‘sound basis’. Furthermore, during the validation process, the user can detect noise and reject certain clusters, or words in a cluster. Although the system was developed on behalf of the company Dassault and was
intended for use with aeroplane manuals, the authors could not test it on this material and used a corpus of cooking recipes instead. They do not present a convincing evaluation, however, of their system as all they could say was that most of the clusters were relevant and that 5% of the corpus was sufficient to ‘cover’ 48% of the remaining text in the corpus.

This does not invalidate their approach; it merely indicates the need for comprehensive evaluation in terms of effort required and resulting utility. It would appear that the ASIUM system might be able to reduce human intervention further by integrating a stage of statistical tests along the lines of Brent/Manning.

6.4.3 Ontolearn

The Ontolearn system has been presented by Navigli and Velardi in a number of papers [Navigli et al., 2003, Navigli and Velardi, 2004, Velardi et al., 2005]. The system is not intended to construct ontologies from scratch but rather to prune an existing ontology such as WordNet. The approach has been innovative in number of respects: a) its emphasis on handling the technical terminology of a domain, b) its attempt to provide ‘semantic interpretation’ of complex technical terms, c) innovative methods for evaluating the overall system (cf. Section 11.6). Their method can be summarised as follows (although it has evolved over time):

1. The extraction of terms from the texts.
   Using the Ariosto system [Basili et al., 1993b,a, 1996a], the corpus is parsed in order to extract candidate terms including noun compounds, adjective noun pairs and ‘base noun phrases’. They then use a ‘domain relevance’ metric which compares the specificity of the terms to that domain as opposed to a wider collection of corpora. A second filter identifies those terms which are sufficiently frequent across the corpus in order to merit inclusion [Velardi et al., 2001]. These procedures are relatively uncontroversial (cf. keyword approach of Scott 1998, ‘weirdness’ of Ahmad 2000).

2. Semantic Interpretation
   First Ontolearn organises concepts into a hierarchy on the basis of simple string inclusion. The next step is to semantically disambiguate each multiword term. If each word in each term is assigned a set of possible WordNet synsets, the task is then to identify the right combination. The result should be as follows:
   \[ S(\text{transport company}) = \{(\text{transportation#4, shipping#1, transport#3)}, \text{company#1}\} \]
   The process for achieving this is quite complex:
   a) A semantic net is constructed for each sense of each word in the term using data from WordNet up to a distance of three steps in WordNet
   b) Each possible net for each term is intersected with each possible net for the other term and a score is computed which
is dependent on the number of semantic patterns which are found. The semantic patterns are 13 predefined patterns including such things as co-occurrence in the SemCor corpus, or if there is a hypernymic path from a word in the WordNet gloss to the other word in the term. This allows the highest scoring intersection to be chosen so that specific pair of WordNet synsets is chosen.

c) The next step reorganises the set of string-included words into a hierarchy using the information derived from WordNet.

d) Following this the semantic relation between the words in the multiword term is identified. They create a set of semantic relation types from a number of sources. Then they use inductive machine learning (specifically C4.5) to learn a set of rules for tagging the relations between words in the terms.

3. Creating a specialised view of WordNet.

The newly derived domain concept forest is attached at the appropriate point to WordNet pruning concepts extraneous to the domain and pruning intermediate nodes.

The Ontolearn system is significant in both what it attempts to achieve and what it avoids. It deals directly with the issue of multiword items and this is a major contribution. However, it does not deal with the acquisition ab initio of ontological knowledge and takes for granted the use of WordNet as the skeleton upon which to build. While this may be an effective methodology in certain cases, there is no evaluation of the suitability of WordNet for any of the domains covered. Furthermore there is no consideration of terms which do not already exist in WordNet (which is necessary for the provision of the ‘Semantic Interpretation’) and consequently there is no method for handling individual terms (i.e. single word terms). Above all their method is not clearly a method for knowledge acquisition. In sum, their approach could be useful as an auxiliary approach when handling multi-word terms and given an existing skeletal ontology.

6.4.4 KnowItAll

The KnowItAll system is one whose prime objective is the collection of ‘facts’ from the World Wide Web. The facts it collects are of the type “Paris is a city” or “Hawaii is a US State” i.e. instances of a generic class. It was developed by a team led by Oren Etzioni and uses a number of techniques in an interesting and innovative manner. The system does not attempt to build an ontology and so is not directly comparable but in many ways the authors are dealing with closely related challenges.

The basic version of the system as described in [Etzioni et al., 2004] has four basic components:

- Extractor This module extracts from instantiations of Hearst type patterns (e.g. ‘NP1 such as NPList2’) the nouns in the list. So from ‘We provide tours to cities such as Paris, Nice, and Monte
Carlo’), this module extracts the three instances of cities. The patterns KnowItAll uses are either derived from Hearst or developed by themselves. The extractor uses the Brill tagger to tag the texts with POS and then uses regular expressions to extract the noun phrases.

- **Search Engine Interface** Given template rules, this module constructs the query (e.g. ‘cities such as’), issues this to multiple search engines and returns the results to the extractor module. The authors use multiple search engines and cache the results. They believe that this is most efficient as it leverages the efficiency of commercial search engines. However in more recent work they have developed their own NLP focussed search engine [Cafarella and Etzioni, 2005].

- **Assessor** This module uses pointwise mutual information (PMI) based on Turney’s approach (2001) to ‘assess the likelihood that the Extractor’s conjectures are correct’. The PMI is calculated between each instance of a city (e.g. Paris) and a number of phrases associated with cities such as ‘the city of Paris,’ which the authors term ‘discriminator phrases.’ The results are combined using a Naive Bayes Classifier.

- **Database** There is persistent storage of the results of the queries issued in a commercial database.

They present results for five classes, city, country, US State, actor and film which they compared to the Tipster Gazetteer and the IMDB database. The results were very impressive with, for example, precision 0.98 up to recall 0.76, and this drops to 0.71 precision for 1.0 recall (Figure 11).

Subsequent versions of KnowItAll extended its functionality by adding pattern learning [Etzioni et al., 2005] i.e. the learning of domain specific extraction rules. For example, “the film <film> starring” is a high precision, high coverage pattern for that specific domain. These patterns increase coverage by learning extractors and accuracy because they are also used as discriminators (cf. Section 7.4.1 for more details). Another later version also adds the capability to learn subclasses. So for example, for the class scientist, they system will a collection of types of scientists biologist, zoologist, astronomer, meteorologist, geologist etc. The basic method is to use the generic rules and apply a common noun test i.e. test that the extracted noun is not a proper noun by testing for the absence of capitalisation. The assessor module in this version tests for morphology and hyponymy in WordNet, in addition to the application of PMI as before [Popescu et al., 2004].

The major contributions of KnowItAll are to emphasise that a great deal of common sense knowledge is available on the Web and to provide a number of techniques to extract that knowledge with very good precision and recall. One of the most important aspects is that they put a set of technologies together to make a very impressive system. Of particular interest from our perspective are their techniques
for learning further extraction patterns and for subclass learning. Also the design of the assessor which determines the likelihood that the fact extracted is correct is a major contribution.

They do not attempt to construct an ontology and thus avoid many of the fundamental issues in such an attempt. They use multiple terms as head words in the instantiated discriminator phrases to avoid ambiguity but other than that the ‘facts’ may come from anywhere or any domain.

### 6.4.5 Heterogeneous Evidence Approach

Philip Cimiano and colleagues have presented an approach based on learning ontologies from a multiplicity of sources [Cimiano et al., 2005]. In this approach, the authors integrate a number of different approaches for the learning of taxonomies proposed in the literature including Hearst patterns, the head matching/NP structure technique (originally proposed by Navigli et al. [2003]), hyponymy information from WordNet, subsumption based on corpus based syntactic features (an extension of Grefenstette [1994]) and document based subsumption [Sanderson and Croft, 1999]. The corpus they used was a tourism-focused corpus and as a gold standard a hand-crafted tourism ontology.

A number of standard classifiers (implemented in WEKA) were used to identify the optimal combination of these different methods.
The best results came from using a SVM classifier which resulted in an F measure of nearly 33%. The relatively low F measure (in view of NLP results in general) is indicative of how great the challenge of ontology learning still remains. The authors set themselves a difficult task in that what they wanted to do was ‘reproduce’ a specific ontology (the Karlsruhe Tourism Ontology) given a specific corpus. The striking difference between this result and the results of KnowItAll reflect the difference between constructing an ontology and learning facts. The importance of this paper lies both in the concept of integrating multiple sources and the manner of implementation, especially the use of the of classifiers to combine the evidence.

### 6.5 From Syntax to Semantics

A significant theme in Basili et al.‘s approach (cf. Subsection 6.4.1) is that “syntactic similarity cannot be used as a clue for semantic similarity” (ibid:130). They provide the examples of where the same syntactic components (with + noun) have different semantic roles. Instead they argue that “semantic similarity is strongly suggested by the observation of verb configurations in which words of the same conceptual type play the same roles” (ibid.). In a certain sense they have avoided facing the issue of how much can be derived directly from plain text by integrating Wordnet categories or manually annotating the words with semantic tags. Faure et al. also appear to have the same perspective, stating that “words occurring together after the same prepositions (or the same function), and with the same verbs represent a same concept [sic]” [Faure and Nédellec, 1999, :4]. The latter authors also sidestep the crucial issue by integrating user validation into their system.

This is an important issue with regard to all efforts to cluster or otherwise predict semantic structure on the basis of syntactic or distributional textual evidence. Considerable work has been undertaken by linguists who argue for an intimate relation between syntax and semantics. Thus, for example, Sinclair [1991], in a detailed analysis of corpus evidence for the verb yield argues that the alignment of sense with structure accounts for over 70% of textual examples (ibid.:57). The so called “distributional hypothesis” can be traced back at least to Zellig Harris (cf. Harris 1968, 1969) and through him to the American Structuralists of the earlier 20th century.

From a very different linguistic background, Levin [1993] provides extensive evidence and arguments that “the syntactic behaviour of a verb [is] to be predicted from its meaning”. This has subsequently been taken up by Dorr and Jones [1996] who conducted experiments trying to show that verbal semantics could be predicted on the basis of syntax. But the efforts of Dorr and Jones show that if the semantic class is known then the syntactic behaviour can be predicted, not that the syntactic behaviour predicts the semantic class.
6.6 Conclusion

In this Chapter we have surveyed a large collection of individual NLP techniques which could contribute to the process of constructing an ontology automatically. We have also looked at some complete systems. While it may be obvious to experienced researchers in the field, the main point is that none of these approaches is infallible, all produce noisy output to varying degrees. The crucial point as we argued in the last section is that the ‘distributional hypothesis’ cannot stand on its own so it is very hard to go in a simple manner from the behaviour of a word in text corpora to its meaning. This is a major reason to take an approach to ontology building which recognises the necessity of manual input, even if it is clearly our objective to minimise it as much as possible.
EXPLICIT KNOWLEDGE AND LEXICO-SYNTACTIC PATTERNS

My family and other animals.

Gerald Durrell

7.1 EXPLICIT KNOWLEDGE IN TEXT

We have considered the distinction between explicit and implicit knowledge in Chapter 2. While we noted a number of philosophical distinctions by philosophers between explicit, implicit and tacit knowledge, these are not easily translated into useful terms for the application of Natural Language Processing techniques. In other words, while these distinctions may have philosophical and even cognitive justification, they cannot be easily made to correspond with the phenomena that NLP techniques recognise and manipulate (such as those reviewed in Chapter 6). From the perspective of texts and natural language processing, there is a great deal of implicit knowledge in the sense of knowledge which can be garnered by using statistical and other methods which associate one term with another. We have surveyed some of these techniques in Section 6.1 above. This knowledge is implicit only in the sense that there is something there but it is not easy to translate this automatically into something useful for the ontology engineer. In terms of building an ontology, knowing that one term is associated or connected with another is a significant step but only part of the journey. It does not allow for propositional knowledge, and it does not enable inferences to be drawn in the manner expected of a formal ontology or knowledge representation. Above all it does not inform us about the nature of the ontological relationship between any given two terms.

Pearson [1998], in an extensive study of defining environments for technical terminology, defines a formal defining expositive as the “rephrasing of an existing definition for the purpose of explanation or clarification” which has a corresponding pattern: \( X = Y + \text{distinguishing characteristics, whereby } X \text{ is subordinate to } Y \) (ibid. 136). She presents a number of such lexical environments, some of which are included in our catalogue below. One important aspect which she stresses is the need to recognise the full (often multiword) term (X and Y) in the process of identifying such contexts. A major issue raised by her research is the complexity and extent of the “distinguishing characteristics” in such defining contexts:

Example 7.1 from Pearson [1998, :145]

A misdelivered frame is a frame transferred from a source user to a destination.
user other than the intended destination user.

Example 7.2  *ibid.*
*An abscess is a cavity filled with pus.*

While ideally automatic systems should be able to extract the ‘differentia’ this is beyond our current capacity. Thus in the current context, our objective for the above examples would be to identify and extract *(misdelivered frame is a frame)* and *(abscess is a cavity).* Pearson makes a distinction between simple and complex formal defining expositives, where the complex ones span more than one sentence. Her examples are all ones where an anaphor needs to be resolved as in the following examples:

Example 7.3  *from Pearson [1998, :152]*
*During sexual reproduction, a male sperm joins with a female egg. This is called fertilisation.*

In this particular case, the antecedent of *this* cannot be resolved because it is not clear that there is one. More generally anaphora resolution is dependent on a great deal of world knowledge which cannot as yet be brought to bear in the process of ontology learning.

In contrast to Pearson, we define explicit knowledge for the purposes of this work to be ontological knowledge which is expressed in such a way that an interpretation rule can be written for the text. That is we can write a rule for a computer of the form given two terms $t_1$ and $t_2$ occurring in a textual environment $E_i$, then add to the database, “knowledge” of the form $<t_1 R_j t_2>$ where $R_j$ is a specific ontological relation previously defined.

In practical terms, this means that for a given ontological relation $R_j$ there will a set $\{E_1 \ldots E_n\}$ of one or more lexico-syntactic environments of the type recognised by Hearst [1992] and others. These lexico-syntactic environments have already been discussed briefly above (cf. Section 6.3.3), and below we will catalogue and describe a number of these patterns in greater detail. Like all linguistic phenomena, our expectation is that even though our definition of explicit knowledge is quite rigorous and limiting, there will be more precise patterns and less precise or vaguer ones; there will be ones which are guaranteed to provide ontological knowledge and those with only a certain degree of certainty.

The definition of explicit knowledge we have provided makes no cognitive or philosophical claims. It is an engineering solution, whose limitations will become apparent in the course of further research. As we noted above (Section 2.1), explicit knowledge in philosophical terms corresponds to propositional knowledge or knowledge which can be expressed verbally. There are many forms of verbal knowledge other than those we will consider or be able to handle here, largely due to the limitations of Natural Language Processing in its ability to interpret language. On the other hand, there is no exact correspondence between an instantiated instance of a lexico-syntactic pattern and a
given proposition. We will use such instances as evidence for the con-
struction of ontologies, which may or may not express propositional
knowledge depending on one’s perspective.

7.2 A PROVISIONAL CATALOGUE OF SOME LEXICO-SYNTACTIC PAT-
TERNs

In this section we catalogue a number of lexico-syntactic patterns
which are cited in the literature. In each case, we have a) specified
what ontological knowledge the pattern provides, b) described the
pattern in an abstract form, c) given a couple of examples, d) tried
to identify research which describes or uses them, and, where ap-
propriate, e) provided comments. Sources in the literature cannot
claim to be complete, but they may be of use in future research. In all
likelihood many of these lexico-syntactic patterns have been identi-
fied independently by a number of authors for a number of different
purposes.

Included here are patterns and examples which appear to be ap-
pllicable only for the identification of instances of a concept (e.g. the
Hilton hotel) because we believe the distinction between concept or
class and instance to be fraught with difficulties.

A further difficulty which needs to be noted concerns the internal
structure of the NPs. In many cases, the NP is a simple lexical item
(one word) e.g. injuries, wounds, proteins in the examples below. How-
ever, in a large number of cases the NPs are complex and the issue
arises as to what is the unit which is being recognised. For example,
in the first examples below do we wish to extract bones or broken bones?
Clearly the the latter because bones are not injuries. In the case of
common-law countries, should one be identifying Canada as a country
or common-law country? In this case again the latter appears to be
acceptable. But in the case of other important civic buildings, we would
want to filter out other important as lexical items which do not form
part of the basic concept or term. The fundamental problem is one of
term recognition which plays a key role in ontology learning because it
is identifying the fundamental units we seek to organise. the present
work has avoided addressing the problem of term recognition directly
as there is substantial research already on the area [EAGLES Lexicon
Interest Group, 1998, Ahmad, 2000] etc. We have chosen for practical
purposes in this thesis to focus just on the head of the NP, while
fully aware of the limitations. We are currently researching this issue
[Zhang et al., 2008], and this will be addressed more fully in future
work.

NOTATION:
NP = Noun phrase (singular or plural), NPS = Singular Noun Phrase,
NPP = Plural Noun Phrase.
The subscripts p and c correspond to ‘parent’ and ‘child’ here:
7.2.1 Patterns for identifying hyponymy or the ISA relation.

1. \{\text{such NPP}_p|\text{NPP}_p \text{ such as } \{\text{NP}_c_1, \text{NP}_c_2\{or|and\} \text{ NP}_c_n\}\ ...
   e.g.
   ... works by such authors as Herrick, Goldsmith, and Shakespeare.
   ... hotels such as the Four Seasons ...
   ... such injuries as broken bones ...
   SOURCE: Hearst [1992]
   USED BY: Cimiano et al. [2004], Etzioni et al. [2004]
   Note: Plural subject

2. \text{NP}_p, \text{ such as } \{\text{NP}_c_1, \text{NP}_c_2\{or|and\} \text{ NP}_c_n\}, \text{ is } ...
   e.g.
   The bow lute, such as the Bambara ndang, is plucked ...
   SOURCE: Hearst [1992]
   USED BY: Cimiano et al. [2004]

3. \text{NPP}_c\{,\text{NPP}_c\}*\{,\} \{and|or\} \text{ other NPP}_p
   e.g.
   ... malignant melanomas and other cancer cell types ...
   ... temples, treasuries, and other important civic buildings.
   ... bruises, wounds, broken bones or other injuries ...
   SOURCE: Hearst [1992]
   USED BY: Cimiano et al. [2004], Etzioni et al. [2004]
   Note: Note plurals

4. \text{NPP}_p\{,\} \{including|especially\} \{\text{NP}_c,\} + \{or|and\}\text{NP}_c
   e.g.
   All common-law countries, including especially Canada and England ...
   ... most European countries, especially France England and Spain.
   SOURCE: Hearst [1992]
   USED BY: Cimiano et al. [2004], Etzioni et al. [2004]
   Note: Note plural

5. \text{NP}_c, \text{ a NP}_p \text{ that } ...
   e.g.
   ... isolation and characterisation of pbp, a protein that interacts.
SOURCE: Saggion and Gaizauskas [2004]¹
Note: Difficult to use in search engines like Google because punctuation is usually filtered out.

6. $NP_c$ is a $NP_p$ . . .
   e.g.
   ... aspartame is a sweetener ...  
   ... bile is a green liquid ...  
   SOURCE: Pearson [1998]
   USED BY: Saggion and Gaizauskas [2004], Cimiano et al. [2004], Etzioni et al. [2004]
   Note: The basic form of the copula.

7. $NPP_c$ are $NPP_p$s . . .
   e.g.
   ... birds are animals ...  
   SOURCE: Saggion and Gaizauskas, 2004

8. $NP_c\{,\}$ a $\{type|kind\}$ of $NP_p\{,\}$ . . .
   e.g.
   ... aspartame, a type of sweetener, ...  
   SOURCE: Saggion and Gaizauskas, 2004
   USED BY: Cimiano et al. [2004]
   Note: This form is called apposition.

9. $NP_c$ is a $\{type|kind\}$ of $NP_p\{,\}$ . . .
   e.g.
   ... aspartame is a type/kind of sweetener, ... 
   SOURCE: Saggion and Gaizauskas, 2004

10. $NPP_c$ are$\{types|kinds\}$ of $NPP_p\{,\}$ . . .
    e.g.
    ... birds are types/kinds of animals ...  
    SOURCE: Saggion and Gaizauskas, 2004

11. the $NP_c$ $NP_p$ . . .
    e.g.
    ... the Hilton hotel ...  
    SOURCE: Cimiano et al. [2004]

¹ Complemented by personal communications.
12. the $\text{NP}_p$, $\text{NP}_c$ ...  
   e.g.  
   ...the hotel Chelsea...  
   SOURCE: Cimiano et al. [2004]  
   NOTE: Cimiano et al. use the above two for ontology population - I am not sure whether they are applicable for ontology construction because distinguishing the two depends on using NE recognition, and I do not think we use patterns such as ‘the bird robin ...’ in English, i.e. $<$common noun$>$ $<$common noun$>$ as opposed to $<$common noun$>$ $<$proper noun$>$.

13. ...the $\text{NP}_p$ [formerly] known as $\text{NP}_c$ ...  
   e.g.  
   The Socioeconomic Group Formerly Known As “Geeks”.  
   The idiot known as Bush  
   SOURCE: [Saggion and Gaizauskas, 2004]  
   Note: Better for Ontology Population.

14. $\text{NP}_c$ consists of $\text{NP}_p$  
   e.g.  
   ...a nephron consists of a cup-shaped, hollow Bowman’s capsule ...  
   Note: Pearson notes that this pattern introduces super-ordinate terms only in certain types of corpora. Otherwise it usually introduces a list of component parts, usually identifiable by number or followed by the conjunction and.

15. $\text{NP}_c$ is defined as $\text{NP}_p$  
   e.g.  
   A circuit is defined as the complete transmission path ...  
   Note: Very dependent on corpus type. An examination of Google results showed much noise.

16. $\text{NP}_p$ is called $\text{NP}_c$  
   e.g.  
   Animals which feed on plants are called herbivores.  
   Transceivers that meet warm-start requirements are called warm-start transceivers.  
   Note: The utility of this is dependent on being able to parse fully the parent NP and extract the head.
7.2.2 Patterns for identifying SIBLINGS

The importance of these patterns lies in that given knowledge about one sibling, it can be assumed to apply to other siblings.

NOTATION:
Here s1 and s2 refer to respective siblings.

1. NP\textsubscript{s1}, also known as NP\textsubscript{s2} {or NP\textsubscript{s3}}, . . .
   e.g. Liming, a Limed wood finish also known as Pickled or “Whitewash”
   SOURCE: [Saggion and Gaizauskas, 2004]

2. NP\textsubscript{s1} is {a|an} NP\textsubscript{p}, and is also known as NP\textsubscript{s2} {or NP\textsubscript{s3}}, . . .
   e.g. Acne is an inflammatory disease of the skin and is also known as Acne Vulgaris.
   SOURCE: [Saggion and Gaizauskas, 2004]

3. NP\textsubscript{s1}, {,|()} also called NP\textsubscript{s2} {or NP\textsubscript{s3}}
   e.g. King Catfish, also called Narmer By Marie Parsons
   e.g. Stomach ulcers (also called peptic and duodenal ulcers)
   Note: Parsing problem with the later example.
   SOURCE: [Saggion and Gaizauskas, 2004]

7.2.3 Patterns for identifying ATTRIBUTES

Patterns for PART/WHOLE attributes

Subscript \textit{w} indicates \textit{whole}, subscript \textit{p} indicates \textit{part}.

1. NP\textsubscript{w}‘s NP
   e.g. ... building’s basement ...
   SOURCE: Berland and Charniak [1999]

2. NP\textsubscript{p} of {the|a}NP\textsubscript{w}
   e.g. ... basement of a building ...
   SOURCE: Berland and Charniak [1999]

3. NP\textsubscript{p} in {the|a}NP\textsubscript{w}
   e.g. ... basement in a building ...
   SOURCE: Berland and Charniak [1999]
4. NP$_{p}$s of NP$_{w}$s
   e.g. ... basement of buildings ...
   SOURCE: Berland and Charniak [1999]

5. NP$_{p}$s in NP$_{w}$s
   e.g. ... basements in buildings ...
   SOURCE: Berland and Charniak [1999]

7.3 THE RELIABILITY OF LEXICO-SYNTACTIC PATTERNS

Let us suppose that we know that two terms are strongly associated for one reason or another, e.g. the application of methods discussed in Section 6.1. In order to establish the nature nature of the relationship between the two terms, one or more queries need to be issued to the corpora which are being used. If no instances are found of a specific instantiated lexico-syntactic pattern, this may be due to inadequacies of the corpus such as that it is too small or of the wrong type for that kind of lexico-syntactic pattern. If instances are found, then what confidence can we have that this represents a valid observation. For example, the instantiated pattern cat is a pet has 171 hits on Yahoo$^2$ On the other hand, pet is a cat has 731 hits. On purely statistical grounds, we should construct an ontology where PET ISA CAT!

Thus it would be of considerable use to be able to issue a query to a corpus and for each lexico-syntactic pattern to have a previously obtained confidence level$^3$. In Table 17, we present the hit counts from instantiating a selection of lexico-syntactic patterns from those presented above. The patterns were instantiated with terms from two mini-ontologies consisting of the terms \{ pet, dog, cat \} and \{ vehicle, car, van \}. There are a number of issues to raise concerning these patterns and the use of any particular one. First, we need to know, given that a pattern has a hit, whether it is a good indicator of the validity of the ontological relation. One way of expressing this is in terms of precision (recall is impossible to calculate because we are using the Web). If we take the data presented in the table as complete, then we can say that the figures in the white columns are true positives (tp) and the figures in the grey columns are false positives (fp). Thus if precision = $\frac{tp}{tp+fp}$ (cf. Manning and Schütze 1999, :267ff), we can rank the precision values for the lexico-syntactic patterns presented in Table 17 as shown in Table 18.

In considering Table 17, we observe that some patterns produce a large number of false positives, while other patterns produce very few or none. A high precision on its own may still result in lots of false positives. For example, the pattern car is a vehicle has 123 hits on Yahoo$^2$ On the other hand, vehicle is a car has 731 hits. On purely statistical grounds, we should construct an ontology where VEHICLE ISA CAR!

Thus it would be of considerable use to be able to issue a query to a corpus and for each lexico-syntactic pattern to have a previously obtained confidence level. In Table 17, we present the hit counts from instantiating a selection of lexico-syntactic patterns from those presented above. The patterns were instantiated with terms from two mini-ontologies consisting of the terms \{ pet, dog, cat \} and \{ vehicle, car, van \}. There are a number of issues to raise concerning these patterns and the use of any particular one. First, we need to know, given that a pattern has a hit, whether it is a good indicator of the validity of the ontological relation. One way of expressing this is in terms of precision (recall is impossible to calculate because we are using the Web). If we take the data presented in the table as complete, then we can say that the figures in the white columns are true positives (tp) and the figures in the grey columns are false positives (fp). Thus if precision = $\frac{tp}{tp+fp}$ (cf. Manning and Schütze 1999, :267ff), we can rank the precision values for the lexico-syntactic patterns presented in Table 17 as shown in Table 18.

In considering Table 17, we observe that some patterns produce a large number of false positives, while other patterns produce very few or none. A high precision on its own may still result in lots of false positives. For example, the pattern car is a vehicle has 123 hits on Yahoo$^2$ On the other hand, vehicle is a car has 731 hits. On purely statistical grounds, we should construct an ontology where VEHICLE ISA CAR!

---

$^2$ June 2005, Yahoo is used in preference to Google because the counts are considered more accurate, cf. http://aixtal.blogspot.com/2005/01/web-googles-counts-faked.html

$^3$ This is partly the motivation for the work of Cafarella and Etzioni [2005].
### 7.3 The Reliability of Lexico-Syntactic Patterns

The reliability of lexico-syntactic patterns can be assessed through hit counts for various instantiated patterns. Here are the hit counts for two sets of patterns involving "pet" and "vehicle" terms:

<table>
<thead>
<tr>
<th>Parent term</th>
<th>Child term</th>
<th>Pattern Description</th>
<th>Hit Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>pet</td>
<td>dog</td>
<td>&lt;Cs&gt;, a type of &lt;Ps&gt;</td>
<td>0 0 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt;, a kind of &lt;Ps&gt;</td>
<td>0 1 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; is a kind of &lt;Ps&gt;</td>
<td>6 2 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; is a type of &lt;Ps&gt;</td>
<td>24 2 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Pp&gt; such as &lt;Cp*&gt;</td>
<td>18300 10400 4 8 0 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; and other &lt;Pp&gt;</td>
<td>312 383 16 327 0 41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; and other &lt;Pp&gt;</td>
<td>43000 88000 44 6010 0 1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; or other &lt;Pp&gt;</td>
<td>443 1380 3 7 0 9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; or other &lt;Pp&gt;</td>
<td>4030 1640 54 318 1 351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TOTAL</td>
<td>66115 101808 121 6670 1 1402</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parent term</th>
<th>Child term</th>
<th>Pattern Description</th>
<th>Hit Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>vehicle</td>
<td>car</td>
<td>&lt;Cs&gt;, a type of &lt;Ps&gt;</td>
<td>4 1 1 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt;, a kind of &lt;Ps&gt;</td>
<td>5 0 0 0 2 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; is a kind of &lt;Ps&gt;</td>
<td>236 2 1 0 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; is a type of &lt;Ps&gt;</td>
<td>74 0 1 1 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Pp&gt; such as &lt;Cp*&gt;</td>
<td>2970 1050 4 4 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; and other &lt;Pp&gt;</td>
<td>878 64 17 43 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; and other &lt;Pp&gt;</td>
<td>44700 1550 104 115 1 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; or other &lt;Pp&gt;</td>
<td>349 37 0 1 0 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;Cs&gt; or other &lt;Pp&gt;</td>
<td>1270 221 1 2 1 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TOTAL</td>
<td>50486 2925 129 166 4 7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OVERALL</td>
<td>116601 104753 250 6836 5 1409</td>
</tr>
</tbody>
</table>

Table 17. Table showing hit counts for various instantiated lexico-syntactic patterns. From Yahoo, June 2005
Table 18. Table ranking the precision and probability of false positives for lexico-syntactic patterns, based on the data presented in Table 17.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;Pp1&gt;</code> such as <code>&lt;Cp&gt;</code></td>
<td>0.99</td>
</tr>
<tr>
<td><code>&lt;Cs1&gt;</code> is a kind of <code>&lt;Ps1&gt;</code></td>
<td>0.99</td>
</tr>
<tr>
<td><code>&lt;Cs&gt;</code> or other <code>&lt;Pp1&gt;</code></td>
<td>0.99</td>
</tr>
<tr>
<td><code>&lt;Cs1&gt;</code> is a type of <code>&lt;Ps1&gt;</code></td>
<td>0.98</td>
</tr>
<tr>
<td><code>&lt;Cp</code> and other <code>&lt;Pp1&gt;</code></td>
<td>0.96</td>
</tr>
<tr>
<td><code>&lt;Cp</code> or other <code>&lt;Pp1&gt;</code></td>
<td>0.90</td>
</tr>
<tr>
<td><code>&lt;Cs1&gt;</code>, a type of <code>&lt;Ps1&gt;</code></td>
<td>0.83</td>
</tr>
<tr>
<td><code>&lt;Cs&gt;</code> and other <code>&lt;Pp1&gt;</code></td>
<td>0.79</td>
</tr>
<tr>
<td><code>&lt;Cs1&gt;</code>, a kind of <code>&lt;Ps1&gt;</code></td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 19. Examples of instantiated lexico-syntactic patterns used as queries on Yahoo.

<table>
<thead>
<tr>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>sovereign, a type of tsar</td>
</tr>
<tr>
<td>fumes is a type of gas</td>
</tr>
<tr>
<td>anatomist and other experts</td>
</tr>
<tr>
<td>person and other anatomists</td>
</tr>
<tr>
<td>anatomist and other persons</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

positives if the number of matches is very high. So it is more useful to know that a given pattern returns no false positives (i.e. a precision rate of 1.0) because this allows one to minimise the total number of queries issued to determine the validity of an ontological relation. There is a problem here because while patterns such as `<Cs1>`, a type of `<Ps1>` and `<Cs1>`, a kind of `<Ps1>` have almost no false positives, their overall frequency is low. The most useful pattern on the basis of this body of evidence appears to be `<Pp1>` such as `<Cp>`.

In order to understand further the accuracy of the lexico-syntactic patterns we conducted two further experiments in order to approximate the precision and recall of the patterns. It should be stressed that in both cases we treat the results as indicative rather than definitive for number of reasons.

**Experiment 1: Precision**  In this experiment, we used the Yahoo API (http://developer.yahoo.net/), and for each of the nine patterns listed in Table 18, we issued the phrase as a query, manually scanned each hit for the pattern, decided whether or not it was actually an example of this pattern (e.g. for “is a type of” that it was preceded and followed by an NP and not a pronoun), and then judged whether the inferred ontological knowledge would be considered valid. Thus the hit/citation “Guttate psoriasis is a type of psoriasis that looks like
7.3 The reliability of lexico-syntactic patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Cs&gt; or other &lt;Pp1&gt;</td>
<td>0.84</td>
</tr>
<tr>
<td>&lt;Cp&gt; and other &lt;Pp1&gt;</td>
<td>0.80</td>
</tr>
<tr>
<td>&lt;Cp&gt; or other &lt;Pp1&gt;</td>
<td>0.79</td>
</tr>
<tr>
<td>&lt;Cs1&gt; is a type of &lt;Ps1&gt;</td>
<td>0.76</td>
</tr>
<tr>
<td>&lt;Cs1&gt;, a type of &lt;Ps1&gt;</td>
<td>0.74</td>
</tr>
<tr>
<td>&lt;Cs&gt; and other &lt;Pp1&gt;</td>
<td>0.70</td>
</tr>
<tr>
<td>&lt;Pp1&gt; such as &lt;Cp*&gt;</td>
<td>0.69</td>
</tr>
<tr>
<td>&lt;Cs1&gt; is a kind of &lt;Ps1&gt;</td>
<td>0.53</td>
</tr>
<tr>
<td>&lt;Cs1&gt;, a kind of &lt;Ps1&gt;</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 20. Precision of lexico-syntactic phrases based on random sample.

small, salmon-pink drops ....” would be accepted as correct (True Positive), the hit “A cross is a type of flag ....” would be marked as false (False Positive), while “This is a type of nuclear scanning test ....” would be ignored as it contains a pronoun (there is a range of phenomena we assume can be filtered out relatively easily). The results are shown in Table 20.

There is clearly a difference between the results obtained in this experiment and those obtained before and presented in Table 18. The reason for this is that the manner of querying is different. In the former case, each lexico-syntactic pattern was instantiated with a specific pair of terms, so the query was more specific. In the latter case, the lexico-syntactic patterns were not instantiated, and so reflect the performance of these phrases in the wild.

**Experiment 2: Recall**

In principle it is impossible to calculate recall figures for knowledge extraction from the web as we cannot know a priori what there is “out there.” However, we can make an estimate which as in the previous experiment, we stress it is indicative and far from definitive. In this experiment we followed the following procedure: We made a random selection of 200 nouns from the BNC corpus with a frequency greater than 5 and checked if they were present in WordNet4. Given that they were present, we identified the three immediate hypernyms of the noun, and constructed pairs of terms - noun + hypernym1, noun + hypernym2, etc. Thus we had a random set of 600 pairs of terms. Each of the nine lexico-syntactic pattern was instantiated with each pair of terms thus giving a 600 “gold standard” instantiated phrases. We then queried the web using the Yahoo API. The results are shown in Table 21.

There are a number of observations to be made from these results: We note there is no lexico-syntactic pattern which gives no false positives at all. Thus we cannot ever be entirely sure that if we issue a

---

4 Using the frequency tables provided by Adam Kilgariff http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html
query and it returns some positive number that it is not wrong. This is obviously due to the size and noise on the web. Further experiments are needed to determine if properly parsing the citations from the web would reduce the error rate and by how much.

Recall is extremely low, ranging from 0.012 to 0.23, which in view of the large scale of the web means that finding ‘explicit’ knowledge is hard and rare even when using such a large resource. This may be due to a number of factors such as the inclusion of a number of proper names in the data set and the type of hypernym pairs that were being used. Thus if we imagine a chain of terms from leaf to root, it may be that random choices of pairs on that chain is not the best way to find ontological relations which are actually expressed. It may be necessary to identify certain specific levels in the hypernym hierarchy which are amenable to this approach.

In general, we can show that some lexico-syntactic patterns are reasonably reliable in confirming the validity of extracted knowledge, but the sparsity with which certain items of knowledge appear means that extraction patterns cannot be used to prove that a specific ontological relationship is absent or invalid. In effect, absence of evidence is not evidence of absence.

### 7.4 Extraction Patterns

Lexico-syntactic Patterns are specific detailed templates which match a sequence of words and annotations - typically POS speech tags but not necessarily. They have an extensive history in NLP, having been used for a variety of purposes including question answering [Saggion and Gaizauskas, 2004] and obtaining facts from the web (cf. KnowItAll Section 6.4.4). However, we propose that they are merely the most precise from range of “Extraction Patterns” all of which obtain information from the text. Under this more abstract view, we see a hierarchy of Extraction Patterns, starting from a very simple level and
moving to a more complex and more ontologically precise level.

The most basic level (or Level 0) is an Extraction Pattern which identifies that a given term is specifically related to the domain. We use the term *domain* loosely and depending on the purpose a level 0 EP identifies a term as belonging to the corpus, to the subject area, or to the ontology. From this perspective, term recognition is the basic level 0 Extraction Pattern. All that term recognition does is establish that there is greater relationship between a certain set of terms and the domain (ontology etc.) than with the language as a whole. Typical term recognition methodologies use a reference corpus to identify the most discriminating terms in the domain corpus [Ahmad, 2000, Velardi et al., 2001]. Of course a variety of methodologies can be applied to perform term recognition. The output of such an EP can be visualised as shown in Figure 12.

![Figure 12. The result of applying the Level 0 Extraction Pattern (Term Recognition)](image)

At the next level in the hierarchy, Level 1, Extraction Patterns establish that there is some sort of relationship between two terms and that that relationship is greater than with other terms. Of course, such Term Association algorithms will often rank the degree of association (cf. Section 6.1 and 6.2.1). The establishment of some sort of relationship is important both in limiting the work of an ontology engineer (i.e. guiding their input) and more importantly in the context of applying Extraction Patterns at Level 2 (lexico-syntactic patterns) in limiting the search space. If a corpus has (only) 100 candidate terms, then theoretically there are 9900 bigrams to potentially consider for possible relations. This set of pairs with potential ontological relations can become rapidly unmanageable if we cannot select the best candidates. Thus following the application of Level 1 EPs, we might theoretically obtain a situation as shown in Figure 13 (ignoring issues of data sparsity and a number of other issues, for the moment).

---

5 We preclude reflexive relations on the basis that “A cat is a cat” would be uninteresting for knowledge acquisition purposes and highly unlikely to occur.
Finally at the last level (Level 2), lexico-syntactic Extraction Patterns specific to particular ontological relations can be applied. These EPs are of the type discussed above and represent the most explicit recognition and definition of ontological relations. Thus for each ontological relation (ISA, PART_OF, LOCATED_IN, etc.) a set of EPs apply, and where they do apply successfully then the ontological relationship between the terms can be labelled. This will tend to be a subset of the terms identified and associated in the earlier steps, resulting in something like Figure 14.

At each level, some degree of knowledge is being extracted from the texts. For example, the level 1 structures of purely word association capture significant knowledge (as Schvanefeldt would argue, cf. Section 6.2.1) even if the lack of labelling still means that the precise nature of the knowledge is still vague. Following the application of relation labelling, we clearly reach a degree of explicitness where reasoning systems can be used more effectively.

### 7.4.1 Learning of Extraction Patterns

The lexico-syntactic patterns catalogued above (Section 7.2) must be considered only a partial set. There are a great many other such lexico-syntactic patterns but few have the same degree of generality. These other patterns are specific to certain styles, subjects or disciplines and usually are able to provide high precision but focussed on a specific
domain. For example, [Etzioni et al., 2005] provide the example of the film <film> starring and headquartered in <city> as being patterns with with precision and (for their topic) high recall. Equally in a legal corpus (“Statutes and Statutory Instruments of England and Wales”), there is repeated use of the pattern “"<someterm> means <defining phrase>". As Etzioni et al. note the ability to learn such domain specific patterns is vital in order to increase the number of sentences from which one can extract ontological knowledge. There are a number of possible approaches to extraction pattern learning which we will review in this section.

The automatic generation of extraction patterns has a long history in Information Extraction going back at least to the work of Riloff’s AutoSlog-TS [Riloff, 1996]. This was an extension of her AutoSlog system that built extraction patterns for Information Extraction using manually tagged text [Riloff, 1993]. Instead in AutoSlog-TS, a corpus of texts was used which had been classified into relevant and non-relevant for the given domain (e.g. terrorism, or management succession). The system applies a syntactic analyser, and then applies a set of 15 ‘heuristics’ which are templates for extraction patterns such as ‘< subj > active − verb dobj’ to construct a dictionary of extraction patterns of the type ‘< subj > bombed embassy’. The training corpus is then used again to calculate relevance statistics for each pattern so as to allow a ranking of these patterns. The approach
112 Explicit Knowledge and Lexico-Syntactic Patterns

proposed by Riloff was not extremely successful as the F-Measure ranged from 0.38 to 0.44 for the terrorism domain on the MUC-4 corpus, but this is a measure of the information extracted not an evaluation of the extraction patterns per se. This work appears to have laid the foundation for the idea of automatically generating extraction patterns. It should be noted that her worked was focussed on IE and thus addressing a slightly different objective from ontology learning although clearly closely related.

In order to reduce the need for manual tagging in information extraction systems, Brin, proposed in the DIPRE system [Brin, 1998], to use a small set of seeds occurrences of which are sought in a large body of texts (typically the Web). These occurrences are then used to learn extraction patterns, which in turn are used to generate new seeds. Such an approach is crucially dependent on the evaluation metrics used to evaluate both the extraction patterns generated and the seeds or knowledge tuples extracted. An important descendant of both AutoSlog-TS and DIPRE is the SnowBall system developed at Columbia University [Agichtein and Gravano, 2000, Agichtein et al., 2001]. This used a similar approach to DIPRE in using seeds which where used to find extraction patterns which then found further seeds (cf. Figure 15).

The authors focussed exclusively on organisations and locations and so were able to improve their performance over previous work by applying a named entity tagger. This allowed them to restrict extraction pattern generation to environments of the form '<ORGANIZATION’s headquarters in <LOCATION>'. For each candidate pattern P they define their ‘confidence’ following the standard precision formula, where \(P_{\text{positive}}\) is the number of positive matches, \(P_{\text{negative}}\) the number of negative matches, and \(\text{Conf}(P)\) the confidence level value:

\[
\text{Conf}(P) = \frac{P_{\text{positive}}}{(P_{\text{positive}} + P_{\text{negative}})}
\]  
(7.1)

For example, the pattern \(P = <{}>, \text{ORGANIZATION}, <"",">,1>, \text{LOCATION},{}>,\) matches the following texts:
Exxon, Irving, said . . .
Intel, Santa Clara, cut prices . . .
invest in Microsoft, New York-based analyst Jane Smith said

then because the tuple \(< \text{Microsoft, NewYork} >\) conflicts with the existing tuple \(< \text{Microsoft, Redmond} >\), it is treated as a negative example and so \( \text{Conf}(P) = 0.67 \). The confidence the system has in each extraction pattern is changed on each iteration of the system using a formula which controls the learning rate and can be set to be conservative and trust new examples less than past examples.

Agichtein and Gravano present a complex metric to evaluate their system in terms of the precision and recall of information extraction tuples. It is based on a gold standard i.e. an ideal set of organisation-location tuples. They do not present figures which directly evaluate the patterns generated by the system but their recall of around 0.85 and precision of about 0.90 show that their approach is very successful. The most important limitation is the restriction of the system to organisation-location pairs so this puts in question just how general their approach is.

KnowItAll (cf. Section 6.4.4) built further on the SnowBall approach seeking to learn extraction patterns for a wider range of ‘facts’ or pieces of knowledge. Their approach to pattern learning involves the following steps:

- Start by using a set \( I \) of seed instances (i.e. terms which are instances of some class - note these are not seed patterns) which have been identified by generic extraction patterns.
- For each instance (or term) \( i \in I \) issue a web query and for each occurrence of \( i \), record a citation line with \( w \) before and after the occurrence of \( i \). They used \( w = 4 \).
- Select the best patterns in accordance with a given metric. A pattern is defined as any substring from the citation line including the term and some other word.

The metric used is dependent on calculating a candidate EP’s precision and recall. Estimating precision is particularly difficult due to the lack of labelled negative examples. So basing themselves on the work of Yangarber, they consider the positive examples for one class the negative examples for another. Thus they propose the following two formulas for calculating precision and recall:

\[
\text{EstimatedPrecision} = \frac{c(p) + k}{c(p) + n(p) + m} \quad (7.2)
\]

\[
\text{EstimatedRecall} = \frac{c(p)}{S} \quad (7.3)
\]

where for a given pattern \( p \), \( c(p) \) is the number of distinct seeds from that class that a pattern as been found for, and \( n(p) \) is the number of
distinct seeds from other classes. \( S \) is the total number of seeds in the target class and \( k \) and \( m \) are constants which are prior estimates of precision and are used to perform a Laplace correction [Etzioni et al., 2005]. In view of this, patterns are selected which occur for more than one seed which means that in Equation 7.3, \( \text{EstimatedRecall} > \frac{1}{S} \). According to the authors this eliminated 96% of all candidate patterns. The remaining patterns are sorted using Equation 7.2. They chose the top 200 patterns which satisfied the first criterion and then were ranked most highly by the second criterion. They reported precision ranging from 0.65 to 0.80.

Automated approaches to pattern learning show great promise and certainly fall within the range of performance commonly accepted for NLP techniques. Challenges remain concerning how much data is needed to learn a new pattern with a reasonable degree of confidence, and also how applicable such techniques are across a range of ontological relations. Some relations are more often explicitly expressed (in the sense we defined earlier in this chapter) while other ontological relations are more often implicit. There is still considerable work to be undertaken to identify the boundaries of what is possible here.

7.5 Conclusions

In this chapter, we have defined explicit knowledge as special form of textual phenomena where the ontological relationship between two terms is sufficiently explicit as to be recognisable by Natural Language Processing techniques. Our objective has been make it possible to label ontological relations between terms. In order to achieve this we have presented an incomplete catalogue of lexico-syntactic patterns, which correspond to specific ontological relations. For each significant ontological relationship, it should be possible to construct, either manually or automatically (cf. Brewster et al. 2002) a set of lexico-syntactic patterns which allow the explicit capture of that form of knowledge.

In order to make the use of these lexico-syntactic patterns more effective, we attempted to identify which had greater precision. We performed a small experiment with only two ontologies of three terms each, and this indicated that there was a clear divergence in the performance of different patterns. A larger experiment was conducted with a greater number of terms and this showed that although reasonable precision could be achieved recall was very low. Significantly there a number of methods in the literature for learning extraction patterns which have a high degree of success and these approaches need to be harnessed in order to make the identification of explicit ontological contexts more effective.
Every philosophy is tinged with the colouring of some secret imaginative background, which never emerges explicitly into its train of reasoning.

Alfred North Whitehead [1926]

8.1 ONTOLOGIES AND TEXT

In previous chapters (Chapter 6 and Chapter 7), we have presented computational techniques and tools which can be used in constructing ontologies from a body of textual data. However, none of the techniques nor the complete systems considered are able to input text at one end and produce ontologies at the other. In the preceding chapter, we identified Hearst-type lexico-syntactic patterns as a specific technique to extract explicit knowledge but we also showed that recall was very problematic. This might be seen as surprising: texts are used for the collection, storage and transmission of knowledge, so there should be a way to extract this knowledge and make it available in some machine-readable format. This has certainly not proved to be the case, but there has been insufficient research into why such knowledge extraction has proved so hard. There has usually been an acceptance by the knowledge acquisition community that any such process is ‘hard’, ‘laborious’ etc., or by those involved in Natural Language Processing that language is too complex or too irregular to permit the easy identification of knowledge items. The most advanced work in this area has tended to be in the form of information extraction (cf. Ciravegna and Wilks 2003, Dill et al. 2003) and yet it is widely accepted to have severe limitations.

Rather than just take all this for granted, in this chapter we will make a first attempt to consider the underlying nature of an ontology on the one hand, and the what is happening in a text on the other. Our challenge is to determine just how feasible it is to derive an ontology from texts given our closer understanding of the respective phenomena.

8.1.1 Ontologies and Background Knowledge

The most widely cited definition of an ontology is that of Gruber [1993] who said that an ontology was “a formal, explicit specification of a shared conceptualisation, used to help programs and humans share knowledge” (cf. Section 3.4 above). The important aspect we wish to focus on here is that it is a shared set of concepts. For any
given domain, the ontology is supposed to represent the concepts which are held in common by the participants in that domain. Thus it would appear that an ontology represents the background knowledge associated with a domain. It is this background knowledge to everyday life that Cyc [Lenat et al., 1994] attempts to encode, and most domain specific ontologies appear explicitly or implicitly to try to capture this knowledge. In a certain sense and without making any claims of psychological reality, there may be ontologies for each and every domain of human knowledge.

The idea of background knowledge is not a new one. It has played a role under this name or other closely related ones in a number of disciplines. There has been much philosophical discussion of what is called common knowledge, especially after Lewis’s work Convention (1969), where it is defined as the knowledge common to agents which is also known to be common, and also known to be known to be common etc. \textit{ad infinitum}. This is contrasted (in the philosophical and economic literature) with mutual knowledge which concerns external reality or so-called first order knowledge [Koessler, 2000]. Most of this work has been concerned with the interaction between ‘agents’ and how to account in philosophical terms for the choices that agents could make [Vanderschraaf and Sillari, Winter 2005]. This is not very far from the psycholinguistic and linguistic pragmatics concept of common ground which is the set of propositions assumed by a speaker to be shared by both the speaker and addressee [Stalnaker, 1978, Clark and Marshall, 1981]. In Clark and Marshall’s classic work, one of the foundations of common ground is ‘community membership’ by which they mean cultural communities such as “soccer fans, pipe fitters, speakers of French”. This conception of community membership appears to be echoed in the ‘shared conceptualisation’ of ontologies, and has re-appeared in the work on communities of practice in the context of the Semantic Web.

The most extensive research on the relationship between background knowledge and texts has been undertaken by educational psychologists concerned with literacy. They use the terms background knowledge and prior knowledge interchangeably. Prior/Background knowledge has been shown to correlate highly with reading comprehension including speed and accuracy of comprehension [Langer, 1984, Stevens, 1980, Dochy et al., 1999]. This we take as evidence that the reader brings to varying degrees some form of ontology to the reading task. Research has also shown that activating background knowledge before reading improves comprehension, which in our terms would be equivalent to helping someone find the correct domain specific ontology. In other cases, due to ‘inaccurate preconceptions’ it was more effective to teach wholesale the necessary background knowledge (cf. review of this research area in Strangman and Hall [2004]). From our perspective, the exact nature of the text and what it does has not been studied adequately by educational researchers either.

\footnote{It remains an open question whether ontologies as currently conceived are adequate representations of that background knowledge.}
8.1.2 Text and Knowledge Maintenance

What is the nature of a text? When a writer creates a text, they assume a number of things. There is a linguistic assumption concerning the language used and a cognitive assumption concerning the ability of the audience to understand the vocabulary and technical terms used. In effect, a writer assumes that the audience shares a large proportion of the knowledge that the author is basing their text upon. Obvious exceptions are in a detective story where the author does not assume the reader knows who the murderer is, and only tells us at the end. The writer assumes the knowledge is shared by default except where the writer is explicitly signalling a difference. “Communication is made efficient by language producers assuming that their audience or listeners will draw on knowledge that is ‘stored away’ within the mind and that a large amount of information can therefore by left unstated” Emmott [1999].

Given that a writer of a text assumes their audience has the same ontology, then the question arises as to what the purpose of an arbitrary text is. We would like to argue that one way of looking at a communicative act, a text, is to see it as an act of knowledge maintenance. If we take a Quinean view of the nature of knowledge, then the text could be said to re-enforce or re-adjust a particular part of the overall ‘force field’ of knowledge (cf. Section 2.2 above). There are three aspects to this:

- One aspect is that the text re-enforces the assumptions of background knowledge. It tells the reader which ontology to use to process the text (if we assume there exist different sub-ontologies) and re-enforces the knowledge structures of that ontology by the particular linguistic juxtaposition of concepts. For example, in the abstract quoted in Example 1, the use of the terms ‘motor neuron’, ‘innervate’, and ‘transcription factors’ immediately identify the domain and the respective background knowledge needed to understand the text.

- A second aspect is that the text alters the links, associations and instantiations of existing concepts. Thus a primary purpose of a text at some level is to change the relationship between existing concepts, or change the instantiations of those concepts. One way that texts provide ‘new’ information to the reader is by asserting a link previously absent, or by deleting a link previously in existence. This kind of activity can be seen as trying to restructure the domain ontology which is clearly another form of knowledge maintenance. Again in Example 16, the phrase “lim3 and islet constitute a combinatorial code that generates distinct motor-neuron identities” restructures the domain ontology.

- The third and most obvious way a text affects a domain ontology is by adding new concepts. The author may propose a new analytical concept or name a new procedure etc. and these acts of
Different classes of vertebrate motor neuron that innervate distinct muscle targets express unique combinations of LIM-homeodomain transcription factors, suggesting that a combinatorial code of LIM-homeodomain proteins may underlie the control of motor-neuron pathway selection. Studies of LIM-homeodomain genes in mouse, Drosophila melanogaster and Caenorhabditis elegans have revealed functions of these genes in neuronal survival, axon guidance, neurotransmitter expression and neuronal function. Our results provide evidence that lim3 and islet constitute a combinatorial code that generates distinct motor-neuron identities.

Figure 16. Example from Thor et al. [1997]

A shell script is nothing more than a sequence of shell commands stuffed into a text file.

Figure 17. Example from Schwartz and Christiansen [1997]

naming label new concepts and indicate their relationship with the rest of the ontology (as in Example 17). This is relatively rare in proportion to the overall quantity of text which is produced and read.

It would appear, if this analysis is correct, that the background knowledge of a domain is part of what the writer takes for granted. Consequently the background knowledge captured in an ontology is rarely if ever explicitly stated in a text. It is implicit and assumed to be part of the cognitive landscape by the author.

The notion of explicit in this context has a very precise meaning. By explicit, we mean that an ontological relationship between two terms is expressed in some lexico-syntactic pattern of the type first identified by Hearst and discussed in greater detail in Chapter 7 above.

If we accept this line of thought, we should find it extremely rare for background knowledge (i.e. the explicit specification of ontological relationships between two terms) to be explicitly expressed in any text (by this definition of explicit). This we would expect to be especially true of a scientific text or academic paper because they are prototypical attempts to alter a community’s accepted ontology. We would not expect this to be true of an introductory textbook, manual or glossary which by their nature do NOT assume the domain specific background knowledge (more accurately they assume a more general, possibly more top level, ontology).

It would follow that one would find a specification of an ontology at the borders of a domain. These borders might be in time or intellectual space. Thus, we might expect that when a concept or rather its cor-
responding lexicalisation is first introduced there will be statements defining or explicating the idea. On the other hand, when a concept is borrowed from one discipline into another, there again the term is likely to be defined. We will return to this subject below in Chapter 10.

8.2 The Assumed Knowledge Hypothesis

There are two points here. The foundation for efforts at automatically building ontologies from texts is the assumption that there are texts which do specify in a coherent manner the ontological relations one is interested in, and that these textual specifications can be read and processed by a machine. However, a number of efforts at automating the ontology learning process have encountered substantial difficulties due to data sparseness [e.g. Cimiano et al., 2005, Wu and Hsu, 2002]. No writer has attempted to explain the fact that on the one hand human beings use texts to communicate knowledge and on the other hand knowledge acquisition from texts is so very difficult.

Our hypothesis is that no matter how large our corpus, if it is domain specific, the major part of the domain ontology will not be specified because it is taken as given, or assumed to be part of the background knowledge the reader brings to the text. We should stress that this hypothesis is assumed to apply most obviously to ordinary texts without didactic intentions (such as textbooks or manuals), because if they are intentionally written to instruct they are likely to provide explicit definitions of terms and explanations. However, there are three potential problems with textbooks and manuals which mean that in practice their inclusion (from a theoretical perspective) does not matter. First, they tend to reflect knowledge at a particular point in time and often fall behind the changes occurring in a particular field. Second, explicit knowledge is very sparse because the writers tend to assume that if a definition is provided once that is sufficient. Finally there are significant areas of endeavour where there are no textbooks (as was the case in the field of nursing for example until the 1970s [Flaming, 2003]).

The hypothesis cannot be proved empirically because one could always imagine a larger collection of texts in a specific domain such that somewhere in it one might find the missing text which expresses the knowledge one is seeking to identify. However, experience has shown that a certain number of textual contexts (citations) are needed for the ontological knowledge to be explicitly available. Finally, we should add that at a certain level it is obvious that background knowledge is presupposed by specialised texts, but it is very important to keep this in mind when designing ontology learning systems and understanding their failings.
8.2.1 Experimental Evidence

In order to provide initial evidence in favour of the hypothesis proposed, we will show that it is impossible to reconstruct a ‘gold standard’ ontology like the Gene Ontology [Ashburner et al., 2000] from a relevant collection of texts. We chose ten arbitrary pairs of terms, five from the ‘higher’ regions and five from the ‘lower’ regions, i.e. at or close to the leaves of the tree. No scientific claims are made for the manner in which the terms were chosen or the distinction between the higher and lower regions of the ontology since the ‘depth’ (i.e. the number of steps between a leaf and the root) is immensely variable. In order to simplify matters further, we limited ourselves to terms which were related by the IS-A relation.

We chose as our corpus all articles from the journal Nature covering a period from 1996 to 2001. Our domain specific corpus was a subset of these which concerned genetic topics (e.g. included the words gene, genome, genetics, etc.). Thus of 13000 texts in the whole corpus the subcorpus consisted of 1300 texts, or about 10%. We chose this ontology-corpus pair, on the basis that it might be reasonable to use the Gene Ontology as a Gold Standard. We attempted to determine how much of such an ontology could be derived a priori from a corpus derived from the journal Nature.

We looked for simple lexico-syntactic patterns of the type first suggested by Marti Hearst. If the corpus was theoretically to be a potential source of knowledge, in order to construct an ontology, then, at least some of the time, ontological relationships would be expressed in the corpus by these sorts of patterns.

The results of our experiment are shown in Table 23 and show the frequency of distribution for our selected terms. It is clear from these figures that there was no question of looking for lexico-syntactic environments since the terms hardly ever occurred in each other’s environment.

So at this stage it appears reasonable to turn to other sources. In order to do this, we looked up each pair of terms in the following sources:

- Google (www.google.com)
- Google Glossary (http://labs.google.com/glossary)
- Encyclopaedia Britannica (www.britannica.com)
- Dictionary of Encyclopaedia Britannica (www.britannica.com)

The results varied enormously depending on the terms. Some pairs were only to be found in online versions of the Gene Ontology, while at the other extreme over 31000 citations could be found on Google. In each case, the citations were checked for Hearst type patterns which could be said to explicitly represent the ontological relation between
the terms\(^2\). The individual results are presented in the Appendix in Chapter A, and the overall results are shown in Table 22.

<table>
<thead>
<tr>
<th>Textual Source</th>
<th>Number of articles found</th>
<th>Clear specification of ontological relation (no. of cases out of the 10 pairs of terms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original corpus</td>
<td>0-2</td>
<td>0</td>
</tr>
<tr>
<td>Google citations</td>
<td>3 - 31,000</td>
<td>6/10</td>
</tr>
<tr>
<td>Encyclopaedia</td>
<td>0-11</td>
<td>2/10</td>
</tr>
<tr>
<td>Dictionary</td>
<td>0-1</td>
<td>3/10</td>
</tr>
</tbody>
</table>

Table 22. Summary Results

Using the internet directly provides the greatest likelihood of finding defining contexts. Using an encyclopaedia or dictionary appeared to be no guarantee that definitions relating the two terms will be found. 60% of terms were found in explicit contexts on the internet which is clearly a great improvement on 0% in the original corpus. The figure of 60% does imply a limit on what is likely to be explicitly expressed, although a more systematic survey may give more reliable results. These figures should be taken merely as indicative.

Even though this is a limited sample, the figures appear to show that data cannot be derived from a corpus in order to demonstrate the expected ontological relationships. Even in the case of ligase/enzyme, where the number of occurrences in the text was significant, the number of contexts where they co-occurred was very small. This could be seen as a simple data sparseness problem and the obvious solution is to increase the size of the corpus. In accordance with Zipf’s observations, the ‘long tail’ of vocabulary will always occur too infrequently to provide enough opportunities for knowledge to occur explicitly.

We conclude that events, where terms co-occur in a domain specific corpus, are too sparse to provide sufficient opportunities for machine interpretable contexts to arise. Only by accessing external sources of information is there a significant increase in usable contexts.

---

\(^2\) This checking was done manually and where there was a large number of citations from Google we read through the first 100 approximately.
### Table 23. Frequency Distribution of the Terms Selected.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Frequency</th>
<th>Common environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>histolysis <em>isa</em></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tissue death</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>flocculation <em>isa</em></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cell communication</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>vasoconstriction <em>isa</em></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>circulation</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>holin <em>isa</em></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>autolysin</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>aminopeptidase <em>isa</em></td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>peptidase</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>death <em>isa</em></td>
<td>654</td>
<td>0</td>
</tr>
<tr>
<td>biological process</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>metallochaperone <em>isa</em></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chaperone</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>hydrolase <em>isa</em></td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>enzyme</td>
<td>672</td>
<td>2</td>
</tr>
<tr>
<td>ligase <em>isa</em></td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>enzyme</td>
<td>672</td>
<td>2</td>
</tr>
<tr>
<td>conotoxin <em>isa</em></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>neurotoxin</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
8.3 SOURCES OF ONTOLOGICAL KNOWLEDGE

There are a number of potential sources of such ontological knowledge, all of which present certain challenges:

- Encyclopaedias: they might appear ideal sources for ontological knowledge. Clearly they include defining or explanatory texts which could be mined. The main problems are the difficulty of access to encyclopaedias and the fact that they are not likely to be very up to date. However, one can expect that the kind of background knowledge we are interested in does not change that rapidly so absolute currency is not necessarily an absolute necessity. One possibility is to use an open access encyclopaedia such as Wikipedia (www.wikipedia.org). This has the advantage of accessibility and in certain regards currency (since the many volunteers tend to keep it up to date). One possible disadvantage is that certain areas of knowledge will be far better represented than other due to the vagaries of fashion and attention. However, first attempts to use Wikipedia are promising. The main challenge is to be able to parse the page and the textual content correctly. Here for example are the first sentences of the articles for a number of terms:

  - HEDGEHOG: “A hedgehog is any of the small spiny mammals of the subfamily Erinaceinae and the order Insectivora.”
  - CHAPERONE: “In biology, chaperones are proteins whose function is to assist other proteins in achieving proper folding.”
  - LORRY: “Lorry may mean: * A truck, a large commercial goods vehicle, * An open railroad car with a tipping trough, often found in mines, * A heavy horsedrawn Wagon”

  One advantage of encyclopaedias is that they have a relatively uniform style in which they begin entries and therefore may be easier to parse.

- Textbooks and manuals associated with the domain also have potential usefulness. Here the main problem is identifying the relevant texts and obtaining them electronically. Furthermore, in both this case and that of the encyclopaedias, there is the problem of data sparseness - we hypothesise that one will tend to find very few defining contexts. This remains to be empirically proven. Textbooks will naturally use every possible way of introducing new terms and defining them with respect to previously introduced knowledge. They will consequently be significantly more difficult to manage from a Natural Language Processing perspective.

- Google Glossary: this is a new experimental service from Google Labs which provides definition texts for the terms one enters. Consider the examples shown in Figure 18. The main current

---

3 Obtained from Wikipedia on 27 December 2005.
• (n) 1. a protein which makes possible or facilitates a chemical reaction under a given environmental condition. 2. a digestive enzyme, an enzyme secreted by the body which helps break down food substances into forms that can be absorbed and assimilated by the body. Digestive enzymes are secreted by the salivary glands (e.g., amylase or ptyalin which breaks down starches); by the stomach (e.g., pepsin which breaks down proteins); by the liver (e.g., bile which help break down fats by emulsifying them); and by the pancreas (e.g., amylase which breaks down starches and lipase which breaks down fats.)
  http://prism.troyst.edu/ tiemeyep/glossary.htm

• Enzymes of the proteinic molecules occur in various reactions. They are biocatalysts, i.e. proteins allowing to increase the speed of a chemical reaction, at temperature compatible with the biological life (37 °C). One of their properties is the specificity of action and reaction: each enzyme can be fixed only on one type of substrate (a molecule) and can catalyse only one chemical reaction. Once the catalysis is finished, the enzyme can enter in reaction again...
  http://library.thinkquest.org/26644/us/Lexique.htm

• Proteins that accelerate the rate of a specific chemical reaction without themselves being altered. Enzymes are generally named by adding the suffix “-ase” to the name of the substance on which the enzyme acts (for example, protease is an enzyme that acts on proteins).
  http://www.sahealthinfo.org/Modules/HIVAIDS/aidsglossary/ aidsglossary.htm

---

Figure 18. Definitions for Enzyme from the web.

problem with using Google Glossary as a source is that there are a great number of technical terms absent from it. Half of the 19 terms we considered above were absent from Google Glossary. Furthermore, we have no information as to how their look-up system works and it is unsatisfactory from a research perspective to use this type of black box.

• the Internet: this is the most obvious source and has both advantages and disadvantages:
  - Advantages a) It is extremely large so one is likely to find what is needed; b) It is continuously growing so recent conceptual developments are likely to be represented; c) It is easily accessed; d) We can understand what we can and cannot do with it.
  - Disadvantages a) For any given term, texts can occur defining it in many different domains. Thus when looking for the genetic definition of ‘chaperone’, on the Internet we find a number definitions (cf. examples in Figure 19). This is one form of noise; b) For narrow domains, even the internet does not cover the terminology - only the Deep Web\(^4\). c) The perspective of a particular corpus may not correspond to that of the web as a whole, thus providing another form of

\(^4\) The Deep Web is the areas of the Internet which are hidden behind a web form and can only be accessed manually i.e. the data is not indexed by the search engines.
8.4 Conclusions

This chapter has argued that there is an inherent contradiction in the desire to build ontologies for a domain from a specific set of documents. There are cognitive reasons why this should be the case. The ontology reflects the assumed background knowledge which the text is ‘maintaining’ i.e. re-enforcing and modifying. Furthermore, there is the practical reality that terms do not co-occur sufficiently

...a chaperone is a dangerous, freethinking individual in ...
Chaperone is a responsible female adult (minimum age of 21 years).
A chaperone is a helper protein that binds to a polypeptide ...

Figure 19. Definition from the web.
frequently so as to make possible the machine interpretation of the requisite knowledge. The beginnings of a solution have been proposed in this chapter by suggesting the use external sources such as the internet, and possibly more domain relevant sources, to compensate for the knowledge gaps in the initial corpus. We have argued that, in one limited experiment at least, we can go from no exemplars to at least 60% of cases having exemplars for the relevant knowledge relations.

There are various potential responses. One approach would be to take this as clear demarcation of the limits of what computational approaches can derive from text. Hays [1997] has argued, with respect to disambiguation, that there are limits as to what is linguistically encoded in a text and the rest is dependent on the world knowledge the reader brings. However, another perspective is that the limits of parsing and processing texts have yet to be reached so it is far too early to say what is and is not interpretable by machine.
The beginning of wisdom is the definition of terms.

Socrates

We have shown in Chapters 7 and 8 that there are important limitations on trying to build an ontology from a domain specific corpus. In Chapter 7, we showed that there were severe problems in finding enough instances of lexico-syntactic phrases in order to be able to acquire meaningful and precise knowledge. In Chapter 8, we argued that this fact was to a large extent inevitable given the nature of an ontology and the nature of texts - especially domain specific scientific or technical texts. These results appear to be rather disheartening, even if they provide both a greater understanding of the technical challenge and a stimulus to look for more fruitful approaches.

We are seeking to find techniques to manage the inherent properties of our material effectively, in this case human language. Many researchers in both language technology and symbolic AI either lament or ignore the inherently dynamic nature of language: the way it changes, slips, slides and transforms itself. As noted elsewhere, at least since the “universal language” movement of the 17th century KNOWLSON [1975], SLAUGHTER [1982] up to the work on Cyc and Mikrokosmos in the late 20th century (cf. Chapter 3), humans have sought to ‘fix’ language, to determine the precise meaning of words. In a certain metaphorical sense, we are still seeking to undo the effects of Babel. It is unusual among engineering scientists to be in such a denial about the nature of their materials. We do not hear of bridge builders complaining that they could build the perfect bridge if only the banks of the river could be made stable, or the shifting sands prevented from shifting. Architects do not complain of the insufficient stability of the ground on which they build. In fact, they have learnt to build under extreme conditions, even at the bottom the ocean or in outer space. Language engineers tend to blame the materials, which makes them almost as bad as the workman blaming his tools. It is in a spirit of finding engineering solutions which can take into account the inherent nature of the material, that the ideas of this chapter are presented.

In Chapter 6, we described the approach taken by the KnowItAll system(cf. Section 6.4.4), which uses the whole web as its source of information. Furthermore, it attempts to avoid ambiguity by using additional terms in its search queries so as to limit the retrieved documents to ones relevant to the domain. The Web is the largest ‘corpus of texts’ currently easily accessible, and there is an honourable tradition in NLP of using it both for research and knowledge acquisition purposes [AGIRRE et al., 2000, BANKO and BRILL, 2001]. Equally there has been the attitude in corpus-based linguistic research that size matters,
i.e. the more text available the better, sometimes leading to the view that the Web is necessarily better because it is so much larger than older corpora such as the BNC (for example). However, there are a number of problems here. First, the web changes continuously in size. The web of today is not the same as the web of tomorrow. Secondly, there is the problem of confidence or trust. What we find today on the web may or may not be true and just because we find it repeated does not make it any more true. Third, all previous research with corpora has found that Zipf’s Law hold i.e. that there is a long tail with many terms or words occurring very very infrequently and often only once [Zipf, 1949]. This fundamental data sparsity remains a major challenge.

In this and the following chapter we propose an approach to ontology learning which depends (in part) on a greater understanding of text type. In this chapter we focus on the importance of specific types of texts in order to acquire knowledge fruitfully, or at least to begin to understand how we might approach the problem more effectively. With respect to texts types, we propose three hypotheses, as follows (discussed further below):

1. Terms are likely to occur in explicitly defining contexts when the reader is being trained for the domain or being instructed in the subject area.

2. Terms are likely to occur in explicitly defining contexts when they move across a cognitive border from one domain to another.

3. Terms are likely to occur in explicitly defining contexts when they first are created, invented or used.

By an explicitly defining context we mean the contexts as defined in Chapter 7, which involve specific lexico-syntactic patterns. According to the above hypotheses, when a writer is writing for the trainee/learner, they are more likely to make the ontological knowledge explicit. Equally, when a writer borrows a term from another domain, they are likely to define the term for the new audience. Finally, the early uses of a term are in an environment where the writer feels obliged to provide an explanation or definition of a term. Essentially we are seeking textual events where the author cannot assume a concept/term is already part of the reader’s background knowledge.

These hypotheses presume a greater understanding of where and how human beings provide evidence for knowledge in their textual communicative acts. Only by understanding in greater detail what types of text actually contain knowledge can we then design systems to access the relevant texts. For example, a great deal of text in electronic format is not directly accessible on the web for various reasons. There has been considerable discussion of the Deep Web i.e. those parts of the Web which are not static but are output by a database as a result of a query [Bergman, 2001, Wright, 2004]. By knowing what

---

1 Cf. also Wikipedia entry on “Deep Web.”
9.1 LANGUAGE VARIETY: SYNCHRONIC AND DIACHRONIC VARIATION

It may appear obvious to researchers in NLP that language varies tremendously depending on different circumstances of creation and the intention of the author. In contrast, historically there has been a long tradition which viewed language as a homogenous whole, to be studied and analysed as one unified entity. This tradition certainly includes Saussure (1916) with his interest in *langue* as opposed to *parole*, continues via the American Structuralists and includes Chomsky’s interest in *competence* as opposed to *performance*. In all these cases, the model of language constructed is idealised in that what is described is an abstract form of the language with all the variation between individuals, topics, styles and genres disregarded. This variation is considered irrelevant to the fundamental structure of the language. It was assumed that all individuals speaking a single language must conform to a single code, a single vocabulary, in order to enable communication between members of that language community. Moreover, Chomsky states “Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interests, and errors (random or characteristic) in applying his knowledge of the language in actual performance” [Chomsky, 1965].

There is, however, another tradition, dating back at least to the Neo-Grammarians, is concerned with language change and dialect variation. This tradition has recognised the variability and diversity of language, celebrating the difference between groups, sub-groups and even individuals of the same speech community. It has come to be recognised that people use language in different ways on different occasions and for different purposes. A combination of philosophers
[e.g. Wittgenstein, 1953], linguistic theorists [e.g. Halliday, 1976] sociolinguists [e.g Labov, 1972], field linguists [e.g Bickerton, 1973] and stylistic researchers resulted in a number of different terms to describe these various types of language. Thus the term register was proposed to differentiate between degrees of formality in language or types of language varying by communicative situation. Topic, retaining its common sense meaning, refers to the subject matter or ‘aboutness’ of a text (however, see Sinclair and Ball [1996] on the importance of ‘internal criteria.’). The dimension that we believe relevant to our needs here is that of genre, which we will consider further below.

The other great concern of the Neo-Grammarians was diachronic variation i.e. how language changes. From our perspective of finding the most appropriate environments for extracting knowledge what is significant is that certain words are created and enter the language and on these occasions certain usage and patterns arise. This we view as the diachronic phenomena which we also consider further below.

9.2 THE SYNCHRONIC DIMENSION

9.2.1 Genre or language for a purpose

Genre has been defined differently by various authors and often confusingly interchanged with register. Lee [2001] has provided a useful survey of various definitions and proposes his own definition whereby genre and register are essentially different perspectives covering the same ground. “Register is used when we view a text as language: as the instantiation of a conventionalised, functional configuration of language tied to certain broad societal situations, that is, variety according to use. ... Genre is used when we view the text as a member of a category: a culturally recognised artifact, a grouping of texts according to some conventionally recognised criteria, a grouping according to purposive goals, culturally defined” [Lee, 2001, :46]. Another way of putting this is that genre refers to what task a text is achieving in a given cultural context. The cultural context is important because, as we noted previously (Chap. 8). In spite of many scientists’ belief to the contrary, we have argued that a fundamental purpose behind an ordinary scientific article (for example) is not to instruct or define the shared ontology, but rather merely to perform an act of knowledge maintenance (and take the instruction, the definition of terms, the shared ontology - all for granted). For our purposes, we are interested in those types of text, those genres, which are especially rich in explicit statements about knowledge because for one reason or another they do not take the background ontology for granted.

The practical point is that varieties of register concern such categories as formal or legal. Genre concerns such categories reports, technical manuals, recipes, wills and advertisements. Our expectation is that certain genres (e.g. encyclopaedias, manuals, textbooks, etc.) will be especially fruitful for the extraction of knowledge.
9.2 THE SYNCHRONIC DIMENSION

| A. PRESS: REPORTAGE |
| B. PRESS: EDITORIAL |
| C. PRESS: REVIEWS |
| D. RELIGION |
| E. SKILL AND HOBBIES |
| F. POPULAR LORE |
| G. BELLES-LETTRES |
| H. MISCELLANEOUS: GOVERNMENT & HOUSE ORGANS |
| J. LEARNED |
| K. FICTION: GENERAL |
| L. FICTION: MYSTERY |
| M. FICTION: SCIENCE |
| N. FICTION: ADVENTURE |
| P. FICTION: ROMANCE |
| R. HUMOR |

Table 24. List of text categories used in the Brown and LOB corpora

9.2.2 Categories of Text

The extensive work done to build text corpora over the last half century might have been expected to result in a coherent set of categories for the types of texts included in these collections. There might even have been a hope for some sort of standardised schema that would prove useful for our purposes. Most public corpora have had extensive categorisation schemes. The earliest electronic corpus was the Brown Corpus of American English consisting of texts printed in 1961 [Francis and Kucera, 1964/1979]. This consisted of 500 texts of about 2000 words each, distributed over 15 categories (shown in Table 24). They are clearly rather heterogeneous classes including topics such as ‘religion’, mediums such as ‘press’ and literary categories such as ‘belles-lettres’.

Later work disentangled the different dimensions along which variations of language could be categorised, defining such dimensions as mode (e.g. written vs. spoken), medium (e.g. book vs. periodical vs. radio vs. telephone), genre (e.g. novel vs. play vs. article vs. horoscope) [Atkins et al., 1992]. Thus we see a progression from the idea of a relatively small number of categories to a multitude of what in librarianship terms would be called ‘facets’ or dimensions. With the creation of the British National Corpus [Aston and Burnard, 1998], a real effort was made to design a ‘balanced’ corpus representative of the whole language and this necessitated the use of such a faceted approach. Thus in the BNC written texts are classified by medium, domain, audience, author, and other categories including place of publication (cf. Figure 20).

The BNC, however, avoids proper genre categories and to compen-
sate for this lack Lee has provided a “BNC Index”\(^2\) which classifies the BNC texts by a number of dimensions including genre. To this end, he provides a list of 70 genres or sub-genres (24 for spoken language, and 46 for written texts). The genre categories he proposes include such labels as \(W_{ac\_humanities\_arts}\) (i.e. written, academic prose, humanities) or \(W_{non\_ac\_humanities\_arts}\) (i.e. written, non-academic/non-fiction, humanities). Although, 70 categories may seem a considerable number, the impression is that they are quite arbitrary, and, like taxonomies, just reflecting the cultural assumptions of that moment.

This is in fact the position which he takes in a well-argued paper on the concept of genre and its application to the BNC [Lee, 2001]. He notes that Biber defines genre in terms of use and topic “Genre categories are determined on the basis of external criteria relating to the speaker’s purpose and topic; they are assigned on the basis of use rather than on the basis of form” [Biber, 1988, :70] but that ‘topic,’ according to the EAGLES report, is concerned with “the lexical aspect of internal analysis of a text . . . [Topics] are subjective, and . . . the resulting typology is only one view of language, among many with equal claims to be the basis of a typology” [Sinclair and Ball, 1996, :17]. He contrasts register and genre as two different ways of categorising the same ‘ground’ where register relates to societal situation and is variety according to use, and genre refers to a ‘culturally recognised artifact’ and is variety according to purpose. He sees genres categories as categories which arise from consensus in a culture and thus are subject to gradual but continuous change [Lee, 2001, :46].

Genre may be a culturally defined range of categories as Lee argues, but an abiding and in all likelihood universal genre category is that concerned with instructional and educational material. While we have not tested the hypothesis in a cross-cultural context, we suspect that at all times and in all cultures there are texts which function to instruct the next generation about the concepts ad terms used by the previous generation and defines those concepts and terms in one manner or another. Thus in Lee’s (self-admittedly arbitrary) categorisation scheme for the BNC, the only clearly useful category for our purposes is that

\(^2\) http://clix.to/davidlee00

Figure 20. Categories used by the BNC (genre is not clearly defined).
of W_instructional (i.e. written, instructional texts/DIY) although the non-academic scientific categories could also be a priori relevant to our purpose of learning ontologies. More generally, our first hypothesis would indicate the need to focus on manuals, textbooks, encyclopaedias but possibly other categories of text which would need to be identified empirically.

The motivation for trying to identify the type of text which is especially rich in explicit knowledge lies in the expectation that there are in fact a much wider range of explicit lexico-syntactic contexts from which knowledge can be extracted than those described in Chapter 7 above. As we have argued previously, ordinary scientific texts most of the time are largely ‘maintaining’ knowledge - they rarely revolutionise the scientific paradigm in the sense of Thomas Kuhn. However, in certain types of texts, the reader brings significantly less background knowledge and it is in these texts that explicit contexts can be identified. The Adaptiva experience [Brewster et al., 2002] showed that it was not easy for human annotators to identify ontologically explicit contexts, and there are major issues with data sparsity as described in Chapter 7. Consequently the identification of types of texts which a priori are known to be rich in knowledge bearing expressions is extremely helpful - effectively it is a way of reducing the search space, something which is important given the relative data sparsity.

It could be argued that all one needs to do is find texts with a large number of existing Hearst patterns. This might have some merit but ignores the fact that knowledge is presented in a variety of ways to the reader, and any assumptions that may be made initially have no guarantee of being borne out. In order to identify the variety of means which writers use to convey knowledge, it would be useful first to have the relevant corpus and then to be in a position to apply adaptive information extraction techniques to this corpus (see Brewster et al. 2002 and Chapter 10) and thereby learn a greater range of lexico-syntactic patterns that those we currently have at our disposal.

It is in the context of the need to identify automatically the appropriate type of text for the mining of the knowledge and the learning of new patterns that we turn to methods to categorise texts. Text classification or categorisation has a long history in NLP [Guthrie et al., 1994, Sebastiani, 2006] [Manning and Schütze, 1999, Ch.16] with a considerable success rate. Traditionally text classification concerned classification by topic (ever since the early work on ‘automatic indexing’ Maron [1961]) but has also extended to other areas such as spam detection, authorship attribution and sentiment detection. More recently there has been a flurry of interest in the possibility of automatically classifying text genres [Stamatatos et al., 2000a]. The usual motivation presented is that users of the Web may wish to restrict a search not only by topic but also by text genre [Boese and Howe, 2005, Finn and Kushmerick, 2006]. The use of only topic clearly does not restrict searches sufficiently and users often have specific types of text.

---

3 No text assumes the total absence of background knowledge because then communication would be impossible.
in mind when they are searching.

The interesting thing is that some people in this community define genre in a very different way from that of corpus linguists or linguists in general. Here are some example definitions:

- “The genre of a document concerns aspects of the document such as the style or readability, presentation layout, and metadata such as words in the title or the existence of graphs or photos” [Boese and Howe, 2005]

- “Genre is an abstraction based on a natural grouping of documents written in a similar style and is orthogonal to topic” [Finn and Kushmerick, 2006]

Although in the case of Finn and Kushmerick this results in unusual genres (fact vs. subjective, positive vs. negative), Boese and Howe use much more recognisable categories such as Abstract, FAQ, How-to, C.V./Resume, Tech Paper, etc. and Lee and Myaeng [2002] used categories such as reportage, editorial, research article, homepages etc. Boese and Howe note that the lack of a standardised set of genre categories and a corresponding standard corpus means that each researcher establishes their own text collection. This clearly means results tend not to be comparable.

There is also a considerable divergence as to the textual features that should be used in constructing classifiers. Lee and Myaeng [2002] construct a genre classifier based on identifying terms which are characteristic of that genre but occur across many subject classes of the genre. They give the example of the term home for the home page genre. Stamatatos et al. [2000b] use the most common words in the language as features i.e. what are usually removed as stop words. Finn and Kushmerick use three different feature sets: bag of words, POS statistics and ‘shallow text statistics’. For a review of various approaches to genre classification consult Boese [2005].

The fundamental point however, from our perspective, is that it is possible to construct a genre classifier with a high degree of accuracy, often reaching over 90% in the papers cited here. Thus we can propose that it should be methodologically coherent to identify those genres which are most didactic and instructive or more generally do not assume too much from the reader and identify these text for further analysis for knowledge extraction.

9.3 THE DIACHRONIC DIMENSION

Language has often been describe as a living entity continuously subject to change, continuously developing and metamorphosing. An essential part of this process is the introduction of new ideas and their associated terms into the language. Thus a word or term may enter a language at a certain moment, its currency may wax and wane, and in some cases may eventually disappear from current usage. Of course, since the creation of the printing press, words used in a
printed text can no longer entirely disappear from the language. Before
that only the extraordinary memory of the epic traditions retained
archaic forms long after they had fallen into disuse in the everyday
language [Rubin, 1995]. In the modern era, words can die out and
sometimes be revived but the greatest impact both in the creation and
maintenance of words is the progress of technology. A great many
new technological inventions or scientific discoveries have demanded
new terms to describe them, whether they are the gluon of physics or
the blog of the Internet. Equally the rise and fall of a technology may
lead to the rise and fall of the corresponding term, as for example in
the case of telex a term many people under the age of thirty will have
difficulty recognising.

There are two types of diachronic events which we will consider.
The first, reflecting out second hypothesis concerns the borrowing of
a term form one domain to another. The second concerns the creation
of new terminology, in effect the birth of a new term.

9.3.1 The Borrowing or Migration of Terms

In contrast to our previous discussion of textual genres, here we will
concern ourselves with language reflecting certain topics or sublan-
guages as they have been termed. Academic disciplines, even in this
age of multidisciplinary research, have relatively clear boundaries
reflecting such institutional structures as academic departments and
social dimensions as specific social networks [Small and Griffith, 1974,
Hopcroft et al., 2004]. Such disciplines have specific sublanguages which
reflect particular terms and even grammatical structures characteristic
of the field. Considerable research was undertaken into sublanguages
even in the 1980s and early 1990s [Kittredge, 1982, Grishman and
Kittredge, 1986, Haas and He, 1993, Haas, 1997]. Some of the early
work on sublanguages tends to use the term to refer to what we would
now term genres (cf. above) for example ‘reports’ or ‘patient histories.’
Our concern here is with sublanguages which reflect the terminology
(in particular) of professional and academic disciplines.

Within a discipline, as it changes and progresses, new theories or
technical advances are made together with corresponding terminology
to describe the relevant concepts. If the term and is associated ideas
are successful or useful the frequency of its use will increase within
the discipline (intra-disciplinary growth) and if the ideas are general
enough or found to be useful elsewhere then the term may migrate,
being borrowed by one or more disciplines (inter-disciplinary growth)
[Losee, 1995].

Our second hypothesis is that in such an environment, that is where
a term is new to a domain, the writer cannot assume that it is part of
the background knowledge of the reader and so is likely to provide
a definition of explanation which may provide an opportunity for
knowledge capture. The explanation may be provided either because
the term or concept is new to that field or because the author wishes
to change some part of the existing definition or understanding so as
to adapt it to their theoretical or practical needs in the context.

Examples of such expression include the following:

Example 9.1  Ginsberg and Harvey [1992]
So we first start by searching the tree with an artificial breadth cut-off of 2, then try with a breadth cut-off of 3, and so on, until an answer is found. We will refer to this technique as “iterative broadening,” borrowing terminology from Korf’s notion of iterative deepening.

Example 9.2  Sjostrom et al. [2001]
Recently, much interest in immunology has focused on posttranscriptional membrane-related events occurring when immune effector cells meet their target cells. Borrowing terminology from the neurosciences, the interphase between effector and target has been termed the immunological synapse.

Example 9.3  Allen et al. [2000]
Our paper demonstrates that this conclusion is an artifact of a modelling strategy which allows capacity to be instantaneously adjusted upward in the post-entry game. Borrowing terminology from the real business cycle literature (Kydland and Prescott (1982)), we call this assumption no time to build.

Arabidopsis is one of several multicellular model organisms (along with Drosophila, Caenorhabditis and zebrafish).

The phrase ‘borrowing terminology’ is quite fruitful in identifying such examples but it is obviously not the only environment in which terminology is actually borrowed and defined. However, the sorts of textual contexts found are not obviously explicit in our sense.

Losee has suggested that there are ‘hard sciences’ (mathematics, physics, etc.) and ‘soft sciences’ (education, psychology, economic, etc.) and that in general terms tend to migrate from the hard sciences to the soft sciences rather than vice versa. He seems to think that terms migrate relatively rarely and but there are no quantitative studies as to the amount of terms borrowed from one discipline to another.

One potential challenge is defining the boundaries between disciplines in order to be able to determine the occurrence of a word migration. At an abstract level we could conceive of a discipline as probabilistic model where terms have a certain probability of occurring. Core terms have high probability, peripheral terms have a low probability, and terms which are borrowed are ones with a high probability in another field, especially at an earlier time period. Previous work on the terminological differentiation of disciplines and sublanguages has shown that automatic categorisation by sublanguage can be quite successful. Essentially it is no different that text classification (cf. above) but once a text has been classified then the terms in that document can be evaluated for their ‘fit’ with the domain. Terms which do not fit are potential immigrants. Terms with major dominance either by frequency of use or intellectual significance are potential emigrants.
9.3.2 New Terminology

On first use, or until the reader can be assumed to understand, our third hypothesis predicts that the writer will take the trouble to explain in some manner or another i.e. instruct the reader to add the term to their background ontology. Indicative evidence for the validity of this proposal can be found by considering the first citations used in the OED. Citations are not chosen for their perspicuity or clarity in defining a term but rather they reflect the earliest occurrence of a term in text: “The usage of each word, meaning, or idiom in the Dictionary is documented through comprehensive examples drawn from quotations taken from printed texts of the present and the past. These quotation paragraphs begin with the earliest recorded occurrence of a term.” Consequently for unambiguous technical terms, the evidence provided by the ‘first citations’ is helpful.

Some examples from the OED are the following: The first citation for baboon is:

Example 9.5 c1400 Mandeville xii. 238
Babewynes, Apes, Marmesettes, and othere dyverse bestes.

Thus it is that we can apply a Hearst pattern on 15th century text. We can say that the following is an ‘explicit context’ for the definition of chromiole:

Example 9.6 1899 G. Eisen in Biol. Centraalblatt XIX. 131
The first-mentioned granules are the chromatin elements which in time will form the chromosomes. For these granules ... I have proposed the name chromioles.

However, in the following example, which is the first occurrence of gene, although the author is clearly providing a definition, it cannot be described as an ‘explicit.’

Example 9.7 1911 W. Johannsen in Amer. Naturalist XLV. 132, I have proposed the terms gene and genotype...to be used in the science of genetics. The gene is nothing but a very applicable little word, easily combined with others, and hence it may be useful as an expression for the unit-factors, elements or allelemorphs in the gametes, demonstrated by modern Mendelian researches.

Further indicative examples can be found considering the first ten uses (for example) of various recently coined terms on the archives of web sites such as BBC News or the Guardian. Here we present some of the first occurrences of blog on the BBC News website and in the Guardian:

Example 9.8 BBC News news.bbc.co.uk 17/06/2002
Weblogs, or blogs, are online journals where cyber-diarists let the world in on the latest twists and turns of their love, work and internal lives.

http://dictionary.oed.com
Example 9.9 Guardian 31 January 2002

Blogger, which he and a small team launched in August 1999, did not invent the web log or “blog” - a frequently updated, diary-like commentary usually filled with links to other websites.

Inspection of such examples in greater numbers show that there is a strong tendency for explicit expression of the ontological knowledge i.e. for defining statements. The examples chosen both explain what a blog is and this type represented about a third of the first ten examples. While 33% of cases may not seem a lot, there is no real methodological need for a large quantity of explicit expressions of ontological knowledge. All that is needed is that there exist some cases which are identifiable.

The main problem with expecting the diachronic first use to reflect a defining context is that words usually have an extensive oral life which usually predates its use in writing considerably. The extent to which this will in fact be an impediment to using this method depends on the nature of the term whose defining context we seek. If it is a very specialised term, with narrow usage, it is more likely to occur in an explicit context. If it is a more general word given a specific meaning, this is less likely to occur, and may be a greater challenge. More generally, the extension of an existing term providing it with another sense either by metaphorical extension or merely because it appears to be appropriate is more difficult to identify and to process automatically.

9.4 CONCLUSIONS

In this chapter we have proposed a way to circumvent to some extent the problems of data sparsity which plague all attempts at knowledge acquisition from texts. Our proposal is that by identifying the specific types of texts or textual events a greater richness of material will be identifiable. Specifically texts which by their nature make fewer assumptions about the reader are likely to contain more explicit ontological phrases. Furthermore, the initial occurrences of terms whether because they are freshly coined or borrowed from a neighbouring discipline will often co-occur with explanatory phrases.

In general by understanding more effectively how authors construct texts and what they assume from the reader it will be possible to find those relatively rare occasions when the author is explicit and explains to the reader what they mean. In the knowledge acquisition game of cat and mouse, we believe we have proposed some effective mouse traps for knowledge.

An interesting example is the history of the term entropy in English. This was a term first proposed in German (Entropie) by Clausius in the context of thermodynamics and first used in English in an opposite sense to the German. Then it reverted to the originally used German sense, but was provided with a different definition in terms of statistical disorder by Boltzmann. The term was further extended by Shannon in information theory, although this sense was argued to be a generalisation of the statistical on (cf. Wikipedia one entropy, and OED entry)
10

A MODEL OF KNOWLEDGE ACQUISITION

He affirms that there is a supreme Deity, by name Abraxas, by whom was created Mind, which in Greek he calls Nous; that thence sprang the Word; that of Him issued Providence, Virtue, and Wisdom;

Tertullian – Against All Heresies [220/1867]

10.1 THE ABRAXAS APPROACH

The purpose of this thesis has been to investigate the relationship between knowledge and its expression in text so as to understand how ontologies can be built from collections of texts. This has necessitated an investigation of knowledge and its nature, range and structure. Subsequently, we have considered a number of methods and techniques used in NLP which can extract varying types of knowledge from texts and which could be used to contribute to an ontology learning system. In connection with these techniques, we have also reviewed and analysed specific systems which claim to build ontologies. A number of conclusions emerge from the relatively broad sweep of this thesis:

- Knowledge is not monolithic, monotonic or universally agreed. It is uncertain, revisable, contradictory and differs from person to person.
- Knowledge changes continuously over time and so what is believed to be true today will be revised and re-interpreted tomorrow. Changes can be more or less rapid due to social and technological factors.
- Only a small part of what is generally considered knowledge can be represented in an ontology.
- Ontologies are useful if incomplete models of domains.
- Texts assume the reader has a certain amount of background knowledge. The great majority of what we would expect to be in an ontology is exactly in this assumed background knowledge.
- While it is easy to establish that some relationship exists between two terms, explicit defining contexts are relatively rare in texts due to data sparsity.
- From the perspective of ontology learning, data sparsity results in the need to use multiple sources of information concerning the ontological relationship between two terms, and furthermore the recognition that in certain cases textual evidence is too weak to say anything more than that there is some relationship.

1 Many of the ideas presented in this chapter arose out of extensive discussions both with Yorick Wilks and José Iria.
In view of this understanding of knowledge, texts and ontology learning, we propose in this chapter a uncertainty-based, dynamic, data-driven account of ontology learning. In this chapter we present an approach to ontology learning which, because it is far closer to the psychological and practical reality of human knowledge processing, will be more effective, more successful in allowing ontologies to be built, maintained and managed.

The standard approach to ontology learning views it essentially as a pipeline with a set of domain specific texts as input and a set of ontological triples as output. This may be augmented by accessing further resources such as WordNet or Google frequency counts [Cimiano et al., 2005]. However, this remains an essentially linear process and as such reflects an underlying philosophical perspective that knowledge is a monolithic structure [Brewster and O’Hara, 2007]. In contrast, we view the knowledge acquisition process and Ontology Learning as iterative and cyclical. We term this the Abraxas approach because that is the name given to the EPSRC research project which has partially funded this work (http://nlp.shef.ac.uk/abraxas/).

10.2 A TRIPARTITE MODEL

Our incremental approach is founded on viewing ontology learning as a process involving three resources: the corpus of texts, the set of learning patterns (of the form discussed in Chapter 7), and the ontology (conceived as a set of RDF triples). Each may be seen in an abstract sense as a set of entities with specific structural relations. The corpus is composed of texts which may (or may not) have a structure or set of characteristics reflecting the relationship between them e.g. they may all come from one organisation, or one subject domain. The learning patterns are conceived as a set of lexico-syntactic textual patterns or more abstractly as a set of functions between textual phenomena and an ontological relationship of greater or lesser specificity. The ontology is also a set of knowledge triples (term - relation - term, or rather domain - predicate - range) whose structure may grow more and more complex as more items of knowledge are collected.

The goal is to create or extend existing resources in terms of one another, with optional and minimal supervision by the user. Any of the three will interact with any of the other two to adapt so that all three remain in equilibrium (Figure 21). The methodology allows, for instance, creating an ontology given an input corpus, extending a corpus given an input ontology or deriving a set of extraction patterns given an input ontology and an input corpus.

The initial input to the process, whether ontology, corpus, patterns or combinations of them, serves both as a specification of the domain of interest and as seed data for a bootstrapping cycle where, at each iteration, a decision is made on which new candidate concept, relation, pattern or document to add to the domain. Such a decision is modelled via three unsupervised classification tasks that capture
the interdependence between the resources: one classifies the suitability of a pattern to extract ontological concepts and relations in the documents; another classifies the suitability of ontological concepts and relations to generate patterns from the documents; and another classifies the suitability of a document to give support to patterns and ontological concepts. The notion of “suitability” or rather Resource Confidence (RC) is formalised within a framework modelling degrees of certainty, one in which the relationship of any resource to the domain is assigned a confidence value (cf. below 10.4.1). Thus, ontology, corpus and patterns grow from the maximally certain core (the initial or manual input) to the least certain fringes as the process iterates. This is conceived as a set of concentric circles of diminishing Resource Confidence (cf. Figure 22), where the position of any item is continuously changeable depending on the addition of newer items. Newly added elements have an initial neutral certainty but as the data they provide is further confirmed confidence in them increases (or decreases) accordingly. This means that subsequently added elements over time can acquire a confidence value equivalent to the manual input. Stopping criteria are established by setting a threshold on the lowest acceptable RC for each resource type, or by setting a threshold on the maximum number of iterations, without any new candidate resources for each resource type being obtained. The incompleteness of the corpus is tackled by iterative augmentation using the web or any other institutional repository as a corpus. Corpus augmentation in our approach consists of a set of methods that aim to incrementally add new documents to the corpus, such that documents with higher relevance to the domain are added first. We present further details about certain aspects of this model in the remaining sections of this chapter.
10.2.1 The Corpus

As we have argued elsewhere, the corpus is the fundamental source of knowledge in text-based ontology learning. It provides an entry into the mind of the authors of texts, and by collecting the documents of a institutional community or subject domain, provides significant insight into the common vocabulary and background knowledge of those communities or domains.

We conceive of the corpus as an open-ended set of documents. This document set may be empty initially, partially filled or filled with a considerable set of documents. In each case, this has different implications for our approach:

- **Empty Set** In such a case, the domain of the ontology must be specified by a seed ontology of greater or lesser size. The greater the seed ontology the more rigorously defined the domain. Documents will be added to this empty set in response to retrieval queries generated from the ontology terms.

- **Partially populated set - the Seed Corpus.** In this case, the documents are seen as defining the domain, either alone or in tandem with the ontology. The domain is defined by the terms or key words in the corpus.

- **Full Corpus.** In principle, the ABRAXAS model assumes that all corpora are open-ended but in some cases the model needs to work with a corpus that is defined as all there is i.e. no further documents are available. This will have an impact on the degree of detail that an ontology derived from the corpus will have.

In the first two the expansion and growth of the corpus is achieved by iteratively adding further documents which fulfil certain criteria, typically of providing explicit ontological information (as defined in Chapter 7). The Seed Corpus can have a further function in playing a role in the stopping criterion. Thus if the ontology grows to the stage that it covers the Seed Corpus adequately, then this can be identified and halt the process (cf. Section 10.4)

Documents can be added to the corpus from a variety of sources. The web is an obvious source and has been widely used in ontology learning work. An institutional or organisational repository may be another source. Furthermore, there may be specific external sources which may be defined to be used, in the spirit of the genre and text type discussion of Chapter 9. In such a case, only texts of a certain genre or chronology would be made available to the system. A key design issue is how documents are selected and queued to enter the corpus. Should a new document be needed, a query is constructed for any one of the possible sources of new documents. The most typical situation in which an additional document needs to be added is when there is an absence of explicit knowledge defining the ontological relationship between two terms. A document query is constructed containing the relevant extraction pattern and query terms, and documents which contain these are identified. These are queued for possible addition to
Each document has an associated Resource Confidence value, range $[0,1]$ (explained further in Section 10.4.1). A document existing in the seed or full corpus will have a confidence value usually of $[1.0]$, although the user may choose to set another value. A document added subsequently will begin by having a confidence value just above neutral $[0.5]$ because it will have fulfilled some criteria for inclusion. If it proves to have a set of terms very close to the seed corpus, confidence increases and equally if a number of extraction patterns are applicable again confidence in this document increases.

10.2.2 The Ontology

As noted above, the ontology, much like the corpus, consists of a set of objects, in this case knowledge triples (term, relation, term or in RDF terminology domain, predicate, range). It is also an open-ended set which initially may be more or less fully specified. Thus like the corpus, the ontology set may be an empty, partially filled or fully filled set of triples.

- **Empty Set.** The ontology in this case provides no guidance to the ontology building process. Its size and scope will be entirely determined by the other two entities in the system i.e. the corpus and the extraction patterns.
- **Partially Filled Set.** This situation arises typically where the ontology is acting as a ‘Seed Ontology’ which will determine the domain for the ontology building process, or where an ontology exists already and the system is being used to update or expand the ontology.
- **Fully Filled Set.** This situation arises only if the user believes the ontology fully represents the domain and wishes either to ex-
pand the document set (using the ontology as a complex source of queries over the domain) or as a bootstrapping knowledge source for the extension of the extraction pattern set.

In the first two cases, knowledge triples are added iteratively as they are obtained by applying the extraction patterns to the corpus. A Seed Ontology can act as input to the stopping criteria if for example the system is required to identify textual evidence for confirmation of the existing knowledge in the ontology and asked to stop once that has been achieved i.e. once the corpus covers the ontology. The partial and fully filled ontology sets can be used to learn additional extraction patterns from the corpus (existing or obtained).

Knowledge triples (for example hedgehog ISA mammal) are added to the ontology set as a result of either manual insertion or as an output of the application of the extraction patterns to the corpus of documents. Manually inserted triples have a confidence value of 1.0 while automatically added triples have a confidence value which is a function of the document from which they are derived and the reliability of the extraction pattern. Over time as additional documents are added to the corpus, the confidence values for any given triple may increase and be adjusted accordingly.

There are three important points in which our conception of the ontology set differs from an ontology as usually envisaged. First, a set of triples does not add up necessarily to an ontology as usually conceived. We envisage a post-processing stage which rationalises the multiple triples and constructs a hierarchy from the multiplicity of triples. This in itself is not a trivial task but it also reflects the cognitive reality that we may have disparate items of information, individual facts, which do not necessarily add up to a coherent whole. Only a sub-part of the ontology set may actually be rationalisable into a normalised ontology. Secondly, confidence values are associated with each triple and these change continuously as further data is processed by the system. These confidence values start a neutral level for every possible term-relation-term triple (0.5) and alter in the light of further evidence both positive and negative. As more and more evidence for a knowledge triple is encountered so the confidence value increases. As evidence mounts that any two terms are never encountered together, so the confidence value decreases for a given candidate triple. Thirdly, we also envisage a neutral ontological relation so as to be able to state the fact that terms $T_x$ and $T_y$ are closely related by some $R$ but the type of that relation is unknown. This allows the ontology to reflect relational information based on term distribution (of the type identified by term association algorithms and sometimes represented by PathFinder Networks) while still capturing the absence of knowledge concerning the specific nature of the ontological relationship.
10.2.3 The Extraction Patterns

We have discussed extraction patterns in greater detail previously in Chapter 7. As in the other two entities in the Abraxas approach, the Extraction Pattern set is an open-ended set. Clearly it is more limited in size and scope than the other two sets and this is due to the relatively limited number of explicit lexico-syntactic patterns which exist in natural language. Nonetheless, new patterns do exist and develop and there are frequently domain specific ones either related to a specific subject area (e.g. legal documents) or to specific types of texts (glossaries or dictionaries for example).

The extraction pattern set, like the other two entities, may be empty, partially-filled or fully filled:

- **Empty Set.** This situation is essentially counter-intuitive in that it means that the system has no predefined means for recognising ontological relations and is entirely dependent on the information provided by the corpus and the ontology. We believe that in principle all extraction patterns should be learnable by the system from the combination of ontology (which provides examples of terms related by labelled relations) and the corpus (which provides examples of instantiations of the knowledge triples in texts). However, this is not any easy process.

- **Partially-filled Set.** This is the typical starting situation. As described in Chapter 7, we assume a default extraction pattern essentially identifying the existence of an ontological association between two terms, and a number of sets of lexico-syntactic extraction patterns. Each set corresponds to a specific ontological relation. Furthermore, each set may expand by identifying further instantiations in text of two terms.

- **Fully specified Set.** In this case all extraction patterns the system is allowed to use are predefined. This may arise if the user wishes to obtain very high precision or only extract a specific type of ontological relation.

It should be clear that we consider the first and third case described here as being exceptional. The standard situation involves a predefined initial set of extraction patterns which are expected to grow over time with exposure to further data. The set of extraction patterns act as stopping criteria because if nothing new can be extracted from the corpus then either a) a new extraction pattern must be added, or b) a new document must be added. If neither of these is justified (e.g. due to the absence of explicit knowledge) then the system can halt.

New extraction patterns are added when a certain ontological relation is instantiated repeatedly in a specific lexico-syntactic environment. This presupposes that the ontological relation has already been made available to the system and example knowledge triples exist. The following steps then occur: a) Given a knowledge triple of the form \( \{T_x, R, T_y\}\), a number of sentential contexts (citations in corpus linguistic terminology) are identified; b) the citations are analysed.
and abstracted over in a manner similar to that used in the Adaptive IE algorithm $LP^2$ (cf. Section 7.4.1); c) the abstracted lexico-syntactic pattern is added to the Extraction Pattern Set with a given confidence value dependent on how many existing (known) knowledge triples it actually extracts.

In conclusion we may state there are potentially 27 states (3 times 3 times 3) in which the Abraxas system could be in. These are shown in Table 25. A number of these are of no interest practically. For example state 1 (all resources empty) and state 27 (all resources fully specified) have no in-equilibrium so no expansion of resources is triggered. Some intermediate states where both corpus and the ontology is empty are very underspecified and so would result in a random ontology. In Chapter 11, we will evaluate two of these starting states.

10.2.4 The Fourth Dimension: The System User

There is a fourth aspect to these three resource sets and that is the system user. Of necessity, the system user must provide some initial seed resources, whether in the form of a corpus, a set of extraction patterns, or a seed ontology. As noted above, there is no need for the user to provide more than just some basic seed data because the system is designed to re-assert its equilibrium by seeking compensatory resources from elsewhere. The consequence of this is that the user must also specify what external resources the system has access to. These may include the whole web, some particular subpart (for example a range of IP addresses), a specifically defined repository, a corporate intranet etc.

When the user defines manually some specific resources, they are designated as having a specific degree of confidence (RC). There are two options, as the model currently stands. In one case the user specifies the resource as maximally belonging to the domain (RC = [1.0]). In the other case, the resource is defined as not belonging to the domain (RC = [0.0]). This allows the user to provide both positive and negative input for the start of the process, and also in the course of the running of the system it will be possible to designate certain resources as [1.0] or [0.0] and thereby ‘guide’ the system as it iteratively progresses in its process of knowledge acquisition.
<table>
<thead>
<tr>
<th>State No.</th>
<th>Extraction Patterns</th>
<th>Ontology</th>
<th>Corpus</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Empty</td>
<td>Empty</td>
<td>Empty</td>
<td>Stable - but nothing to do</td>
</tr>
<tr>
<td>2</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td>Data in only one resource</td>
</tr>
<tr>
<td>3</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td>Data in only one resource</td>
</tr>
<tr>
<td>4</td>
<td>&quot;</td>
<td>Partial</td>
<td>Empty</td>
<td>Data in only one resource</td>
</tr>
<tr>
<td>5</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>&quot;</td>
<td>Fully specified</td>
<td>Empty</td>
<td>Data in only one resource</td>
</tr>
<tr>
<td>8</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Partial</td>
<td>Empty</td>
<td>Empty</td>
<td>Random ontology - not domain specific</td>
</tr>
<tr>
<td>11</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>&quot;</td>
<td>Partial</td>
<td>Empty</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>&quot;</td>
<td>Fully specified</td>
<td>Empty</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Fully specific</td>
<td>Empty</td>
<td>Empty</td>
<td>Random ontology - not domain specific</td>
</tr>
<tr>
<td>20</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>&quot;</td>
<td>Partial</td>
<td>Empty</td>
<td>Abraxas Evaluation - cf Chap 11</td>
</tr>
<tr>
<td>23</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td>Abraxas Evaluation - cf. Chap 11</td>
</tr>
<tr>
<td>24</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>&quot;</td>
<td>Fully specified</td>
<td>Empty</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Partial</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>&quot;</td>
<td>&quot;</td>
<td>Fully specified</td>
<td>Stable - but again nothing to do</td>
</tr>
</tbody>
</table>

Table 25. The 27 different starting states the Abraxas system could be in.
10.3 THE BASIC METHOD

The basic or core method in Abraxas involves an iteration over the three resources mentioned above triggered by some measure of inequilibrium. The starting state necessarily involves some initial manual input to the resources (cf. Table 25). The system will then test for “equilibrium” as mentioned above, i.e. whether the corpus provides evidence for the knowledge triples given the extraction patterns, whether the extraction patterns provide evidence for the knowledge triples given the corpus, whether the knowledge triples provide evidence for the corpus given the extraction patterns. At an abstract level we can visualise this process as shown in Figure 23.

In Algorithm 1, we present pseudo code showing the core loop of the system. The procedure involves some subprocedures described in Procedures 2, 3 and 4. The following notation is used: D is a set of documents d, T is a set of triples t, P is a set of extraction patterns p.

We can describe the procedure presented in Algorithm 1 as follows: Three resources exist in the system, which are sets of entities i.e. the Corpus consisting of documents, the Extraction Pattern Set consisting of extraction patterns, the ontology consisting of knowledge triples. Any one of these may be empty but, in order to run, two must have some seed data (cf. Table 25). There follows a loop controlled by the stopping criterion. Note that the StoppingCriterion may be a number of things. It may be the number of iterations, or achieving a certain level in the Knowledge Gap/Explicit Knowledge Gap measure, or a certain number of items above a certain threshold in Resource Confidence (cf.
Algorithm 1: The Abraxas Top level Algorithm

Data: $D_0, P_0, T_0$, such that $D_0$ is a set of documents, $P_0$ a set of
e EXTRACTION PATTERNS, and $T_0$ a set of triples, and where at most
one of $D_0, P_0, T_0$ is empty.

Result: $D_j, P_j, T_j$, such that iterations $j > 0$, and some
STOPPING CRITERION is satisfied.

1. Ranked Term list $W \leftarrow \emptyset$
2. for $d \in D_i$ do;
   /* Apply term recognition. */
3.   $W \leftarrow \text{terms}(d)$
4. end

begin

5. i = 0
6. while not STOPPING CRITERION do
7.   Candidate list $D', P', T' \leftarrow \emptyset$
8.   for $p \in P_i$ do;
   /* Apply $P_i$ to $D_i$, return triples */
9.     for $d \in D_i$ do
10.    $T' \leftarrow E(p, d)$
11. end
12. if $T' = \emptyset$ then
13.   $T', D' \leftarrow TE(D_i, P_i, T_i, W)$
14. end
15. for $t \in T_i$ do;
   /* Apply $T_i$ to $P_i$, return documents */
16.   for $p \in P_i$ do
17.     $D' \leftarrow E'(p, t, D_i)$
18. end
19. end
20. for $t \in T_i$ do;
   /* Apply $T_i$ to $D_i$, return patterns */
21.   for $d \in D_i$ do
22.     $P' \leftarrow E''(t, d)$
23. end
24. end
25. Score resources in $D', P', T'$ with RC
26. Get resource $r$ with highest RC from $D', P', T'$ and add to
27. whichever set from $D_i, P_i, T_i$
28. if $r = d \in D'$ then
29.   merge($W, \text{terms}(d)$)
30. end
31. i = i + 1
32. end
end
Algorithm 2: Extraction E(p, d)

Data: p is a pattern, d is a document.
Result: T', where T' is a set of candidate triples.

\[ T' \leftarrow \emptyset \]
\[ \text{begin} \]
\[ \text{forall matches of } p \text{ in } d \text{ do} \]
\[ T' \leftarrow T' \cup \{t\} \]
\[ \text{end} \]
\[ \text{return } T' \]
\[ \text{end} \]

Algorithm 3: Extraction E'(p, t, D)

Data: p is a pattern, t is a triple, and D a set of documents.
Result: D', where D' is a set of candidate documents.

\[ D' \leftarrow \emptyset \]
\[ \text{begin} \]
\[ \text{ip} = \text{Instantiate}(p, t) \]
\[ \text{for } d \in D \text{ do} \]
\[ \text{if } \text{ip found in } d \text{ then} \]
\[ D' \leftarrow D' \cup \{d\} \]
\[ \text{end} \]
\[ \text{end} \]
\[ \text{if } D' \neq \emptyset \text{ then} \]
\[ \text{return } D' \]
\[ \text{else} \]
\[ \text{search external repository} \]
\[ \text{if } \text{ip found in external } d \text{ then} \]
\[ D' \leftarrow D' \cup \{d\} \]
\[ \text{end} \]
\[ \text{end} \]
\[ \text{return } D' \]
\[ \text{end} \]

Algorithm 4: Extraction E''(t, d)

Data: t = <w_1, R, w_2>, d is a document.
Result: P', where P' is a set of candidate patterns.

\[ P' \leftarrow \emptyset \text{ begin} \]
\[ \text{foreach sentence } s \text{ in } d \text{ do} \]
\[ \text{if } w_1, w_2 \text{ in } s \text{ then} \]
\[ p = \text{patternLearn}(w_1, w_2, s) \]
\[ P' \leftarrow P' \cup \{p\} \]
\[ \text{end} \]
\[ \text{end} \]
\[ \text{return } P' \]
\[ \text{end} \]
Algorithm 5: External Triple Extraction TE(T, D, P)

Data: T, D, P, W

Result: T′, a set of candidate triple, D′, a set of corresponding documents

1. T′ ← Ø
2. D′ ← Ø
3. while T′ == Ø do
   4. for w ∈ W do
      5. for t = <w₁, R, w₂> ∈ T do
         6. for w′ ∈ {w₁, w₂} do
            7. for p ∈ P do
               8. ip ← Instantiate(p, w, w′)
               9. Match ip in external repository
               10. if ip found in d ∈ external repository then
                    11. T′ ← T′ ∪ {t} where t is a triple corresponding to ip, D′ ← D′ ∪ {d}
                    12. Delete w from W
               end
            end
         end
      end
   end
4. return T′, D′

below). The loop creates three empty sets for a candidate queue, and then performs three steps. It applies the lexico-syntactic patterns to the corpus in order to identify all knowledge triples to be found, it instantiates each triple with each lexico-syntactic pattern to identify documents either in the existing corpus or in an external repository which provide evidence for that triple, and finally learns new lexico-syntactic patterns from the triples and documents in the resource sets. Note: this is an abstract description which does not take into account of programmatic techniques to avoid repeatedly issuing the same query etc.

Formally, let A be a set, consisting of three sets {T, D, P} where T is a set of knowledge triples, D is a set of Documents, P is a set of Extraction Patterns. Starting state S₀ = {T₀, D₀, P₀} where T₀, D₀ or P₀ may be empty, but at least two of S₀ must be non-empty. The Resource Confidence (RC) of any t ∈ T₀, d ∈ D₀, or p ∈ P₀ is given manually for the first iteration. The output of Abraxas, at any given iteration, is a set of triples, a set of patterns, and a set of documents, together with a corresponding set of confidence levels for each resource item.
10.4 METRICS FOR MEASURING IN-EQUILIBRIUM

Our approach to managing knowledge is based on using a number of metrics to measure the difference between resources and the confidence values attached to each resource. The difference measures concern the difference between what the resource provides and its fit with the other resources. Thus a corpus may have more terms or keywords in it than present in the ontology – this defines a system level in-equilibrium and thus triggers certain actions. The confidence value measures also provide a measure of the system state and thus may trigger a certain action.

These metrics are used together with external measures, such as defining the number of iterations, to determine when the system will either stop or request input from the user. In this section we describe the Knowledge Gap and Explicit Knowledge Gap measures. In the longer term, we believe a number of metrics need to be identified which reflect the internal dynamics of knowledge processing. For example, how do human beings decide when they have enough knowledge concerning a given subject? How many examples of the usage of a term are needed for comprehension of the term and consequent addition to the body of knowledge? How much contradictory data is needed before a piece of knowledge is rejected entirely? These are questions for further research.

10.4.1 The Resource Confidence Measure

A key measure in the Abraxas system is the Resource Confidence Measure (RCM) which measures the confidence that the system has in a given resource (e.g. document, extraction pattern, or knowledge triple). From an intuitive perspective, the RCM provides a means for the confidence in any given item of knowledge to be determined, and in a complementary fashion the confidence for any given extraction pattern or document. The RCM value for a knowledge triple reflects how confident the system is that it is a correct piece of knowledge (in the sense that the system has found evidence for its assertion in the text). The RCM value for an extraction pattern reflects how confident the system is that the extraction pattern will extract accurate pieces of knowledge. The RCM value for a document reflects how confident the system is that the document belongs to the domain of interest and provides valid knowledge triples by providing at least one sentence where an extraction pattern matches.

As noted above, in the Abraxas approach, manual intervention can take the form of specifying that a resource belongs to the domain (and therefore has an RCM [1.0]) or that it does not belong (and therefore has an RCM [0.0]). Equally, the user could specify that they do not know or have an opinion about a given resource and specify it as having an RCM of [0.5]. However, system added resources, whether

2 The initial formalisation of this measure is largely due to José Iria cf. Iria et al. [2006].
documents, knowledge triples or extraction patterns are assumed to have varying degrees of confidence associated with them. The confidence value is a function of the success or suitability of a given resource in deriving the corresponding other resource. For example, we evaluate a given Extraction Pattern in terms of known knowledge triples and documents. If the precision of an extraction pattern is high in the known/learning context, then we can assign it a high confidence value so that the system can be confident that any further knowledge triples extracted by that Extraction Pattern will be correct.

In a similar manner, we can assign a certain Resource Confidence Measure to a knowledge triple depending on its source. If it has been added to the Ontology Set manually its RCM is \([1.0]\), if it has been identified by using a certain Extraction Pattern over a specific set of texts, then the confidence value is a function of the confidence values assigned to the Extraction Patterns and Documents involved. Equally, for a given document, we can assign a confidence value dependent on how fitting to the ontology the knowledge triples extracted are given the existing Extraction Patterns. Thus for each resource set, confidence for any resource item is defined in terms of the other resource sets. This means that for any given resource, there is a corresponding set of resource pairs with which it interacts.

The formulae for calculating the RCM of any given resource are designed so that a) a single measure combines the effect of the other types of resources; b) the greater the sum of the confidence/RC values of the other resource pairs a given resource is associated with, the greater is the RC of that resource; c) the measure should take into account resource pairs not covered. As will be seen below, the calculation of the RCM is rather complex. First we will present a justification for the measure as it is constructed and then we will present a more explicit re-formulation.

**JUSTIFICATION:** The RC measure is designed as a conceptual extension of the precision measure used in Information Retrieval and NLP. Thus it uses an equivalent of “true positives” and “false positives” to calculate the Resource Confidence. The measure is not a calculation over discrete values but rather real numbers between 0.0 and 1.0 reflecting the confidence the user or the system has in the classification of a resource. Typically users will specify a resource as having an RC of 0.0 or 1.0. Our equivalent of “true positives” we symbolise as \(r\) (right) and of “false positives” as \(w\) (wrong). Each resource has some positive contribution to the classification at hand that is equal to its resource confidence value, and a negative contribution that is equal to one minus its RC value. For example for some document \(d\), its confidence reflects how likely it is to belong to the corpus of the

---

3 The use of a different terminology from the standard one of false positive, false negative etc. is motivated by the fact that the Resource Confidence measure is trying to measure something somewhat different from precision, recall or accuracy of the system. We will note however the approximate equivalence as we explain the details below.
domain. If we are certain, then \( r = 1.0 \) and \( w = 0.0 \). However, if we are less certain, say our confidence is 0.33, then \( r = 0.33 \) and \( w = 0.67 \). In consequence, we can describe the basic Resource Confidence measure as shown in Equation 10.1 where RT is some resource type such as patterns, documents, or triples.

\[
RC(\text{RT}) = \frac{r}{r + w}
\]  

(10.1)

A basic concept in the Abraxas approach, and the conceptualisation of the Resource Confidence measure, is that resources are scored based on their complementary resource types. This means that given an arbitrary resource \( \text{RT}_i \), the resource confidence of that resource is calculated based on all the resources in the system up to that iteration which are not of that resource type. The simplest way to do this is to sum the contribution of the other resource types. Thus:

\[
\begin{align*}
\text{RT}_i &= \sum_{k \neq \text{RT}_i} r_k \quad \text{(10.2)} \\
\text{RT}_i &= \sum_{k \neq \text{RT}_i} w_k \quad \text{(10.3)}
\end{align*}
\]

where RT is some resource type (triples, patterns, documents). For example, let us suppose a pattern \( p_1 \) matches \( \{(t_1, d_1), (t_2, d_1), (t_1, d_2)\} \) (i.e. triple 1 in document 1, triple 2 in document 1, and triple 1 in document 2). Then:

\[
\begin{align*}
\text{r}_{p_1} &= \text{r}_{t_1} + \text{r}_{t_2} + \text{r}_{d_1} + \text{r}_{d_2} \\
&= \text{r}_T + \text{r}_D \quad \text{(10.4)} \\
\text{w}_{p_1} &= \text{w}_{t_1} + \text{w}_{t_2} + \text{w}_{d_1} + \text{w}_{d_2} \\
&= \text{w}_T + \text{w}_D \quad \text{(10.6)}
\end{align*}
\]

where

\[
\text{r}_T = \sum_{t_j \in T} \text{r}_{t_j} \quad \text{(10.8)}
\]

and correspondingly for \( \text{r}_D, \text{w}_T \) and \( \text{w}_D \). To recap so far, taking a pattern as the resource type in question:

\[
\begin{align*}
\text{RC}(p) &= \frac{r}{r + w} \\
r &= \text{r}_T + \text{r}_D \quad \text{(10.9)} \\
w &= \text{w}_T + \text{w}_D \quad \text{(10.10)}
\end{align*}
\]
This means that the RC of pattern \( p \) is calculated by dividing RC of the correct complementary resources \( r \) by the sum of the correct and incorrect complementary resources \( r + w \). In this case (as we are calculating for a pattern), \( r \) is defined as the sum of the resource confidence of the correct triples identified by the pattern and the correct documents identified by the pattern. Correspondingly, \( w \) is defined as the sum of the complement resource confidence of the triples and documents identified by the pattern.

There is one further complication to the RC measure. We need the measure to take into account the amount of resource space covered. For example, for a pattern, ideally the pattern would match all high confidence documents and triples and none of the low confidence ones. As the formula stands a pattern that matches one confidence \([1.0]\) document-triple pair and zero confidence \([0.0]\) document-triple pairs will have the same score as a pattern that matches one thousand confidence \([1.0]\) document-triple pairs and zero confidence \([0.0]\) document-triple pairs. We need some sort of term in the formula which modifies the result in proportion to the size of the data set. In order to achieve this we add to the simple formula defining \( r \) (10.10) and \( w \) (10.11) an extra term which makes the formulation take into account the size of the whole population, or resource space covered. In the following explanation we will present a version of the formula which is simpler in notation but more abstract. Below we will present a second formulation which is more complex in notation but possibly more explanatory.

The extra term provides a formalisation of a “coverage ratio” i.e. the ration of set of resources selected to the whole set of resources e.g. documents matched to total documents. So for any resource type where:

\[
\begin{align*}
\text{r}_{\text{matched}} &= \sum_{i \in \text{matched}} r_i \quad (= r_{RT}) \\
\text{w}_{\text{matched}} &= \sum_{i \in \text{matched}} w_i \quad (= w_{RT}) \\
\text{r}_{\text{total}} &= \sum_{j \in \text{total}} r_j \quad (= p_{RT}) \\
\text{w}_{\text{total}} &= \sum_{j \in \text{total}} w_j \quad (= n_{RT})
\end{align*}
\]

we can then define a ratio which acts to “shift” or “modulate” the final Resource confidence value (note the addition of 1 to the denominator to avoid division by zero):

\[
\begin{align*}
\text{modulator of } r &= \frac{(n_{RT} - w_{RT})}{(n_{RT} - w_{RT}) + (p_{RT} - r_{RT}) + 1} \quad (10.16) \\
\text{modulator of } w &= \frac{(p_{RT} - r_{RT})}{(p_{RT} - r_{RT}) + (n_{RT} - w_{RT}) + 1} \quad (10.17)
\end{align*}
\]

Thus we can refine the formulas in (10.10) and (10.11) (still defining it
for a pattern, as an example) as follows:

\[
\begin{align*}
    r &= r_T + \frac{(n_T - w_T)}{((n_T - w_T) + (p_T - r_T)) + 1} + r_D + \frac{(n_D - w_D)}{((n_D - w_D) + (p_D - r_D)) + 1} \\
    w &= w_T + \frac{(p_T - r_T)}{((p_T - r_T) + (n_T - w_T)) + 1} + w_D + \frac{(p_D - r_D)}{((p_D - r_D) + (n_D - w_D)) + 1}
\end{align*}
\]  

Figure 24. The conceptual space over which the RC measure is calculated.

We have avoided using the terms false positive, true negative etc. for the reasons mentioned above that we wish the Resource Confidence measure to be seen as a more abstract measure even if it bottoms out to actual calculations over resources matched. With this caveat in mind, we can visualise the resource space as shown in Figure 24 using the notation used above, and show the correspondence to traditional IR/NLP terminology by comparing it with Figure 25. The correspondences can be seen as follows:

- \( r_{RT} \approx TP(\text{true positives}) \)
- \( w_{RT} \approx FP(\text{false positives}) \)
- \( p_{RT} \approx FN + TP(\text{false negatives + true positives}) \)
- \( n_{RT} \approx TN + FP(\text{true negatives + false positives}) \)
10.4 Metrics for Measuring In-Equilibrium

Figure 25. The corresponding space expressing in usual NLP terms. A list of measures used in NLP, clinical research and other research domains is provided as well.

- $p_{RT} - r_{RT} \approx FN(\text{false negatives})$
- $n_{RT} - w_{RT} \approx TN(\text{true negatives})$

Although the formulation of equations above was arrived at independently, the formula in (10.16) is equivalent to the Negative Predictive Value formula used in clinical testing [Altman and Bland, 1994]. This is defined as shown in Figure 25. No equivalent has been found of the formula in (10.17).

**Rephrasing for Explicitness:** In this section, we rephrase the equations presented above so as to be more explicit as to the terms involved. For a given extraction pattern $p$, we aim to combine in one single measure the effect of both triples that the pattern extracts, and the documents that pattern extracts over. A triple-document pair is defined as the instance of a knowledge triple extracted from a given document. The measure favours extraction patterns that result in many knowledge triple-document pairs (instances) and favours extraction patterns that cover highly likely knowledge triple-document pairs. The Resource Confidence (RC) for a given Extraction Pattern (for example $p$) is shown in Eq. (10.20).

$$RC(p, D_I, P_i, T_i) = \frac{r}{r + w}$$ \hspace{1cm} (10.20)

We will define $r$ and $w$ below, but because the RC is a complex abstract measure we will first define and motivate the component
parts. We need $r$ to combine the contribution of the RC of both triples and documents correctly extracted over by $p$ and $w$ to combine the contribution of the RC of both triples and documents incorrectly extracted by $p$. In order to do this, first, we will present a number of subsidiary formulae which will be used:

Let $P_i$ be the set of all patterns in iteration $i$, $T_i$ the set of all triples in iteration $i$, and $D_i$ the set of all documents in iteration $i$. $\text{PRC}(10.21)$ is the sum of the Resource Confidence for all patterns calculated based on the Resource Confidence of the triples extracted:

$$\text{PRC}(P, T, D) = \sum_{p_j \in P} \text{pTRC}(p_j, T, D)$$

$pRC$ (10.22) is the sum of the Resource Confidence for a particular pattern calculated based on the Resource Confidence of the triples extracted:

$$pTRC(p, T, D) = \sum_{d \in D} \sum_{t \in E(p, d)} \text{RC}(t)$$

Note the use of the extraction function $E(p, d)$ which is defined in the algorithm descriptions above (Algorithms 2-4).

$\overline{\text{PRC}}(10.23)$ is the sum of the Resource Confidence all patterns calculated based on the complement of the RC for all the triples extracted:

$$\overline{\text{PRC}}(P, T, D) = \sum_{p_j \in P} \overline{\text{pTRC}}(p_j, T, D)$$

$\overline{\text{pTRC}}$ (10.24) is the sum of the Resource Confidence for a particular pattern calculated based on the complement of the Resource Confidence of the triples extracted:

$$\overline{\text{pTRC}}(p, T, D) = \sum_{d \in D} \sum_{t \in E(p, d)} (1 - \text{RC}(t))$$

The corresponding formulae for the RC of the documents extracted by the pattern are as follows: $\text{pDRC}$ (10.25) is the sum of the resource confidence for a particular pattern calculated based on the Resource Confidence of the documents identified by the pattern:

$$\text{pDRC}(p, D, T) = \sum_{t \in T} \sum_{d \in E(p, t, D)} \text{RC}(d)$$

$\overline{\text{pDRC}}$ (10.26) is the sum of the resource confidence for a particular pattern calculated based on the complement of the Resource Confidence of the documents identified by the pattern:

$$\overline{\text{pDRC}}(p, D, T) = \sum_{t \in T} \sum_{d \in E(p, t, D)} (1 - \text{RC}(d))$$
PDRC (10.27) is the sum of the Resource Confidence for all patterns calculated based on the Resource Confidence of all the documents identified:

$$\text{PDRC}(p_i, D_i, T_i) = \sum_{p_j \in P_i} \text{pDRC}(p_j, D_i, T_i)$$

(10.27)

PDRC (10.28) is the sum of the Resource Confidence for all patterns calculated based on the complement of the Resource Confidence of all the documents identified:

$$\text{PDRC}(p_i, D_i, T_i) = \sum_{p_j \in P_i} \overline{\text{pDRC}}(p_j, D_i, T_i)$$

(10.28)

At a simple level, Equation 10.20, could be constructed using merely the values from Equations (10.22), (10.25), (10.24) and (10.26). Thus the equivalent of Equations (10.10) and (10.11) would be:

$$r = \text{pTRC}(p_j, T_i, D_i) + \text{PDRC}(p_j, D_i, T_i)$$

(10.29)

$$w = \text{pTRC}(p_j, T_i, D_i) + \text{PDRC}(p_j, D_i, T_i)$$

(10.30)

but to take into account the factors described above which resulted in the equations (10.18) and (10.19), we reformulate the definitions of r and w as shown in equations (10.31), (10.32):

$$r = \text{pTRC}(p_j, T_i, D_i)
+ \frac{\text{pDRC}(p_j, D_i, T_i)}{((\text{pTRC}(p_j, T_i, D_i) - \text{pTRC}(p_j, T_i, D_i)) + \text{pDRC}(p_j, D_i, T_i)) + 1}$$

(10.31)

$$w = \text{pTRC}(p_j, T_i, D_i)
+ \frac{\text{pDRC}(p_j, D_i, T_i)}{((\text{pTRC}(p_j, T_i, D_i) - \text{pTRC}(p_j, T_i, D_i)) + \text{pDRC}(p_j, D_i, T_i)) + 1}$$

(10.32)

Following the same principles corresponding equations can be constructed for the RC of a triple or a document (cf. the worked example below in Section 10.5).
10.4.2 The Knowledge Gap Measure

Let us suppose we have defined initially a seed corpus, which may be either a given set of domain texts or a set of texts retrieved using a seed ontology to provide the query terms. Whether retrieved from the Web or from a specific text collection is immaterial. At an abstract level this seed corpus contains a certain quantum of knowledge, or more accurately, in the light of Brewster et al. [2003], one must assume background knowledge which is the relevant ontology we wish to build. The knowledge gap (KG) is the difference between an existing ontology and a given corpus of texts[^4]. The knowledge gap is measured by identifying the key terms in the corpus and comparing these with the concept labels or terms in the ontology. Such an approach depends on being able to successfully identify all the key terms in a given corpus and also assumes that all the key terms should be included in the ontology[^5].

We define the knowledge gap measure (KG) as the normalisation of the cardinality of the difference in the set of terms in the ontology ($\Omega$) vs. the set of terms in the corpus ($\Sigma$) as shown in Eq. (10.33) (a fully worked example to justify this formulation is presented in Appendix B):

$$KG = \frac{|(\Omega \cup \Sigma) \setminus (\Omega \cap \Sigma)|}{|\Omega \cup \Sigma|}$$  (10.33)

At the beginning of the ontology learning process, KG will initially be either 1 (maximal) or very close to 1, indicating a large ‘gap’ between the knowledge present in the ontology and that which needs to be represented which is latent in the corpus. As the ontology learning process progresses, the objective is to minimise KG as much as possible while realising, in accordance with Zipf’s law, that this is an asymptote.

There are a wide variety of methods for identifying the salient or key terms in a corpus of texts (e.g. Maynard and Ananiadou [2000] or Ahmad [1995]) but the real challenge is to learn automatically the ontological relationship between terms (cf. Chapter 5 and Brewster and Wilks [2004]). It is also relatively un-contentious to use distributional methods to show that there exists a relationship between two terms, and in a large proportion of these cases that relationship will be an ontologically relevant one. In Chapter 7, explicit knowledge is defined as textual environments where ontological knowledge is expressed in a lexico-syntactic pattern of the type identified by Hearst [1992]. In such a case, the ontological relationship (domain - predicate - range) is automatically extractable. However, for any given corpus, a minority of the terms will occur as explicit knowledge and thus can be automatically added to the set of ontological knowledge. The difference between the set of pairs of terms whose ontological relationships are

[^4]: This is closely related to the notion of ‘fit’ we have proposed elsewhere [Brewster et al., 2004]

[^5]: In practice, we believe this to be a relative scale i.e. terms are recognised as key terms and therefore to be included in the ontology with varying degrees of confidence
known (because they are in the ontology) and those which need to be added to the ontology (because they are key terms in the corpus) and whose ontological relationship is unknown, we will term the Explicit Knowledge Gap (EKG). The absence of explicit knowledge may be between two terms both absent from the ontology or between a new term and a term already assigned to the ontology set. Let $\Pi$ be the set of triples based on terms existing in the corpus, which are known to have some kind of ontological relationship on distributional grounds. Each triple consists of $t_i, r, t_k$ where $t_i, t_k \in T$ the set of all terms in the corpus, and $r \in R$ the set of ontological relations. The set $\Pi$ consists of triples where the ontological relationship is unknown (let us represent this by $r_0$). In contrast, $\Omega_3$ is the set of pairs of terms whose ontological relationship is explicit, and these are assumed to be the set of knowledge triples in the ontology. Again obviously these can be represented as triples consisting of $t_i, r_j, t_k$ where, in this case, $t \in \Omega$ i.e. the set of terms in the ontology.

Thus EKG is defined analogously to Eq. 10.33, where $\Omega$ is replaced by $\Omega_3$, and $\Sigma$ is replaced by $\Pi$ (for a worked example, cf. Appendix B).

$$EKG = \frac{|(\Omega_3 \cup \Pi) \setminus (\Omega_3 \cap \Pi)|}{|\Omega_3 \cup \Pi|}$$

(10.34)

While EKG will never be 0, in a similar manner to KG, one objective of the system is to minimise this Explicit Knowledge Gap. The seed corpus will have relatively high KG and EKG measures. The expansion of the corpus occurs in order to reduce the two respective knowledge gaps and consequently learn the ontology. As this is an iterative process we can conceive of the expanding corpus and the consequently expanding knowledge like ripples in a pool (Figure 22).

The exact formulation of these equations will change and further metrics are likely to be added as our understanding develops of what it means to represent uncertain knowledge and the degrees of uncertainty. For example, we are still unclear as to how to model correctly the confidence values which depend on the nature of the source of the knowledge. How do we globally make knowledge derived from source $X$ more confidence worthy than source $Y$, especially if the latter has greater quantity so a given knowledge triple may be derived more frequently and yet may be incorrect. This raises the issue of trust of internet based sources, a topic researchers are only beginning to understand [O’Hara et al., 2004, Artz and Gil, 2006].

10.5 A WORKED EXAMPLE

As an example of our approach, we present here a small walk-through of the process. At each step we will describe what actions are being performed by the system and how this affects the certainty in the model as the steps proceed. The model and its representation of certainty is still work in progress because it is quite complex. There are
Hedgehog From Wikipedia, the free encyclopaedia Jump to: navigation, search For the anti-submarine weapon, see Hedgehog (weapon). For the area denial weapon, see Czech hedgehog. For the video game character, see Sonic the Hedgehog (character). *Hedgehogs* European Hedgehog European Hedgehog *Scientific classification* Kingdom: Animalia Phylum: Chordata Class: Mammalia Order: Insectivora Family: Erinaceidae Subfamily: *Erinaceinae* Genera Atelerix Erinaceus Hemiechinus Mesechinus A hedgehog is any of the small spiny mammals of the subfamily *Erinaceinae* and the order Insectivora. There are 15 species of hedgehog in four genera, found through parts of Europe, Asia, Africa, and New Zealand. There are no hedgehogs native to the Americas or Australia. Contents *1 Physical description* *2 Behaviour* *3 Diet* *4 Reproduction* *5 Domesticated hedgehogs* *6 Pest control* *7 Hedgehog diseases* *8 Edibility* *9 Trivia* *10 See also* *11 Genera and species* *12 External links* 0.1 Hedgehog information 0.2 Hedgehog breeders-supplies

Hedgehog

Physical description: Hedgehogs are easily distinguished by their spines, which are hollow hairs made stiff with keratin. Their spines are not poisonous or barbed and, unlike the quills of a porcupine, cannot easily be removed from the animal. However, spines normally come out when a hedgehog sheds baby spines and replaces them with adult spines around the first year. When under extreme stress or during sickness, a hedgehog will lose spines. Hedgehogs are most closely related to gymnures and other insectivores, including moles, shrews, tenrecs and solenodons. Behaviour: A defense that all species of hedgehogs possess is the ability to roll into a tight ball, causing all of the spines to point outwards. However, its effectiveness depends on the number of spines, and since some of the desert hedgehogs evolved to carry less weight, they are much more likely to try to run away and sometimes even attack the intruder, trying to ram into the intruder with its spines, leaving rolling as a last resort. This results in a different number of predators for different species: while forest hedgehogs have relatively few, primarily birds (especially owls) and ferrets, smaller species like long-eared hedgehogs are preyed on by foxes, wolves and mongooses. All hedgehogs are primarily nocturnal, although different species can be more or less likely to come out in the daytime. The hedgehog sleeps for a large portion of the daytime either under cover of bush or grass or rock or in a hole in the ground. Again, different species can have slightly different habits, but in general hedgehogs dig out dens for shelter. All wild hedgehogs can hibernate, although not all do; hibernation depends on temperature, abundance of food and species. Hedgehogs are fairly vocal, and communicate not only in a series of grunts and snuffles, but sometimes in loud squeals (depending on species). Hedgehogs occasionally perform a ritual called ‘anointing’. When the animal comes across a new scent, it will lick and bite the source and then form a scented froth in its mouth and paste it on its spines with its tongue. This camouflages the hedgehog with the new scent of the area and provides a possible poison or source of infection to any predator that gets poked by their spines.

Table 26. Document d1 - part of the entry on Hedgehog from Wikipedia.

different types of resource confidence (confidence in a triple, confidence in an extraction pattern, confidence in a text, confidence that a term belongs in the ontology, confidence that one term is ontologically related to another term, etc.). Our three initial resources are populated as follows:

- Corpus: one text shown in Table 26 (RC: [1.0])
- Ontology/Triples: one triple mammal ISA animal (RC: [1.0])
- Extraction Patterns: Word association and the lexico-syntactic patterns as discussed in Chapter 7. (The RC for each pattern is empirically derived from the general results in that chapter.)

This is a minimal set of resources with which the system can slowly, iteratively build an ontology. This an arbitrary example because as we have noted above in principle each resource can be empty as long as the two others provides a seed set of data.

6 It remains for us an open research issue whether this is true of the extraction pattern set. In principle, for reasons of aesthetic completeness, we would like this to be true.
Confidence value: An item introduced by the user is assigned an RC by the user [0.0 - 1.0] with [0.5] indicating “unknown” i.e. no evidence in favour or against. Typically seed resources will be assigned [1.0] but not necessarily.

**STEP 1: TERM EXTRACTION** Extract the key terms from the text. In this case we apply the simple approach proposed by Scott [1997] to identify key words i.e. comparison with a reference corpus. From this we obtain, the list shown in Table 27.

<table>
<thead>
<tr>
<th>Term</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>hemiechinus</td>
<td>6.07e+23</td>
</tr>
<tr>
<td>spines</td>
<td>1382.56</td>
</tr>
<tr>
<td>hedgehog</td>
<td>28376.5</td>
</tr>
<tr>
<td>species</td>
<td>126.51</td>
</tr>
<tr>
<td>erinaceinae</td>
<td>1.82e+23</td>
</tr>
<tr>
<td>hibernate</td>
<td>8672.44</td>
</tr>
<tr>
<td>subfamily</td>
<td>5203.47</td>
</tr>
<tr>
<td>domesticated</td>
<td>1110.5</td>
</tr>
<tr>
<td>snails</td>
<td>731.41</td>
</tr>
<tr>
<td>predator</td>
<td>318.95</td>
</tr>
<tr>
<td>fatty</td>
<td>308.16</td>
</tr>
<tr>
<td>poison</td>
<td>186.6</td>
</tr>
<tr>
<td>liver</td>
<td>110.04</td>
</tr>
<tr>
<td>weapon</td>
<td>91.89</td>
</tr>
<tr>
<td>desert</td>
<td>83.93</td>
</tr>
</tbody>
</table>

Table 27. Top keywords from the hedgehog text above. Number reflect relative ‘keyness’ to the BNC.

**STEP 2: MEASURE THE KNOWLEDGE GAP** This is an example application of an equilibrium measure. In this case, the ontology contains one triple i.e. two terms *(mammal, animal)* and (for the sake of this example) the text contains 15 terms. Applying the formula shown in Equation (10.33), we obtain a figure of 1.0. This indicates that there is maximal state of dis-equilibrium and the system needs to take steps to achieve an equilibrium. The KG and EKG measures are discussed further below in Section 10.5.1.

**STEP 3: LIST TERMS TO BE PROCESSED** Identify the terms in the text absent from the ontology and rank them by frequency. Then for each term:

**STEP 3.1: ATTEMPT EXTRACTION PATTERN APPLICATION** Try to apply the existing lexico-syntactic patterns in order to associate the new term \( W_n \) with an existing term \( W_e \) in the ontology. This is applied in a cascade of ever more precise extraction patterns. As noted above in Section 7.4, the most basic semantically neutral extraction pattern is word association or clustering. The second category is generic ontologically explicit lexico-syntactic extraction patterns and the third category is domain specific learned extraction patterns. At this stage, we can only apply word association and generic lexico-syntactic extraction patterns. So if our new term \( W_n \) is *hedgehog* and

---

7 For a review of a number of term recognition approaches useful in this context cf. Zhang et al. [2008]
the existing terms $W_i \in \{\text{mammal, animal}\}$, then we check for all possible instantiations of our lexico-syntactic patterns in the text. Some instantiated lexico-syntactic patterns are shown in Table 28.

The size of this set depends on how rich the set of starting extraction patterns is. This set is relatively conservative (cf. discussion in Chapter 7). If we find such an explicit context (for example $\text{hedgehog is a kind of mammal}$\textsuperscript{8}, then we would be able to add the appropriate triple to the ontology (Step 4). As we have not found anything we need to expand the corpus (Step 3.2).

Confidence value: The confidence the system has in an ontological fact is a function of a. the reliability of the extraction pattern, b. the reliability of the specific source, and c. the frequency of occurrence in the resources the system has (i.e. documents in the corpus in this case).

### Table 28. Instantiated extraction patterns.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>mammal, a type of hedgehog</td>
<td>animal, a kind of hedgehog</td>
</tr>
<tr>
<td>mammal, a kind of hedgehog</td>
<td>animal is a kind of hedgehog</td>
</tr>
<tr>
<td>mammal is a kind of hedgehog</td>
<td>animal is a type of hedgehog</td>
</tr>
<tr>
<td>mammal is a type of hedgehog</td>
<td>hedgehogs such as animals</td>
</tr>
<tr>
<td>hedgehogs such as mammals</td>
<td>animal and other hedgehogs</td>
</tr>
<tr>
<td>mammal and other hedgehogs</td>
<td>animals and other hedgehogs</td>
</tr>
<tr>
<td>mammals and other hedgehogs</td>
<td>animal or other hedgehogs</td>
</tr>
<tr>
<td>mammal or other hedgehogs</td>
<td>hedgehogs or other mammals</td>
</tr>
<tr>
<td>mammals or other hedgehogs</td>
<td>hedgehog or other hedgehogs</td>
</tr>
<tr>
<td>hedgehog, a type of mammal</td>
<td>hedgehog and other animals</td>
</tr>
<tr>
<td>hedgehog, a kind of mammal</td>
<td>hedgehogs and other animals</td>
</tr>
<tr>
<td>hedgehog is a kind of mammal</td>
<td>hedgehog, a type of animal</td>
</tr>
<tr>
<td>hedgehog is a kind of mammal</td>
<td>hedgehog, a kind of animal</td>
</tr>
<tr>
<td>hedgehog is a kind of mammal</td>
<td>hedgehog, a kind of animal</td>
</tr>
<tr>
<td>mammals such as hedgehogs</td>
<td>hedgehogs or other animals</td>
</tr>
<tr>
<td>hedgehog and other mammals</td>
<td>hedgehog or other animals</td>
</tr>
<tr>
<td>hedgehogs and other mammals</td>
<td>hedgehog and other animals</td>
</tr>
<tr>
<td>hedgehog or other mammals</td>
<td>hedgehogs and other animals</td>
</tr>
<tr>
<td>hedgehogs or other mammals</td>
<td>hedgehogs or other animals</td>
</tr>
</tbody>
</table>

---

\textsuperscript{8} This is the second most precise lexico-syntactic pattern according to our investigation in Section 7.3.

\textsuperscript{9} For example, the term \textit{stock}, which is ambiguous between 27 senses according to WordNet, occurs almost exclusively in the financial sense in the Wall Street Journal (p.c. Nancy Ide).
10.5 A WORKED EXAMPLE

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>hedgehogs such as mammals</td>
<td>0</td>
</tr>
<tr>
<td>mammals such as hedgehogs</td>
<td>193</td>
</tr>
<tr>
<td>hedgehogs such as animals</td>
<td>0</td>
</tr>
<tr>
<td>animals such as hedgehogs</td>
<td>289</td>
</tr>
<tr>
<td>mammal is a kind of hedgehog</td>
<td>0</td>
</tr>
<tr>
<td>hedgehog is a kind of mammal</td>
<td>0</td>
</tr>
<tr>
<td>animal is a kind of hedgehog</td>
<td>0</td>
</tr>
<tr>
<td>hedgehog is a kind of animal</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 29. Two instantiated extraction patterns with corresponding Yahoo document counts.

are appropriate to the domain. If we find a text containing a given instantiated Extraction Pattern, this text is added to the corpus, and we are able to a) add the triple to the ontology (Step 4) and b) apply term extraction (Step 1) to the expanded corpus, c) measure the knowledge gap (Step 2), and then d) create a new list of terms to be processed (Step 3).

In order to make this step as efficient as possible, the instantiated patterns are ranked by precision of pattern and frequency of the terms in the corpus. (Note the terms have been extracted due to their unusual relative frequency but once extracted, we deal with the most frequent first in order to reflect their significance in the corpus. This means in our case that the term ‘hedgehog,’ which occurs 123 times in document d1, will be treated first, in relation to the term ‘animal’ and ‘mammal’ (which already existed in our ontology). We can thus rank our list in Table 28, as shown in Table 29:

<table>
<thead>
<tr>
<th>Mammals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apart from two skeletons of whales (due to limitation of space) this confined mostly to small mammals such as hedgehogs, moles, bats, squirrels, gerbils, mice, rabbits, wolf, Jackals, fox, badger, wild cat and seal. Apart from few whole mounted ungulate such as gazelle and dear, only skulls of the others have been kept as a part of collection.</td>
</tr>
</tbody>
</table>

Table 30. Document d2, from the Tehran Museum website.

If we issue the first of these queries ("hedgehogs such as mammals"), we find no documents available. The second of these queries ("mammals such as hedgehogs") has nearly 200 URLs on Yahoo. So we start considering each document in turn to determine if it ‘belongs to the domain’ and if so add it to the corpus. This can be achieved by having a continuously growing queue of documents which are evaluated for their RC given the existing state of the other two resources, or more simply a vector space comparison can be made with the existing

---

10 The exact number varies depending on the moment of query but this does not impact on our methodology.
corpus\textsuperscript{11}. Let us say for simplicity’s sake that the top ranked text is found to be most similar to our existing corpus (strangely enough from the museum of the University of Tehran) \textsuperscript{12} shown in Table 30.

**STEP 4: ADD TRIPLE TO ONTOLOGY** The ontology is represented internally as a set of RDF type triples with associated confidence values. Given the steps we have undertaken so far, and the ‘evidence’ derived from our newly added document, the triple *hedgehog ISA mammal* can be added to the set of triples.

Confidence value: The confidence value for a given triple is dependent on its provenance. If the triple has been added manually, confidence would be assigned manually, if it is the result of an extraction pattern then this is a function of the (known) precision of the EP and the application of the RC formula discussed in Section 10.4.1.

**STEP 5: ADD AN EXTRACTION PATTERN** The addition of a new triple (*hedgehog ISA mammal*) as a result of adding the text shown in Table 30, opens another possibility. We recognise *animal* and *mammal* as linked by a given ontological relationship so if we can find a (usually sentential) context where these two terms co-occur, then we can ‘learn’ a new Extraction Pattern. In this case, in document d1 the sentence with these two terms is: “A hedgehog is any of the small spiny mammals of the subfamily Erinaceinae . . .” from which we can extract a candidate pattern *NP is any of the . . . NPP . . .*. This can be described as <Cs> is any of the . . . <Pp> using the formalism of Chapter 7.

Confidence value: The confidence value for an extraction pattern is again dependent on provenance. As in this case it is the result of the extraction pattern learning process, its RC is a function of the documents and triples it extracts.

**POSTPROCESSING: ONTOLOGY SET NORMALISATION** This step occurs once the stopping criterion is reached and an output ontology is desired. As the Ontology set grows to include more and more knowledge triples, it becomes necessary to ensure that all triples are somehow linked together and where this does not occur to seek a remedy. For example, let us suppose the Ontology set $O = \{(mammal ISA animal), (hedgehog ISA mammal), (snail ISA animal), (liver ISA organ)\}$, we can see that the first three knowledge triples are linked in some way, while the 4th is not (Figure 26). This means either the 4th triple must be linked in some way (by seeking further knowledge triples to add) or it must be discarded from the ontology set.

The reader should note that:

\textsuperscript{11} Determining whether a document ‘belongs to the domain’ would normally be done by training on an existing set of documents using text categorisation methods and setting a relevant threshold, but in order to achieve maximum flexibility we have just used the top ranked document from the top n documents returned by the query (where n = 20) and thus no threshold is needed.

\textsuperscript{12} http://www.fos.ut.ac.ir/text/museums/musem/mammals.htm
The set of terms which need to be processed is a queue which inevitably will never be emptied. The decision to stop processing the queue of terms depends on a stopping criterion such as the degree of knowledge gap that is acceptable - this KG may be measured with respect to the total corpus collected or with respect to the initial seed corpus.

The ontology is described as a set of triples but in reality in order to have a real semblance of an ontology it needs to be put together as a coherent hierarchy. This process of ontology construction will identify a number of defects including:

1. ambiguity of hierarchical value e.g. A ISA B, C ISA B - what are C and A to each other? Siblings, parent/child etc.?
2. orphan triples i.e. pairs of terms which have occurred in explicit contexts but which have not been successfully merged into the ontology as a whole. In both these cases, new data needs to be obtained from the external sources to resolve these problems.

10.5.1 Calculating the measures

Every event in the Abraxas system has the potential to alter the equilibrium between the components and events are often complex in their repercussions. In our worked example, we merely calculated the KG and EKG and decided to add a document. This immediately resulted in the addition of a new triple, and triggered the learning of a new Extraction Pattern. Each event we will refer to as an iteration, and depending on the type of event in that iteration, it may trigger different responses from the system, including stopping, adding further resources, learning new EPs, adding triples etc. The responses are intimately linked to the calculations of the measures at each stage. As noted above there are two types of measure, system wide measures such as KG/EKG and individual resource measures such as the RCM.
168 A MODEL OF KNOWLEDGE ACQUISITION

| State   | $\Omega$ | $\Sigma$ | $\Omega \cup \Sigma$ | $\Omega \cap \Sigma$ | $|\Omega \cup \Sigma \setminus (\Omega \cap \Sigma)|$ | KG (10.33) |
|---------|----------|----------|------------------------|----------------------|---------------------------------|------------|
| Initial State | 2       | 15       | 17                     | 0                    | 17                               | 1.00       |
| Iteration 1 | “”      | “”       | “”                     | “”                   | “”                               | “”         |
| Iteration 2 | 3       | 15       | 17                     | 1                    | 16                               | 0.94       |
| Iteration 3 | “”      | “”       | “”                     | “”                   | “”                               | “”         |

Table 31. The calculation of the KG measure

**Stage A: Calculate KG and EKG**  We already noted above that in the seed state the Knowledge Gap measure (KG) was determined to be [1.0] and so clearly in complete dis-equilibrium. The KG for the seed state was calculated as follows as shown in Table 31. $\Omega$ is the set of terms in the knowledge triple set, so in the seed case $\Omega = \{\text{mammal, animal}\}$ and its cardinality is [2]. $\Sigma$ is the set of terms in the corpus and here we specify those as shown in table 27 (arbitrarily the top 15) and so the cardinality is [15]. There is no overlap between these sets of terms as shown in the value for $\Omega \cap \Sigma$. At Iteration 1 a document is added to the corpus and term extraction occurs again. Given that the new document is very small unsurprisingly this does not change the top terms extracted so $\Sigma$ remains the same. At Iteration 2, however, a new triple is added to the ontology so $\Omega$ changes: $\Omega = \{\text{mammal, animal, hedgehog}\}$ which has a cardinality of [3]. Now we see that $|\Omega \cap \Sigma| = 1$ because the term hedgehog is in both sets. Again nothing changes for the KG measure in Iteration 3 because this adds no new terms either to the corpus or the ontology sets.

With respect to the Explicit Knowledge Gap, the seed situation is one where $\Pi$ is the set of explicit knowledge triples in the corpus, and $\Omega_3$ is the set of knowledge triples in the Ontology. In the seed state, $\Pi = \emptyset$ while $\Omega_3 = \{(\text{mammal ISA animal})\}$ so their respective cardinalities are [0] and [1]. As there is no overlap between the two sets, EKG is [1.0] showing total dis-equilibrium. Iteration 1 adds a document in which there is an explicit knowledge triple (but not yet added to the ontology set) and thus $\Pi = \{(\text{hedgehog ISA mammal})\}$ whose cardinality is [1]. However, due to the absence of overlap between the two sets, EKG remains [1.0]. At Iteration 2, the newly identified knowledge triple is added to the ontology/knowledge triple set, and so $\Omega_3 = \{(\text{mammal ISA animal}), (\text{hedgehog ISA mammal})\}$ whose cardinality is [2]. As a result, the EKG becomes [0.5]. This indicates a significant reduction is dis-equilibrium in terms of explicit knowledge. At Iteration 3, no new explicit knowledge is identified nor any triple added to the triple set, so the EKG remains the same.

As further iterations occur in the system, more items of explicit knowledge are identified in the corpus and then added to the ontology triples and so $\Pi$ and $\Omega_3$ will tend to be almost in sync and EKG will
Tend towards $[0]$. The importance of the EKG lies when dealing with a more complex seed ontology which may have considerable explicit knowledge but where this is not reflected in the corpus and thus will guide the development of the corresponding corpus.

**Stage B: Calculate RC for Each New Resource** The initial or seed state of the system was specified by the user and so the one document had RC $[1.0]$, the one triple $[1.0]$ and the 9 extraction patterns the values derived empirically in Chapter 7 as shown in Table 34 in the column headed ‘seed state.’

**Iteration 1**

We can then apply the formulas presented above to calculate in the first instance the RC of the document added in Step 3.2 above. In order to calculate this RC of a document, we need to determine the corresponding RCs for all triples and extraction patterns which apply. In order to calculate equation 10.20 for a document, the equations 10.20, 10.31 and 10.32 have to be re-cast in terms of triples and patterns, as shown in 10.35, 10.36, and 10.37.

\[
\text{RC}(d, P_i, D_i, T_i) = \frac{r}{(r + w)} \tag{10.35}
\]

\[
r = \text{dTRC}(d_i, T_i, P_i)
\]

\[
+ \frac{|\text{dTRC}(D_i, T_i, P_i) - \text{dTRC}(d_i, T_i, P_i)|}{(|\text{dTRC}(D_i, T_i, P_i) - \text{dTRC}(d_i, T_i, P_i)| + (\text{dTRC}(D_i, T_i, P_i) - \text{dTRC}(d_i, T_i, P_i)) + 1}
\]

\[
+ \frac{|\text{dPRC}(d_i, P_i, T_i)|}{(|\text{dPRC}(D_i, P_i, T_i) - \text{dPRC}(d_i, P_i, T_i)| + (\text{dPRC}(D_i, P_i, T_i) - \text{dPRC}(d_i, P_i, T_i)) + 1} \tag{10.36}
\]

\[
w = \text{dTRC}(d_i, T_i, P_i)
\]

\[
+ \frac{|\text{dTRC}(D_i, T_i, P_i) - \text{dTRC}(d_i, T_i, P_i)|}{(|\text{dTRC}(D_i, T_i, P_i) - \text{dTRC}(d_i, T_i, P_i)| + (\text{dTRC}(D_i, T_i, P_i) - \text{dTRC}(d_i, T_i, P_i)) + 1}
\]

\[
+ \frac{|\text{dPRC}(d_i, P_i, T_i)|}{(|\text{dPRC}(D_i, P_i, T_i) - \text{dPRC}(d_i, P_i, T_i)| + (\text{dPRC}(D_i, P_i, T_i) - \text{dPRC}(d_i, P_i, T_i)) + 1}.
\]

| State     | $|\Pi|$ | $|\Omega_3|$ | $|\Pi \cup \Omega_3|$ | $|\Pi \cap \Omega_3|$ | $|\Pi \cup \Omega_3 \setminus (\Pi \cap \Omega_3)$ | EKG $(10.34)$ |
|-----------|--------|-------------|-----------------|-----------------|----------------------------------|-------|
| Initial State | 0      | 1           | 1               | 0               | 0                                | 1.0   |
| Iteration 1  | 1      | 1           | 2               | 0               | 2                                | 1.0   |
| Iteration 2  | 1      | 2           | 2               | 1               | 1                                | 0.5   |
| Iteration 3  |        |             |                 |                 |                                  |       |

Table 32. The calculation of the EKG measure
Corresponding to equations 10.21 to 10.28, we need to provide the appropriate equations for the component parts of \( r \) and \( w \) defined above:

\[
\begin{align*}
DTRC(D_i, T_i, P_i) &= \sum_{d_j \in D_i} dTRC(d_j, T_i, P_i) \quad (10.38) \\
dTRC(d_j, T_i, P_i) &= \sum_{p \in P_i} \sum_{t \in E(p, d_j)} RC(t) \quad (10.39) \\
DTRC[D_i, T_i, P_i] &= \sum_{d_j \in D_i} dTRC[d_j, T_i, P_i] \quad (10.40) \\
dTRC(d_j, T_i, P_i) &= \sum_{p \in P_i} \sum_{t \in E(p, d_j)} 1 - RC(t) \quad (10.41) \\
DPRC(D_i, P_i, T_i) &= \sum_{d_j \in D_i} dPRC(d_j, P_i, T_i) \quad (10.42) \\
dPRC(d_j, P_i, T_i) &= \sum_{t \in T_i} \sum_{p \in E''(d_j, t)} RC(p) \quad (10.43) \\
DPRC[D_i, P_i, T_i] &= \sum_{d_j \in D_i} dPRC[d_j, P_i, T_i] \quad (10.44) \\
RC(d_j, P') &= \sum_{t \in T_i} \sum_{p \in E''(d_j, t)} 1 - RC(p) \quad (10.45)
\end{align*}
\]

In Table 33 Equation (10.39) is calculated by multiplying the resource confidence of each triple times its existing RC and sums the total. In this case, we have one triple which as part of the seed state has had manually assigned an RC of 1.0, so 1 * 1.0 = 1.0. Equation (10.41) is the complement of this. Equations (10.38) and (10.40) are similar in this case because the set of triple-pattern matches contains only 1 element. Equation (10.43) sums over all the patterns matched in this document - in our case one pattern matches which has a manually entered RC of 0.99 (cf. Table 34 EP no.1). Equation (10.45) is the complement of this. Equations (10.42) and (10.44) operate over the total patterns in the set of triple-pattern pairs. Using these values, we can solve equations 10.36 and 10.37 and thus solve the overall RC value, shown in the RESULT row. This is the value presented in the Iteration 1 column in Table 34.
<table>
<thead>
<tr>
<th>Equation No.</th>
<th>Equation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10.39)</td>
<td>$dTRC(d_j, T_i, P_l) = \sum_{p \in P_i} \sum_{t \in E(p, d_j)} RC(t)$</td>
<td>1</td>
</tr>
<tr>
<td>(10.41)</td>
<td>$dTRC(d_j, T_i, P_l) = \sum_{p \in P_i} \sum_{t \in E(p, d_j)} 1 - RC(t)$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence in the triple matched in the specific document</td>
<td></td>
</tr>
<tr>
<td>(10.38)</td>
<td>$DTRC(D_i, T_i, P_l) = \sum_{d_j \in D_i} dTRC(d_j, T_i, P_l)$</td>
<td>1</td>
</tr>
<tr>
<td>(10.40)</td>
<td>$DTRC(D_i, T_i, P_l) = \sum_{d_j \in D_i} dTRC(d_j, T_i, P_l)$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence in the triple matched in all documents</td>
<td></td>
</tr>
<tr>
<td>(10.43)</td>
<td>$dPRC(d_j, P_i, T_l) = \sum_{t \in T_i} \sum_{p \in E^p(d_j, t)} RC(p)$</td>
<td>0.99</td>
</tr>
<tr>
<td>(10.45)</td>
<td>$RC(d_j, P', t) = \sum_{t \in T_i} \sum_{p \in E^p(d_j, t)} 1 - RC(p)$</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence in the patterns matched in the specific document</td>
<td></td>
</tr>
<tr>
<td>(10.42)</td>
<td>$DPRC(D_i, P_i, T_l) = \sum_{d_j \in D_i} dPRC(d_j, P_i, T_l)$</td>
<td>0.99</td>
</tr>
<tr>
<td>(10.44)</td>
<td>$DPRC(D_i, P_i, T_l) = \sum_{d_j \in D_i} dPRC(d_j, P_i, T_l)$</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence in all the patterns matched in all the documents</td>
<td></td>
</tr>
<tr>
<td>RESULT</td>
<td>Overall RC for $d_2$ ($RC(d_2) = \frac{r}{r + w}$)</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Table 33. Calculating the RC of $d_2$
<table>
<thead>
<tr>
<th>K-triple set</th>
<th>Resource Confidence at each stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triple #</td>
<td>Definition</td>
</tr>
<tr>
<td>1</td>
<td>mammal ISA animal</td>
</tr>
<tr>
<td>2</td>
<td>hedgehog ISA mammal</td>
</tr>
<tr>
<td>Corpus/ Set of documents Doc #</td>
<td>Definition</td>
</tr>
<tr>
<td>1</td>
<td>/Corpora/test/Hedgehog.out</td>
</tr>
<tr>
<td>2</td>
<td>/Corpora/test/mammals.out</td>
</tr>
<tr>
<td>Extraction Pattern Set EP #</td>
<td>Definition</td>
</tr>
<tr>
<td>1</td>
<td>&lt;Pp1&gt; such as &lt;Cp&gt;</td>
</tr>
<tr>
<td>2</td>
<td>&lt;Cs1&gt; is a kind of &lt;Ps1&gt;</td>
</tr>
<tr>
<td>3</td>
<td>&lt;Cs&gt; or other &lt;Pp1&gt;</td>
</tr>
<tr>
<td>4</td>
<td>&lt;Cs1&gt; is a type of &lt;Ps1&gt;</td>
</tr>
<tr>
<td>5</td>
<td>&lt;Cp&gt; and other &lt;Pp1&gt;</td>
</tr>
<tr>
<td>6</td>
<td>&lt;Cp&gt; or other &lt;Pp1&gt;</td>
</tr>
<tr>
<td>7</td>
<td>&lt;Cs1&gt;, a type of &lt;Ps1&gt;</td>
</tr>
<tr>
<td>8</td>
<td>&lt;Cs&gt; and other &lt;Pp1&gt;</td>
</tr>
<tr>
<td>9</td>
<td>&lt;Cs1&gt;, a kind of &lt;Ps1&gt;</td>
</tr>
<tr>
<td>10</td>
<td>&lt;Cs&gt; is any of the… &lt;Pp&gt;</td>
</tr>
</tbody>
</table>

Table 34. Resource confidence for the items in the different resource sets, iterations 1-3.
ITERATION 2

As noted in Step 4 above, as a result of the addition of the new document, we now have a new knowledge triple to add to the triple set/ontology i.e. *hedgehog ISA mammal*. We can now set about calculating the RC for this new resource. We now have a new match \( \{d_2, p_1, t_2\} \) because the new triple \( t_2 \) is found in \( d_2 \) by the application of pattern \( p_1 \). In the basic equation 10.46, \( r \) and \( w \) will depend on pattern and document confidence as shown in Equations 10.47 and 10.48.

\[
RC(t, P_i, D_i, T_i) = \frac{r}{r+w} \tag{10.46}
\]

\[
r = tDRC(t, D_i, P_i) + \frac{(TDRC(T_i, D_i, P_i) - tDRC(t, D_i, P_i))}{(TDRC(T_i, D_i, P_i) - tDRC(t, D_i, P_i)) + (TDRC(T_i, D_i, P_i) - tDRC(t, D_i, P_i)) + 1} + tPRC(t, P_i, D_i) \tag{10.47}
\]

\[
w = tDRC(t, D_i, P_i) + \frac{(TDRC(T_i, D_i, P_i) - tDRC(t, D_i, P_i))}{(TDRC(T_i, D_i, P_i) - tDRC(t, D_i, P_i)) + (TDRC(T_i, D_i, P_i) - tDRC(t, D_i, P_i)) + 1} + tPRC(t, P_i, D_i) \tag{10.48}
\]

The equations for obtaining the component parts of Equations (10.47) and (10.48) are shown below:
\[
TDRC(T_i, D_i, P_i) = \sum_{t_j \in T_i} tDRC(t_j, D_i, P_i)
\] (10.49)

\[
tDRC(t_j, D_i, P_i) = \sum_{p \in P_i} \sum_{d \in E'(p, t_j)} RC(d)
\] (10.50)

\[
TDRC(T_i, D_i, P_i) = \sum_{t_j \in T_i} RC(t_j, D_i, P_i)
\] (10.51)

\[
tDRC(t_j, D_i, P_i) = \sum_{p \in P_i} \sum_{d \in E'(p, t_j)} 1 - RC(d)
\] (10.52)

\[
TPRC(T_i, P_i, D_i) = \sum_{t_j \in T_i} tPRC(t_j, P_i, D_i)
\] (10.53)

\[
tPRC(t_j, P_i, D_i) = \sum_{d \in D_i} \sum_{p \in E''(t_j, d)} RC(p)
\] (10.54)

\[
TPRC(T_i, P_i, D_i) = \sum_{t_j \in T_i} tPRC(t_j, P_i, D_i)
\] (10.55)

\[
tPRC(t_j, P_i, D_i) = \sum_{d \in D_i} \sum_{p \in E''(t_j, d)} 1 - RC(p)
\] (10.56)

The values used for these equations are presented in Table 35 with the corresponding results for each stage and the overall result, where \(RC(t) = 0.993\), which is the value shown in the Iteration 2 column in Table 34.

This value makes sense given that the document was chosen specifically because a particular extraction pattern applied, and that pattern was the one with the highest confidence. Note further that we now use the RC for \(d_2\) that we obtained in Iteration 1.
<table>
<thead>
<tr>
<th>Equation No.</th>
<th>Equation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10.54)</td>
<td>( tPRC(t_j, P_i, D_i) )</td>
<td>0.99</td>
</tr>
<tr>
<td>(10.56)</td>
<td>( tPRC(t_j, P_i, D_i) = \frac{1}{</td>
<td>P_i</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence for the matches of patterns for the specific triple.</td>
<td></td>
</tr>
<tr>
<td>(10.53)</td>
<td>( TPRC(T_i, P_i, D_i) = \sum_{t_j \in T_i} tPRC(t_j, P_i, D_i) )</td>
<td>0.99</td>
</tr>
<tr>
<td>(10.55)</td>
<td>( TPRC(T_i, P_i, D_i) = \sum_{t_j \in T_i} tPRC(t_j, P_i, D_i) )</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence for the matches of the patterns for all triples.</td>
<td></td>
</tr>
<tr>
<td>(10.50)</td>
<td>( tDRC(t_j, D_i, P_i) )</td>
<td>0.995</td>
</tr>
<tr>
<td>(10.52)</td>
<td>( tDRC(t_j, D_i, P_i) = \frac{1}{</td>
<td>P_i</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence for the document matches for the specific triple.</td>
<td></td>
</tr>
<tr>
<td>(10.49)</td>
<td>( TDRC(T_i, D_i, P_i) = \sum_{t_j \in T_i} tDRC(t_j, D_i, P_i) )</td>
<td>0.995</td>
</tr>
<tr>
<td>(10.51)</td>
<td>( TDRC(T_i, D_i, P_i) = \sum_{t_j \in T_i} RC(t_j, D_i, P_i) )</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence for the document matches for all the triples.</td>
<td></td>
</tr>
<tr>
<td>RESULT</td>
<td>Overall (( RC(t_2) = \frac{r}{r+w} ))</td>
<td>0.993</td>
</tr>
</tbody>
</table>

Table 35. Calculating the RC of \( t_2 \)
ITERATION 3

In order to complete the examples given in this chapter, we now present the calculation of confidence in the newly identified Extraction Pattern noted in Step 5 above. Given that the RC of an extraction pattern depends on the RC of the documents and triples it operates over, the relevant equations are those presented above (10.20) to (10.32).

The overall result for Extraction Pattern $p_{10}$ is $\text{RC}(p_{10}) = 0.666$, which value is noted in the column for Iteration 3 in Table 34 above. This value is reasonable since it is a newly found Extraction Pattern for which we have no guarantee for its future performance. Note that it has a significantly lower RC than the lowest of our standard set described and analysed in Chapter 7.
### 10.5 A Worked Example

<table>
<thead>
<tr>
<th>Equation No.</th>
<th>Equation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10.22)</td>
<td>$p_{TRC}(p_j, T_i, D_i) = \sum_{d \in D_i} \sum_{t \in E(p_j, d)} RC(t)$</td>
<td>0.993</td>
</tr>
<tr>
<td>(10.24)</td>
<td>$p_{TRC}(p_j, T_i, D_i) = \sum_{d \in D_i} \sum_{t \in E(p_j, d)} 1 - RC(t)$</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence of the triple matched by the <em>specific</em> pattern.</td>
<td></td>
</tr>
<tr>
<td>(10.21)</td>
<td>$p_{TRC}(P_i, T_i, D_i) = \sum_{p_j \in P_i} p_{TRC}(p_j, T_i, D_i)$</td>
<td>1.993</td>
</tr>
<tr>
<td>(10.23)</td>
<td>$p_{TRC}(P_i, T_i, D_i) = \sum_{p_j \in P_i} 1 - p_{TRC}(p_j, T_i, D_i)$</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence of the triple matched by <em>all</em> the patterns.</td>
<td></td>
</tr>
<tr>
<td>(10.25)</td>
<td>$p_{DRC}(p_j, D_i, T_i) = \sum_{t \in T_i} \sum_{d \in E'(p_j, t, D_i)} RC(d)$</td>
<td>1</td>
</tr>
<tr>
<td>(10.26)</td>
<td>$p_{DRC}(p_j, D_i, T_i) = \sum_{t \in T_i} \sum_{d \in E'(p_j, t, D_i)} 1 - RC(d)$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>This sums the confidence of the document matched by the <em>specific</em> pattern.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_i = d_p + d_n = \text{total documents}$</td>
<td>2</td>
</tr>
<tr>
<td>(10.27)</td>
<td>$p_{DRC}(P_i, D_i, T_i) = \sum_{p_j \in P_i} p_{DRC}(p_j, D_i, T_i)$</td>
<td>1.995</td>
</tr>
<tr>
<td>(10.28)</td>
<td>$p_{DRC}(P_i, D_i, T_i) = \sum_{p_j \in P_i} 1 - p_{DRC}(p_j, D_i, T_i)$</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>This sums matches of the document with <em>all</em> the patterns.</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>Overall ($RC(p_{10}) = \frac{r}{r+w}$)</td>
<td>0.666</td>
</tr>
</tbody>
</table>

Table 36. Calculating the RC of $p_{10}$ cf. Table 34
This chapter has presented a different approach to ontology learning from that previously found in the literature\(^{13}\). The Abraxas approach has tried to capture the characteristics of knowledge analysed in previous chapters and summarised in the introduction to this chapter. Above all the approach recognises the dynamic continuously changing nature of human knowledge and tries to model this aspect by allowing each new entity in each set of resources to affect the overall confidence the system has. This confidence (formalised as Resource Confidence) is attached to each item whether it is a knowledge triple (the actual formal knowledge we obtain), or a document (which act as evidence for the knowledge we obtain) or the extraction patterns (which allow us to extract the knowledge from the evidence).

The system is designed to allow varying degrees of confidence in each type of resource. Evidence may mount up that a particular Extraction Pattern is a poor indicator of knowledge. Other evidence may mount that the knowledge extracted from a particular document is has no or insufficient evidence from elsewhere. Evidence can mount that a particular knowledge triple has little or no re-enforcement from other documents. Or quite the contrary, evidence can mount to re-enforce an initial resource item\(^{14}\). Part of the longer term challenge is to determine what exactly should be considered sufficient evidence and how can this be measured.

It is not difficult to see that if RCs are attached to every resource item, then overall Resource Confidence can be calculated for the existing ontology/triple set, or for some specific subset (for example the initial 10% of triples acquired or the seed ontology). And equally analogously for the other resource sets. The KG and EKG measures in contrast are by their nature global measures for the current state of the system. Their role is to determine whether the current state of the system ‘covers’ the data adequately in view of the initial parameters. When the KG and EKG measures reach a certain threshold the system can be deemed to have succeeded and terminate.

In working through the measures in this Chapter, we used the notion of ‘iteration’ and this allows the system to have a sense of diachrony. Thus as time goes on there are more iterations and any given ‘moment’ in the system can be specified by its iteration number. In this mechanism, we see the potential for a diachronically sensitive process of ontology learning (cf. Chapter 9). For example, it may be important the relative timing of different knowledge triples and their corresponding evidence. If there is a burst of evidence at a certain time, and a burst of different evidence at another time then there is a case for a change in the usage of a term or the knowledge associated

\(^{13}\) Systems such as SnowBall [Agichtein and Gravano, 2000] and KnowItAll [Etzioni et al., 2005], which share many similarities with our approach, are not designed to construct ontologies.

\(^{14}\) As the system stands there is no capacity to handle contradictions in texts and resolve them.
with it. The detailed conceptualisation of such an approach and its formalisation will have to be part of future research, however.

There remains one major issue and that is the question of evaluation of the Abraxas approach. The evaluation of ontologies is in themselves a very fraught topic, let alone the evaluation of ontology learning systems. The fundamental problem lies in that we do not have a simple way of deciding if any given piece of knowledge is good or not. The following chapter is devoted to attempting to disentangle these issues.

10.7 Glossary of Terms

**Confidence Value** The confidence value for a given resource item (knowledge triple, extraction pattern, or document) is a measure of confidence in the item derived by using the Resource Confidence Measure.

**Corpus** A corpus is a resource in the Abraxas system, consisting of documents.

**Explicit Knowledge Gap (EKG)** This measure measures the difference between the set of pairs of terms whose ontological relationships are known (because they are in the ontology) and those which need to be added to the ontology (because they are key terms in the corpus) and whose ontological relationship is unknown. Cf. Appendix B.

**Knowledge Gap (KG)** This measure measures the gap between the terms in the corpus and the terms in the ontology. Cf. Appendix B.

**Knowledge Triple/Ontological Triple** A term-relation-term structure extracted from text, such as hedgehog, ISA, mammal, where hedgehog and mammal are terms found in the text, and ISA is an ontological relationship associated with a particular extraction pattern. A set of knowledge triples make up an ontology.

**Lexico-Syntactic Patterns** Lexico-syntactic patterns are sequences of lexical items and/or parts of speech categories which act as templates for the recognition/extraction of knowledge triples from text. They are the most precise form of extraction pattern (cf. Chapter 7).

**Ontological Relationship** The ontological relationship between terms or concepts is the kind of relationship described in ontologies such as is a, part of etc. In this work we have exclusively focussed on is a but the methodology is expected to be generally applicable.

**Ontology** An ontology is a resource in the Abraxas system, consisting of a set of knowledge triples.

**Resource** In Abraxas, there are three main types of resource: the corpus, or set of documents; the ontology, or set of knowledge triples; the Extraction Pattern Set, consisting of different types of extraction patterns including lexico-syntactic patterns.
**RESOURCE CONFIDENCE MEASURE (RCM)** This measure calculates the 'confidence' that the system has in a particular item in a resource set, e.g. in a knowledge triple, in a document, or in an extraction pattern. It is a recursive measure which calculates the value of the resource confidence as a function of the confidence in the complementary resource, i.e. to calculate the RCM of a knowledge triple, we sum the confidence in the document or documents in which it is found together with the sum of the confidence in the extraction pattern or patterns which extracted that knowledge triple.
11 ONTOLOGY EVALUATION AND SELECTION

Evaluation is creation: hear it, you creators! Evaluating is itself the most valuable treasure of all that we value. It is only through evaluation that value exists: and without evaluation the nut of existence would be hollow. Hear it, you creators!

Friedrich Nietzsche [1883/1978]

11.1 INTRODUCTION

The tradition in Natural Language Processing, more than any other sub-discipline of AI, has been for rigorous evaluation of the results of research. Possibly due to the influence of the major evaluation conferences sponsored by DARPA (TREC and MUC), most research in NLP is presented with comparative evaluations. This has not been the case for ontology engineering so far. Although ontology evaluation has a history stretching back to the mid-1990s [Gómez-Pérez, 1994], most research on ontology learning and building does not present evaluations.

There are two ways to understand this absence. On the one hand, AI as a whole has aimed to get things to work. If it works that is good enough. Evidence of this can be seen in the articles of publication such as IEEE Intelligent Systems, which rarely include evaluations. On the other hand, there has been a strong avoidance in computer science in general to undertake experimentation as Tichy [1998] notes. Even though it claims to be a science, much of what is undertaken in computer science is more correctly described as craft. However, enormous progress has been made in areas such as information retrieval and those areas of NLP affected by TREC and MUC because of the obligation to standardise to a given data set and the comparability of the results between researchers.

The difficulty with ontology engineering is far greater than with other areas of AI or NLP. Ontologies are supposed to represent knowledge and knowledge is an amorphous ill-defined substance. It cannot easily be enumerated, catalogued or defined a priori so as to allow some sort of comparison to be made with what the ontology engineer has achieved with their tools. One important distinction must be made and that is between the evaluation of ontologies versus the evaluation of ontology learning systems. There has been a considerable lack of work in both these areas but they must be distinguished. In the former case, one or more ontologies are evaluated by some criteria (cf. below) irrespective of whether they are manually or automatically created. In the latter case, one or more ontology learning systems are evaluated by some
criteria (e.g., a reference ontology). Systems which learn ontologies can be evaluated by a large number of criteria such as productivity gains, precision and recall of terminology covered, etc. In the final analysis, the ontology produced must be evaluated. It is because of this that most of this chapter focuses on the former problem. In Chapter 2, we surveyed the nature of knowledge in general from the perspective of building ontologies. In this chapter, we begin by surveying some relevant ideas concerning knowledge and evaluation before considering specific approaches to evaluating ontologies and making our own proposals.

11.2 Types of Evaluation and the Evaluation of Knowledge

The field of AI and knowledge representation is not concerned with philosophical issues about whether knowledge is allowable or believable. Instead, it is concerned with whether a given representation is valid, good, fitting for the purpose or merely better than an alternative. We now consider a number of distinctions drawing from several disciplines which we believe are relevant to an understanding of how to evaluate ontologies as artefacts of knowledge.

Across a number of disciplines, there is a standard contrast between quantitative and qualitative research methods [Creswell, 2002]. The history of quantitative methods strictly speaking goes back as far as human beings have measured and compared things in order to understand them and make decisions. The early growth of mathematics in Ancient Egypt, Mesopotamia, and Ancient Greece was significantly influenced by the need to measure and keep accounts, for the purposes of obtaining knowledge and making decisions. Quantitative methodologies assume that the object of study can be measured either directly (e.g., the number of apples) or indirectly via some appropriate model. More generally, quantitative methodologies have been associated with the application of pure science methodologies to social sciences, especially for the construction of testable theories which make quantifiable predictions. Thus in social sciences and marketing, the use of questionnaires is a typical methodology which provides raw data to which statistical techniques can be applied and quantitative comparisons and evaluations made. Equally, a great deal of educational testing reflects this paradigm in an attempt to evaluate student performance and provide metrics for comparison (see especially the Educational Testing Service in the US - www.ets.org).

Qualitative research, by contrast, is an approach which has gained recognition only in the last thirty years [Taylor and Bogdan, 1998]. It is characterised by an interest in describing phenomena in detail often in a subjective manner in contrast to the (claimed) objectivity of quantitative methods. Typically it has been used in such fields as ethnography, anthropology, and psychology, and is characterised by interviews and an attempt at in-depth understanding. Even though the approach has received an official label in the last 30-40 years, its origins go back as far as the origins of historical research (such
11.2 Types of Evaluation and the Evaluation of Knowledge

as Herodotus’ Histories, parts of which reflect first hand interviews with the participants in events). Quantitative methodologies include approaches such as narrative inquiry which depends on the subject constructing a narrative of their experience or life-story [Clandinin and Connelly, 2000]. Qualitative approaches permit a greater depth of understanding, a greater concern for the context of an activity or phenomenon, and in general are richer and more detailed. Quantitative methods are appropriate for the testing of hypotheses, for identifying correlations and cause and effect, and the results have more potential to be generalisable.

Significantly, one of the main traditions that ontology engineering derives from - that of expert systems and knowledge bases - has used as its main knowledge acquisition technique protocol analysis which can be seen as a fundamentally qualitative technique combining methods characteristic of ethnography and narrative inquiry. Protocol analysis sought to enter into the mind of the expert sufficiently so as to understand the mental processes and the decision making criteria so as to be able to model these [Ericsson and Simon, 1996, Cooke, 1999].

The only discipline that has a fully developed approach to the evaluation or assessment of knowledge is education. One way of viewing education is to say that its purpose is the communication or transfer of knowledge. Teachers and educators have developed a range of techniques and methods to evaluate and assess the success of their enterprise and there is a whole science of educational testing. The Chinese were the first to evaluate the suitability of their civil servants by public examination (about 11th Century BC) but most modern examination systems derive from the medieval exams in universities. Modern educational theorists have made a number of distinction concerning evaluation which could potentially be applied to ontologies. In a certain sense the analogy which we wish to draw is that like the pupil is assessed for their knowledge by various means so the ontology may be assessed for its knowledge in a number of ways.

Educators make a distinction between formative assessment and summative assessment [Airasian, 2000]. The former refers to assessment that is carried out throughout a course or learning process and is intended to aid the learning (or knowledge acquisition) process, while the latter concerns the an assessment at the end of the course of learning, typically the assignment of a grade (often termed diagnostic assessment). Objective assessments are ones where questions have one correct answer, as opposed to subjective assessments such as essays or open-ended answers which may be answered correctly in a number of ways. Criterion referenced assessments are ones where pupils are assessed in relation to a defined and objective measure, for example in a driving test. In contrast, norm-referenced assessments (otherwise known as ‘grading on the curve’) are assessments which make comparisons with the other pupils in the cohort [Bond, 1996]. Many entrance tests to major US universities are norm-referenced meaning that entry levels can vary
from year to year. Yet another distinction is made between formal assessments which provide a numerical score as opposed to informal assessments which include observation, interview and self-assessment. This later distinction is parallel to the quantitative and qualitative contrast described above.

An important issue with all forms of assessment is validity. If an assessment method is valid then it measures what it is intended to measure. IQ test have often been criticised for not being valid because they depend on too much cultural knowledge rather than intelligence. For an assessment to be acceptable it should both valid and reliable i.e. it should give the same results each time it is applied to the same set of pupils (in the context of education).

Teachers and educators appear to find the process of testing an individual for their level of knowledge a relatively routine process. In contrast, the evaluation of ontologies (as we will show below) is fraught with contrasting approaches and views. At a certain abstract level a pupil is asked to master a body of knowledge (with the help of the teacher) and that knowledge is often embodied in one or more text books. Equally, in this thesis as in much other recent work, a certain body of knowledge is identified in a set of texts and a supposedly corresponding ontology is constructed. However, here the parallel ends because an assessment can be constructed (an exam or test) and the pupil will respond directly to the questions posed. In contrast, an ontology cannot ‘reply’ to a question - it is a passive entity which has to be manipulated by querying mechanisms and reasoners.

One of the best known projects in recent times which lies at the intersection of these issues was Project Halo. This was a project funded by Paul Allen’s Vulcan Venture’s to try to create ‘digital Aristotle’ i.e. a system in which there was encoded a body of knowledge and which could then answer questions. In the pilot project, three research groups were asked to encode into a knowledge base 70 pages of a chemistry textbook used for the Advanced Placement (AP) exam (a popular standardised test in the US). Then 100 questions were posed to each system derived from the AP exam. They included multiple-choice, fill-the-blank and open short essay type questions. For each question the system had to provide an appropriate answer and an account of its reasoning in natural language. A detailed description of
the contest and the performance of the teams is provided in Friedland et al. [2004]. From our perspective, the relevance of this project was the application of educational testing methodologies in evaluating three expert systems. In each case an ontology was used for the purposes of knowledge representation, but obviously as the ontology was only one component of a larger system, the evaluation procedures applied in this case cannot tell us very much about the comparative quality of the different ontologies. This highlights the fundamental difference between evaluating someone by means of a test and evaluating an ontology. The person is a whole ‘system’ with reasoning capabilities, language input/output and other capabilities such as common sense which make any assessment tool such as a test valid (in the sense defined above) only to a certain extent and largely only valid as a comparative instrument with other pupils. The problem which arises when ontologies are integrated into systems will be touched on again below in Section 11.4.2.

More generally this problem raises a number of questions about how we set about evaluating ontologies:

- Can and should knowledge (in the form of ontologies) be evaluated?
- If yes, can knowledge (as ontologies) be evaluated using quantitative methods? And what form should these take?
- If quantitative methods are inappropriate (due to the nature of knowledge in the form of ontologies), then are there qualitative evaluations which can be designed and which can be useful for engineers?

In the following sections, we present survey some of the techniques proposed in the literature as a basis for considering these questions.

11.3 THE EVALUATION OF A REPRESENTATION OF KNOWLEDGE

There are inherent problems in trying to evaluate an ontology as it is not clear what exactly one is trying to evaluate. Going back to Gru- ber’s definition of an ontology, we can note that in spite of the ‘shared’ aspect, an ontology is still extremely subjective, representing the time, place and cultural environment in which it is created. For example, the labels used for concepts are an interpretation over the information available. Thus, when a company’s taxonomy categorises some subject under the heading ‘business opportunity’ this will be entirely dependent on the nature of their business and their interpretation of external events. Business opportunities are not the same for Disney as they are for British Petroleum. Equally, the seemingly simple notion of ‘school’ is differently conceived (for cultural reasons) in the US as opposed to the UK.

---

1 Parts of this section were first published in Brewster et al. [2004, 2005] and reflect work done collaboratively with the respective authors.
Ontology evaluation cannot be compared to the evaluation tasks in Information Retrieval or classic Natural Language Processing tasks such as POS tagging, because the notion of precision and recall cannot be easily used. One would like precision to reflect the amount of knowledge correctly identified (in the ontology) with respect to the whole knowledge available in the ontology. One would like to define recall to reflect the amount of knowledge correctly identified with respect to all the knowledge that it should identify. But we have yet to develop any clear notion of what “correctly identified knowledge” is, or how to determine what “all the knowledge that an ontology should identify” is. Precision and recall are concepts which depend on there being a clear understanding of the set of items concerned. There is a clear set of Parts of Speech (for example), or one can arbitrarily define one, but there is no clear set of “knowledge to be acquired.” It is relatively obvious that the same set of facts can give rise to very different interpretations and therefore different kinds of “knowledge.”

There is a potential contradiction in the concept of evaluating an ontology. Should one treat it as model in line with other scientific models? In such a case, it is the predictive power of the model that needs to be tested. Is the ontology a good model of the knowledge domain so that one can use it to reason and infer (this is the ‘predictive’ aspect in knowledge representation). Or do we treat an ontology as a cultural artifact, much like a dictionary or an encyclopaedia, which has to be evaluated on an entirely different set of criteria. Following Davis et al. 1993 with respect to knowledge representations, we have argued that ontologies are expected to play a multitude of roles (see Section 2.3 and Brewster et al. [2005]). Consequently it is important to bear in mind which of the different roles an ontology is supposedly playing in the context of a given mechanism for evaluation. Here we adapt and re-cap Davis’ categories and consider the implication for evaluation:

An Ontology is a Surrogate. In this case, the fidelity of model to object needs to be measured. The object is our knowledge of the world (not the world itself). The model will inevitably ‘lie’ in the sense that it can never be a perfect model. This means that a great deal depends on the knowledge engineer judging the model as appropriate or ‘valid.’ Much as an educational test must be valid (see above) so an ontology must also be valid in actually representing the aspects of our knowledge which are relevant to the engineering problems at hand. This approach however will never ‘disprove’ an ontology (in the manner of a scientific theory) because there can be degrees of fidelity or validity.

An Ontology is a set of Ontological Commitments. The ontological commitment is the set of concepts and relations chosen to be used in an ontology. So a measure of an ontology would be whether its coverage of concepts and relations in the domain is sufficient. The measure of coverage is simultaneously a measure of fidelity, i.e. if there are significant concepts absent then both the coverage
and fidelity are insufficient. The problem lies in what mechanism is used to determine the ‘significant’ concepts in a domain. For example, in the Halo Project the three contestants had ontologies of enormously different size (only 41 concepts in the case of Ontoprise) [Friedland et al., 2004]. What criteria can be used to determine which approach was right? Or is this undecidable?

An Ontology is a Fragmentary Theory of Intelligent Reasoning. An ontology, especially when combined with a specific representational language such as OWL, presents a particular view of intelligent reasoning. It claims that with this model and these constructs an adequate component of Artificial Intelligence is possible. By determining what kinds of inferences are possible and what are excluded, a specific set of assumptions are expressed. For example, in the Halo Project, chemistry was chosen as the domain because it does not involve graphics or reasoning with uncertainty [Friedland et al., 2004].

An Ontology is a Medium for Efficient Computation. Efficient computation is not an important concept in AI or the Semantic Web in recent years largely due to the perception that computers are so fast this is no longer an issue. Again the results from Project Halo may make one question this. The three teams had very different results in terms of the time taken to complete the task (Ontoprise 9 minutes, SRI 30 minutes, and Cyc Corp. 27 hours) (ibid.). For practical purposes such differences would be immensely significant, but, as noted above, there is no indication in the published material as to whether these differences were due to the nature of the ontologies involved or other components in these systems.

An Ontology is a Medium of Human Expression. Ontologies function as a means for human beings to communicate with machines and vice versa. Here one has to make a careful distinction between the ontology itself (i.e. the concepts and relations) and the language it is expressed in (such as OWL or KIF). For example, one might find OWL rather impenetrable even if the ontology expressed is quite simple. On the other hand, the ontology may highly complex but expressed in a more simple formalism (e.g. semantic nets). An ontology with great complexity may mean that knowledge is represented with fine grained distinctions. From the perspective of evaluation there is no obvious mechanism for measure the complexity or adequacy (by some criterion) of a ‘medium of human expression.’

Taking such a multi-faceted view of ontologies, while true to life, appears to complicate the dimensions on which an ontology could potentially be evaluated and does not make the task any easier. It would appear that most existing efforts in the literature to present methods for the evaluation of ontologies are relatively one dimensional. It is to an account of these that we now turn.
11.4 APPROACHES TO ONTOLOGY EVALUATION

Since the pioneering work of Gómez-Pérez in the early 1990s [Gómez-Pérez, 1995], the literature on ontology evaluation has grown considerably. Its main focus has been on guidelines to be followed or tests to carried out during the construction of an ontology (this could be termed formative evaluation or assessment). Little was done in the early years to evaluate existing ontologies which has become more relevant as the Semantic Web has grown and more and more ontologies are publicly available (see www.swoogle.com, for example). Most of the work described below concerns formal ontologies, expressed in formal ontology languages, and consequently tends to be chiefly concerned with the logical properties of the ontologies under consideration. We believe that ontologies lie on a continuum (see Figure 4) and thus the exclusive focus on formal ontologies is problematic especially now that informal taxonomies and folksonomies are becoming more widespread (in view of the rise of social network software) and thus there needs to be a place for fuzzy informal taxonomies and ontologies.

11.4.1 Structural and Methodological Evaluation

There are two groups of researchers who have pursued an approach to ontology evaluation based on the principles used in its construction. The one has been led by Asuncion Gómez-Pérez of the Universidad Politécnica de Madrid and her focus over a number of years has been on the principles involved in ontology construction. Her approach can be seen as a type of formative evaluation since the emphasis is on continuous evaluation throughout the ontology construction process. A number of criteria for ontology assessment are proposed:

CONSISTENCY. The refers to whether an inconsistent conclusion is derivable from the ontology or not. Inconsistencies may arise within the definition of a concept (‘individual consistency’), for example, if the formal and informal definitions do not coincide. Inconsistencies may arise from errors in the taxonomy e.g. circularity errors, subclass partitions with common classes, etc. [Gómez-Pérez, 2004].

COMPLETENESS Completeness refers to whether all the knowledge that should be included in the ontology is either explicitly stated or can be inferred. Gómez-Pérez notes that completeness cannot be proven and that only the incompleteness of an ontology can be proven (ibid.). This makes completeness not dissimilar from Popper’s scientific method of falsificationism (see above Section 2.2).

CONCISENESS. An ontology is concise if it does not store knowledge which is unnecessary or useless, and if there are no redundancies in the definition of terms or that can be inferred (ibid.). In theory,
it is extremely difficult (if not highly subjective) to determine what may be unnecessary in an ontology - just as subjective as determining what is needed.

**expandability** This concerns the ease with which new definitions and concepts are added to the ontology without re-organising the ontology as a whole. Gómez-Pérez does not provide examples of this. If an ontology reflects a model or theory about a domain then the ease with which new concepts or phenomena are added is a good indicator of the validity of the model. When new concepts cannot be added easily we reach a Kuhnian revolutionary moment (see above 2.2).

The concept of ‘sensitivity’ is also noted but Gómez-Pérez does not explain this sufficiently so as to make it a coherent evaluation criterion [Gómez-Pérez, 1999, 2004].

Some aspects of Gómez-Pérez’s suggestions have turned into actual tools used by ontology developers. There are tools which can check **consistency** (up to a point) such as ReTAX+ [Lam et al., 2004] and the OWL plug-in to Protégé which integrates reasoners [Knublauch et al., 2004, Sirin and Parsia, 2004]. Some research on **conciseness** has been undertaken under the term ‘winnowing’ i.e. identifying which concepts are actually being used in the context of an application [Alani et al., 2005, 2006b]. Alani et al.’s work showed that 59.5% of classes and 51.4% of properties could be removed and still the ontology could be used successfully in the same application. This may indicate the potential for future tools but clearly the approach is dependent on analysing an ontology in the context of an application, unlike the consistency checking tools. With respect to **expandability**, there is no research implementing this in software, to our knowledge.

A different type of formative evaluation is provided by the OntoClean approach to ontologies [Guarino and Welty, 2002]. This approach has its foundations in philosophy, especially the logical tradition of Frege, Russell and Carnap. It proposes a set of “metaproperties” which can be used to describe the classes, properties and relations in an ontology. These properties include:

**rigidity** “A property is rigid if it is essential to all its instances.” Properties are rigid if something cannot be that thing without having that property. Rigid properties include “being a person” for a person. This is in contrast with “being a student” for a person in that a person does not stop being a person if they stop being a student.

**identity** Identity concerns the ability to determine that two things are the same. Thus identity criteria “are conditions used to determine equality (sufficient conditions) and that are entailed by equality (necessary conditions).” The authors argue that using identity criteria enables the ontology engineer to determine if two concepts are one or two. They discuss the example of distinguishing *time interval* as opposed to *time durations*. According
Guarino and Welty, “two time duration of the same length are the same duration ... on the other hand, according to the identity criteria for time intervals, two intervals occurring at the same time are the same, but two intervals occurring at different times, even if they are the same length, are different” [Guarino and Welty, 2002, 62]

**Unity** The notion of unity helps understand the parts of an object and the determining of what is part of a given object and what not, and when an object is whole. The example the authors give is *water* where an instance of water is an amount of water but not (according to Guarino and Welty) a whole object. This is in contrast with *ocean* where an instance such as the Atlantic Ocean does represent a coherent whole.

Guarino and his colleagues present OntoClean as an approach appropriate for the ‘cleaning up’ of existing ontologies. For example, they provide an extensive critique of WordNet [Fellbaum, 1998] pointing to a number of ‘flaws’ [Gangemi et al., 2001, Oltramari et al., 2002]. They identify confusion between concepts and individuals where for example *organisation* includes both types of organisations (*company*, *alliance* etc.) and instances of organisations such as *Red Cross*, *Irish Republican Army*. They propose that what WordNet lacks is an *instance_of* relation and that this is just a lack of expressivity in the WordNet structures. They identify confusion between the object-level and meta-level by which is meant the inclusion of what they call object level concepts (*Set*, *Time*, *Space* and meta-level concepts *Attribute*, *Relation* in the same category *Abstraction*). A further area of ‘error’ is in the violation of the OntoClean principles described above in particular the principle of rigidity. So the authors object to describing a *person* as both *Organism* and also *Causal Agent*. The latter is a role while the former is a type and these have been confused. More generally they identify heterogeneity of levels by which they mean again a confusion of different types of objects as hyponyms of a given category. For example, under *Animal* there exist as hyponyms *Chordate* (a biological class), *Fictional_Animal* (which are not real), as well as *Work_Animal* and *Domestic_Animal*. Oltramari et al. consider the latter two roles as opposed to types.

We can see OntoClean either as a methodology to be applied in the process of building ontologies (as it has been generally promoted) or as a way to evaluate the ‘cleanliness’ of existing ontologies and propose ways to make them more consistent and rigorous (within the principles and objectives that Guarino and his colleagues specify). Clearly, because their approach is dependent on a deep semantic and philosophical understanding of the concepts in the ontology, it is not an approach that can be automated.

There are a number of problems with this approach to ontologies. Even though the authors claim to want to capture the “ontological categories lying behind natural language and human common sense”, and claim to be descriptive in their approach, fundamentally they are prescribing a certain perspective, a certain view of what is real
with respect to language, cognition and the world. Fundamentally
OntoClean applies a logician’s view of the world as having discrete
categories, with necessary and sufficient criteria for membership of
those categories.

The most fundamental critique of OntoClean has come from Yorick
Wilks (2002). In particular, Wilks objects to the OntoClean attempts
to ‘clean up’ or improve WordNet and similar lexical databases. For
example, Guarino et al. object to the multiple placement of a term like
apple under both fruit and food, or country under both location and legal
agent. In their view, this would result in conflicts in identity criteria
(ICs) because the two hyponyms of fruit and country have different
ICs and according to OntoClean’s principles. Wilks argues that the
multiple appearance of items in a structure like WordNet is necessary
in order to allow for the interpretation of certain phrases such as
“Smith tripped and fell on the guinea pig and killed it”. The latter
example makes sense only if we understand a person (‘Smith’) as being
a physical object. Guarino denies a person is a physical object because
“as a physical object, a body has different persistence conditions
from a living being” [Guarino, 1998]. Guarino denies many seemingly
uncontroversial statements including that ordered sets are sets. Wilks
notes that “they take as technical terms, on which to base a theory,
words of a language (English) which will not and cannot bear the
interpretations required” (ibid.). Wilks fundamentally objects to the
use of Identity Criteria, noting that apart from the absurd conclusions
it results in (for example, on this view it follows that Canaries cannot be
birds because they have different ICs), the whole OntoClean approach
ignores a considerable body of philosophical thought. In particular,
the critique of Wittgenstein [1953], Quine [1951] and Putnam [1970]
which have rejected the Aristotelian tradition of defining something
by the features it has. In the Aristotelian tradition for something to be
what it is, it must have a certain set of features which are necessary
properties. Quite apart from the philosophical tradition which Wilks
cites, there is also the empirical research by psychologists [Rosch, 1973,
1975] and linguists [Lakoff, 1986, Taylor, 1989] which has strongly
argues for the same position.

Fundamentally OntoClean applies a set of philosophical principles
to linguistic phenomena. If we consider ontologies as abstract enti-
ties which have a certain Platonic existence then it may be valid to
take such an approach but such a view tends to ignore the reality,
the contextualised use of ontologies to interact with human beings.
There is a mis-representation lying at the heart of the OntoClean
approach. A basic justification for the work is to construct artefacts
i.e. ontologies which are usable in computational systems specifically
the reasoners which have been developed to reason over OWL on-
tologies [Horrocks, 1998, Haarslev and Möller, 2001]. These reasoners
and the corresponding OWL ontologies are based on the principles
of Description Logic and that imposes severe constraints on what is
allowable in the representation format without resulting in a contra-
diction\(^4\). Thus if we look at the OntoClean methodology as an attempt to extract a logically coherent subset of knowledge from entities like WordNet or Mikrokosmos, then their objectives are more comprehensible than any supposed attempt to “clean up” WordNet and other such taxonomies/thesauri. One of the areas where ontologies appear to have great promise is bioinformatics where OWL ontologies are being used to allow interchange and reasoning over disparate databases (cf. BioPAX \url{http://www.biopax.org/} and [Luciano, 2005]) and yet even here, as Wilks notes, there is in fact a continuously changing landscape where if (for example) genes were identified by the identity criteria such as what they encode for, then there is an empirical problem because these expressions are changing all the time. The only thing genuinely constant is the name of a gene which is essentially a linguistic construct. It is vital to remember in this context that a knowledge representation functions also as a form of communication between human beings and computer systems [Davis et al., 1993]. This is usually forgotten in the pursuit of appropriate formalisation.

In conclusion, we would argue that the approach espoused by Guarino and his colleagues is only appropriate for the ‘evaluation’ and/or ‘cleaning up’ of an ontology specifically intended to be used with the DL reasoners currently available. It is certainly not clear that forcing ontologies to fit within those criteria is necessarily going to produce the best, the most appropriate or the most complete representation of a the knowledge of a domain. Furthermore, as Wilks has argued concerning the case of WordNet, these ‘imperfect’ structures do not prohibit extensive practical applications.

11.4.2 Evaluation in Use

Another approach would be to evaluate how effective a particular ontology is in the context of an application. Currently, little work has been done to take a given application environment and test a number of similar but different ontologies in order to evaluate which is most appropriate for the application concerned and determine why that is so. The establishment of a clear set of simple application suites which allow a number of different ontologies to be ‘slotted in’ in order to evaluate the ontologies is an important research step. Another way of achieving the same result would be to set up the detailed tasks presented in TREC or MUC. From the perspective of the machine-readability vision of the Semantic Web, where ontologies are an enabling technologies for interoperability of processes, it may even be entirely inappropriate for humans to read and assess the ontology; only technological effects of the ontology are to be judged [Brewster et al., 2004].

Velardi and Navigli have done relevant work in this field in the context of their OntoLearn system (cf. Section 6.4.3). They presented one of the earliest systems which constructed an ontology and used it

---

4 Personal communication with Aldo Gangemi.
11.4 Approaches to Ontology Evaluation

in a specific application, in this case automated terminology translation [Navigli et al., 2003]. This is a challenging task which has been typically undertaken by using bilingual aligned corpora or else using the Web [Grefenstette, 1999]. Navigli et al. used OntoLearn to extract the technical terminology and associate each component element of complex terms with the appropriate synset in EuroWordNet () and then select the correct corresponding synset in the other language. This task was performed for English-Italian translation of terms concerning tourism. However, although the results were respectable, the system does not actually use a stand alone ontology in order to produce its output. The ontology is generated automatically as part of the process and thus the terminology translation task is not well-suited to comparing one ontology against another, or comparing successive versions of an ontology.

Velardi and her colleagues have extended their approach, largely with the purpose of evaluating the OntoLearn system as an ontology learning system (cf. below Section 11.6). As part of this, they presented a technique for generating natural language definitions [Velardi et al., 2005]. As their approach to ontology construction is entirely centred on linking and extending WordNet, and presenting an ontology engineer with a combination of WordNet synset numbers is not very useful, the automated generation of glosses provides human readable access to some conceptual choices that the ontology generation system has made. Their evaluation of this gloss generation system shows it to be up to 53% fully acceptable or reasonably acceptable for two domain experts in two different domains. This evaluation is very limited but it raises interesting questions. Velardi et al. say that a generated gloss is needed when “a definition of a multiword expression is not found in glossaries or documents, and when attaching a root node of domain concept tree to the appropriate WordNet node” (ibid. p. 103). Thus it appears they believe that if a term has not been recognised sufficiently by the domain so as to be defined, an automatic definition should be generated based on the disambiguation of the component parts. A major question here is to what extent such an approach can capture the subtlety of terminological change as technology and language usage change. The relatively low acceptability figures indicate that this approach is highly problematic even if interesting and thought provoking.

The most relevant work in trying to construct an application-based evaluation methodology has been undertaken by Porzel and Malaka [Porzel and Malaka, 2005]. The application in their scenario is the identification of correct speech recognition hypotheses in dialogue systems, where the correct hypotheses have been identified by hand and act as the gold standard. They use a sophisticated approach to analysing the types of errors, identifying, for example, insertions, deletions and substitutions needed at the level of vocabulary, isa-relation and other semantic relations. This is based on the notion of “error rates” derived from approaches used in speech recognition and dialogue. Table 37 summarises Porzel and Malaka’s classification of
Error Type → insertions deletions substitutions

<table>
<thead>
<tr>
<th>Error Level</th>
<th>Vocabulary</th>
<th>irrelevant concepts</th>
<th>omitted concepts</th>
<th>ambiguous concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vocabulary</td>
<td>isa too coarse</td>
<td>isa too fine</td>
<td>isa too ambiguous</td>
<td></td>
</tr>
<tr>
<td>2. ISA-relations</td>
<td>irrelevant relations</td>
<td>missing relations</td>
<td>indirect relations</td>
<td></td>
</tr>
</tbody>
</table>

Table 37. Ontology Learning Errors (adapted from Porzel and Malaka [2005])

This approach allows a more fine grained understanding of the different dimensions along which an ontology could be evaluated. The authors present a specific application in which (according to them) the results solely depend on the ontological model [Porzel and Malaka, 2005, :115]. This is the ranking of speech recognition hypotheses within a large dialogue system (SMARTKOM - Wahlster 2003) where the rank of each hypothesis is determined by a coherence score derived using the ontology. The ontology they used was one specifically developed for the SMARTKOM project covering both general concepts and more specifically the tourism domain. The coherence scoring algorithm identifies the shortest path between all concepts lexicalised in the speech recognition hypothesis. Because words are ambiguous, the algorithm attempts to identify the cheapest using some scoring mechanism for different types of ontological paths. In order to perform the experiment, the authors narrowed the task further to be correct identification of the ontological relation between two concepts. In comparison with a gold standard using human annotators, they were able to show that the ontology they used was 76% accurate as opposed inter-annotator agreement of 79.5%. Because they were able to break down the errors into substitution, deletion or insertion errors, they were able to make a very detailed analysis of the failings of the ontology they used. Overall their approach is extremely interesting and rigorous, but is also highly complex. In this it is a further indication of the complexity of thorough ontology evaluation.

Since these authors’ papers, we are not aware of further (published) attempts to create an evaluation scenario of a similar type i.e. one where different ontologies or different versions of the same ontology could be compared and lessons drawn from the system performance in that context.

11.4.3 Data-driven Ontology Evaluation

There is a third approach to ontology evaluation which attempts to evaluate the congruence or ‘fit’ between an ontology and a specific domain of knowledge. Such an approach begs the fundamental ques-
tions as to how this ‘fit’ can be measured, and what it is an ontology is ‘fitting.’ Clearly it is impossible to automatically evaluate directly the fit between a knowledge artefact such as an ontology and a person’s knowledge of a domain, let alone the knowledge of a group.

One standard approach would be to compare a new ontology with an existing ‘gold standard’ one. Such has been the approach espoused by authors such as Grefenstette [1994] who used the Roget and Macquarie thesauri in his evaluations. However, the problem here is that one is developing a methodology to imitate a result that could be flawed. If the results differ from the gold standard, it is hard to determine whether that is because the corpus is inappropriate, the methodology is flawed or there is a real difference in the knowledge present in the corpus and the gold standard. In any case, this approach is more applicable when one is trying to evaluate ontology learning methodologies. However, in the context of the Semantic Web, it is likely that one has to choose from a range of existing ontologies the most appropriate for a particular domain, or the most appropriate to adapt to the specific needs of the domain/application.

Elsewhere we have argued that a corpus of texts might be the most effective source of information for the construction of a large proportion of ontologies (Sections 2.4, 5.5.5 above and Brewster et al. 2001). The traditional methods of ontology construction of protocol analysis or introspection are extremely time-consuming, laborious, and expensive. We wish to argue that equally for the evaluation of an ontology (however built) a major part of the evaluation should be to identify the ‘fit’ of the ontology with a domain specific corpus. We purposefully use the word ‘fit’ because there are a number of ways in which this congruence can be assessed. This approach we term data-driven ontology evaluation on the basis that essentially we wish to identify means to compare the ontology and the actual knowledge of a domain. This knowledge in practise will nearly always be present in the texts people generate concerning a domain.

At its most simple, the class labels/concepts in the ontology can be compared to the terms in the corpus. This is just a matter of set comparison. For example, we chose the art and artists domain, for which colleagues had developed the ARTEQUAKT application [Alani et al., 2003]. Using this we collected 41 arbitrary texts from the Internet on a number of artists. The ARTEQUAKT ontology was compared with four others: The Ontology of Science [Ontology of Science, n.d.] was a revised version of the KA2 ontology, the AKT Reference Ontology [AKT, 2003] concerns the academic domain, the CIDOC Conceptual Reference Model (CRM) [CIDOC, 2003] is an ontology representing cultural heritage, and SUMO is the Suggested Upper Merged Ontology [IEEE P1600.1 Standard Upper Ontology Working Group, 2003]. Another approach is to use a vector space representation of the terms in both the corpus and the ontologies under evaluation. This permits an overall measure of the ‘fit’ between the ontologies and the corpus. Thus for example, when comparing the five ontologies mentioned above with our corpus of artist related
texts, we obtained the figures shown in Table 38.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artequakt</td>
<td>0.0518</td>
</tr>
<tr>
<td>CRM</td>
<td>0.0460</td>
</tr>
<tr>
<td>AKT</td>
<td>0.0259</td>
</tr>
<tr>
<td>Science</td>
<td>0.0221</td>
</tr>
<tr>
<td>SUMO</td>
<td>0.0355</td>
</tr>
</tbody>
</table>

Table 38. Vector comparison of five ontologies

This fits with our intuitive and objective understanding of the Artequakt ontology as having the closest fit with the selected corpus. We have developed the idea of measuring the congruence between a corpus and a specific ontology, albeit not using the vector space measure, in Section 10.4.2, where we propose its use in ontology learning.

There is an obvious limitation in this approach. We can imagine two or more ontologies which cover the same domain (in the above sense) but structure the domain entirely differently. Good examples can be found in the ontologies which were used in our ontology ranking section (cf. Section 11.5 below). In some cases a university was defined as an employer, in other cases as an organisation, in others as an educationorganisation, etc. All the ontologies we considered contained the terms student and university but each reflected a different view of the world with a different way of organising very similar categories. To evaluate the suitability of the organisational structure of an ontology is much more difficult because fundamentally there are always a large number of different ways to structure our understanding of the world. It may be that from our data-driven perspective the only thing that can be determined is a range of possible ways of organising terms/concepts. For example, we can find evidence from texts that (for example) a cat is a pet, a cat is a mammal, a cat is a vertebrate but no evidence that a cat is a fish, a cat is a whale, a cat is a piece of furniture.

Such an approach would be in accord with the model of ontology learning described in Chapter 10 with its emphasis on degrees of certainty. On this approach for any given set of terms/concepts, a priori any ontological organisation of terms/concepts is potentially possible. However, for only a subset of these may be there positive evidence, and for another subset there may be negative evidence. And for yet another subset there may be no evidence at all, either negative or positive, which on an open world assumption means they must not be excluded. We believe that this is the direction in which a probabilistic data-driven account of ontology evaluation must proceed, although at this stage we assign this into the category of future work.
11.4.4 Gold Standard Ontology Evaluation

The use of Gold Standards has a considerable history in AI and especially NLP and IR. Classic examples of gold standards are using a predefined tagset when using machine learning in a POS tagger (e.g. the Brill tagger [Brill, 1994]) or the used of predefine human annotator derived ‘correct’ retrieval results in the TREC competitions. There has always been a slight question mark over the use of gold standards because it has not always been uncontroversial what the gold standard is and who determines it. In the educational testing terminology we reviewed above (Section 11.2), the use of a gold standard corresponds to a criterion referenced assessment where the evaluation metric has been specified beforehand.

In the literature relevant to the ontology learning from texts, a number of gold standards have been used or proposed, as well as a number of methodologies for measuring the extent to which the ontology (or output of a given system) corresponds to the chosen gold standard. In the early 1990s, considerable use was made of Roget’s thesaurus as it was available electronically. For example, Grefenstette used Roget for the evaluation his SEXTANT system (described in Section 6.1.1) using the then available 1911 edition from Project Gutenberg [http://www.gutenberg.org/] [Grefenstette, 1993, 1994]. Other researchers have since used the Open Directory (DMOZ) (for example Brank et al. 2006), or have constructed their own ontology to use such as the Tourism Ontology used by successive researchers from Karlsruhe [Cimiano et al., 2005].

There are two fundamental problems with the use of gold standards. One is that as a criterion referenced assessment it is somewhat arbitrary. It is comparable to demanding a student to know a certain textbook (or part of a textbook) off by heart for an exam in e.g. the history of the French Revolution, when the real objective of the assessment is to find out if the student knows about the French Revolution. The student may know about the topic in a variety of ways without knowing the exact wording of the textbook. In educational testing terms, this puts into question the validity of the assessment, and so comparably a gold standard approach to ontology evaluation can be considered of questionable validity. Essentially this depends on whether one is evaluating the closeness of the correspondence between one ontology and another or whether one is evaluating whether or not a given ontology is a reasonable representation of the knowledge of a domain. This is a significant distinction.

The main reason gold standards are used is because they are easy and not too complicated. But excessive reliance on this approach is indicative of the immaturity of ontology evaluation in general.

The other fundamental issue with a gold standard approach is how to measure the difference and similarity between any two ontologies. For example, Maedche and Staab [2002] propose to compare any two ontologies at two levels. One is the lexical level where they compare the lexical labels in the ontologies using the Levenshtein edit distance.
[Levenshtein, 1966] and derive an averaged string match. The other level is a conceptual one (which claims to thereby explicitly disambiguate). They propose the notion of semantic cotopy as a means to express the semantics of a concept. Semantic cotopy is defined as the set of all super- and sub-concepts and is used to calculate the taxonomic overlap between two ontologies. The define a relations overlap measure where again a relation is defined both lexically (by its string label) and conceptually as a pair \((C_1, D_1)\) describing “the concept \(C_1\) that the relation belongs to and its range restriction \(D_1\).” Some experimental evaluation of their measures show great disagreement between people constructing ontologies irrespective of the amount of material being predefined. This raises the issue of how effective it is to compare ontologies in this manner as the lexicalisation of both concepts and relations can vary considerably, while the semantics that Maedche and Staab propose for concepts and relations are both unnecessary and insufficient to establish semantic similarity let alone identity. An extension of this work is presented in Dellschaft and Staab [2006] where the authors change the manner of measuring the taxonomic overlap. Instead of using Semantic Cotopy as the criterion they propose common semantic cotopy which is defined as the set of all super- and sub-concepts common to both ontologies.

Dellschaft and Staab provide a number of measures to compare an automatically generated ontology with a reference or gold standard one. Lexical Precision (LP) and Recall (LR) measure the coverage of terms in the Gold Standard by the generated ontology; Taxonomic Precision (TP) and Recall (TR) aims to identify a set of characteristic features for each term in the output ontology which express the taxonomic position of that term and compare them with those of the corresponding term in the GS; Taxonomic F-measure (TF) is the harmonic mean of TP and TR, while Overall F-measure (TF’) is the harmonic mean of TP, TR, LP and LR. These measures are defined as follows by Dellschaft and Staab (for further details including the formal definition of an ontology please refer to their paper):

\[
LP(O_C, O_R) = \frac{|C_C \cap C_R|}{|C_C|}
\]

(11.1)

where \(O_C\) is the computed ontology and \(O_R\) the reference ontology. Corresponding lexical recall is defined as follows:

\[
LR(O_C, O_R) = \frac{|C_C \cap C_R|}{|C_R|}
\]

(11.2)

The common semantic cotopy, a measure which excludes all concepts not common to both ontologies under comparison, is used to measure taxonomic precision and recall. This is defined as follows: First local taxonomic precision is defined:

\[
tp_{csc}(c_1, c_2, O_C, O_R) = \frac{|csc(c_1, O_C, O_R) \cap csc(c_2, O_C, O_R)|}{|csc(c_1, O_C, O_R)|}
\]

(11.3)
where \(csc(c, O_C, O_R)\) is defined as:

\[
csc(c, O_1, O_2) = \{c_i \mid c_i \in C_1 \cap C_2 \land (c_i <_1 c \lor c <_1 c_i)\}
\]  (11.4)

where \(c \in C\). Given the local taxonomic precision, global taxonomic precision and recall can be defined as follows:

\[
TP_{csc}(O_C, O_R) = \frac{1}{|C_C \cap C_R|} \sum_{c \in C_C \cap C_R} tp_{csc}(c, c, O_C, O_R)
\]  (11.5)

\[
TR_{csc}(O_C, O_R) = TP_{csc}(O_R, O_C)
\]  (11.6)

From these measures they are able to propose a taxonomic F-measure which combines precision and recall, and for the combination of lexical and taxonomic measures in a measure called Taxonomic F’ defined as follows:

\[
TF(O_C, O_R) = \frac{2 \cdot TP(O_C, O_R) \cdot TR(O_C, O_R)}{TP(O_C, O_R) + TR(O_C, O_R)}
\]  (11.7)

\[
TF'(O_C, O_R) = \frac{2 \cdot LR(O_C, O_R) \cdot TF(O_C, O_R)}{LR(O_C, O_R) + TF(O_C, O_R)}
\]  (11.8)

They argue that this provides better results in a set of evaluation experiments. Certainly if one has no choice but to undertake a GS evaluation, then this approach appears the most useful.

Brank et al. [2006] extend Maedche and Staab’s approach to provide a more general account which allows the inclusion of instances as part of the evaluation. In contrast to Maedche and Staab, they do not rely on the natural language descriptors for concepts thus eliminating the string edit distance approach described above: “No assumptions are made regarding the representation of instances, only that we can distinguish one instance from another” (ibid.). Extending the Rand index used in the comparison of clusters [Rand, 1971], they propose an OntoRand index based on measuring the distance between the different clusters (i.e. concepts) containing any two instances and then comparing that distance across the two ontologies. They propose that by using different functions to measure the distance between the placement of different instances, they can provide a family of measures comparing different ontologies. The authors’ attempt to eliminate all reference to the lexical label is interesting and in effect reflects a philosophical position that intension (or meaning) can be determined by extension.

In general, methods for the measurement of the similarity between any given ontology and a gold standard one remain an open issue. This makes the seeming simplicity of a gold standard approach not quite so obvious as it may appear at first sight. In the following section, we discuss ontology ranking which also involves ontology comparison metrics and is closely related.
Ontology ranking is a closely related task to ontology evaluation. It is similar to ontology evaluation in that an attempt to rank ontologies will try to list in some order a set of ontologies from best to worst by some criterion. It differs from ontology evaluation in general in that essentially it provides a single vector of information, while some of the ontology evaluation approaches discussed above can provide more complex evaluations which can (for example) guide the ontology engineer in their subsequent efforts. There are a number of ways of approaching this task depending on what the objective is. It may be that various outputs of an ontology learning system need to be ranked. Or successive versions of the same ontology need to compared and ranked. However, the scenario that is envisaged in the context of the Semantic Web is one where there are many ontologies available for access on the Internet and some ranking needs to be made to choose the best one for a specific task at hand.

One of the major advantages of ontologies on the Semantic Web is the potential for knowledge re-use. In theory, publicly available ontologies are to be reused, modified, extended, and pruned as required, thereby avoiding the huge effort of starting from scratch. A number of ontology libraries currently exist, hosting various ontology files. Examples of such libraries include Ontolingua, the DAML library, the Protégé OWL library, etc. The ontology search facilities provided by these libraries are at best limited to term search, making it difficult for the user to select the relevant ontologies. Furthermore, these more traditional libraries have in effect been superseded by dynamically populated ontology repositories which are in effect search engines specialised in ontology retrieval Sabou et al. [2006]. In particular, there now exists both Swoogle which contains over 10,000 ‘semantic documents’ (i.e. OWL or RDF documents)[Finin et al., 2005], and OntoSelect which contains about 800 ontologies [Buitelaar et al., 2004].

There is clearly a need for more and more ontologies to be constructed and made available partly in the expectation that this will reduce the ‘knowledge acquisition bottleneck’ so long bemoaned in AI. However, as this occurs, so the reuse of this knowledge becomes an ever greater problem. We will term this the knowledge re-use conundrum. In order to find a ‘piece of knowledge’ to re-use in the existing search engines, the user has to represent their request merely by means of one or more search terms. While this is clearly a very poor and approximate description of the knowledge required by a user (let us say a Semantic Web Service agent developer), it is the most usual situation.

It is in this context that in collaboration with colleagues in Sheffield

---

6 Parts of this section were first published in [Alani and Brewster, 2005, 2006, Alani et al., 2006a] and reflects work done collaboratively with Harith Alani. Implementation and architecture were undertaken by Harith alone.
7 http://www-ksl-svc.stanford.edu:5915/
8 http://www.daml.org/ontologies/
9 http://protege.stanford.edu/plugins/owl/owl-library/
and Southampton, we developed a prototype of an ontology ranking system which applies a number of analytic methods to rate each ontology based on how well it represents the given search terms. We present a description of the AKTiveRank system and its metrics in the rest of this section.

### 11.5.1 AKTiveRank

Figure 28 shows the current architecture of AKTiveRank. The main component (no. 2 in the figure) is a Java Servlet that receives an HTTP query from a user or an agent (no. 1). The query contains the terms to search for. Currently it is only possible to search for concepts. In other words, search terms will only be matched with ontology classes, and not with properties or comments.

![AKTiveRank Architecture](image)

Figure 28. AKTiveRank Architecture

When a query is received, AKTiveRank queries Swoogle (no. 3) for the given search terms and retrieves the ontology URIs from the results page returned by Swoogle. Swoogle only allows for keyword-based queries to be made when searching for ontologies. AKTiveRank does not search for ontologies, but only attempts to rank them. Currently, AKTiveRank relies entirely on Swoogle for searching ontologies, and therefore queries submitted to AKTiveRank are meant to resemble Swoogle queries by being based on keywords.

Once a list of ontology candidates is gathered from Swoogle, AKTiveRank starts to check whether those ontologies are already stored in a Jena MySQL database back-end (no. 4), and if not, download them from the web (no. 5) and add them to the database. The Jena API is used here to read the ontologies and handle the database storage. Some of the analysis of ontology structures that AKTiveRank performs...
for the ranking is also undertaken using Jena’s API.

Existing RDF query languages are not well suited for graph queries [Angles and Gutierrez, 2005]. To this end, the current version of AKTiveRank is connected to a purpose-built JUNG servlet (no. 6), which receives an ontology URI and sends back results of JUNG queries in RDF. JUNG (Java Universal Network/Graph framework) is a software library for analysing and visualising network graphs.

AKTiveRank then analyses each of the ontology candidates to determine which is most relevant to the given search terms. This analysis will produce a ranking of the retrieved ontologies, and the results are returned to the user as an OWL file containing the ontology URIs and their total ranks.

11.5.2 The Ranking Measures

AKTiveRank applies four types of assessments (measures) for each ontology to measure the rankings. Each ontology is examined separately. Once those measures are all calculated for an ontology, the resulting values will be merged to produce the total rank for the ontology.

In a previous version of AKTiveRank which was reported in Alani and Brewster [2005], one of the measures applied was the Centrality Measure (CEM). That measure aimed to assess how representative a class is of an ontology based on the observation that the more central a class is in the hierarchy, the more likely it is for it to be well analysed and fully represented [Rosch, 1978]. However, in some experiments we found a few ontologies that placed our concept of interest as a near-top-level concept. Those few ontologies were entirely focused around the concept we were searching for. This meant that even though such ontologies can be highly relevant to our search, they scored very low in CEM. Furthermore, we also found that CEM values corresponded in most cases to the values of the Density measure, and renders CEM somewhat redundant. The Density measure calculates the information content of a class. This observation backs Rosch’s studies [Rosch, 1978] which showed that classes at mid-hierarchical levels tend to have more detail than others.

The new implementation of AKTiveRank also introduces a new measure; the Betweenness measure, and extends the Class Match measure as described in the following sections.

Class Match Measure

The Class Match Measure (CMM) is meant to evaluate the coverage of an ontology for the given search terms. AKTiveRank looks for classes in each ontology that have labels matching a search term either exactly (class label identical to search term) or partially (class label “contains” the search term).

An ontology that contains all search terms will obviously score higher than others, and exact matches are regarded as better than partial matches. For example if searching for “Student” and “Uni-
versity", then an ontology with two classes labelled exactly as the search terms will score higher in this measure than another ontology which contains partially matching classes, e.g. “UniversityBuilding” and “PhDStudent”.

This measure has been extended from its previous version used in Alani and Brewster [2005] by allowing it to take into account the total number of partially matching classes irrespectively of whether an exact match has been found or not. In other words, if we are interested in the concept “student”, then the CMM value for this ontology will be higher the more classes it has with the given word appearing in their labels or URIs.

Definition 1. Let \( C[o] \) be a set of classes in ontology \( o \), and \( T \) is the set of search terms.

\[
E(o, T) = \sum_{c \in C[o]} \sum_{t \in T} I(c, t) \quad (11.9)
\]

\[
I(c, t) = \begin{cases} 
1 & \text{if} \text{ label}(c) = t \\
0 & \text{if} \text{ label}(c) \neq t 
\end{cases} \quad (11.10)
\]

\[
P(o, T) = \sum_{c \in C[o]} \sum_{t \in T} J(c, t) \quad (11.11)
\]

\[
J(c, t) = \begin{cases} 
1 & \text{if} \text{ label}(c) \text{ contains } t \\
0 & \text{if} \text{ label}(c) \text{ not contain } t 
\end{cases} \quad (11.12)
\]

where \( E(o, T) \) and \( P(o, T) \) are the number of classes of ontology \( o \) that have labels that match any of the search terms \( t \) exactly or partially, respectively.

\[
\text{CMM}(o, T) = \alpha E(o, T) + \beta P(o, T) \quad (11.13)
\]

where \( \text{CMM}(o, T) \) is the Class Match Measure for ontology \( o \) with respect to search terms \( T \), \( \alpha \) and \( \beta \) are the exact matching and partial matching weight factors respectively. Exact matching is favoured over partial matching if \( \alpha > \beta \). In the experiments described here, \( \alpha = 0.6 \) & \( \beta = 0.4 \), thus putting more emphasis on exact matching.

**Density Measure**

When searching for a “good” representation of a specific concept, one would expect to find a certain degree of detail in the representation of the knowledge concerning that concept. This may include how well the concept is further specified (the number of subclasses), the number of attributes associated with that concept, number of siblings, etc. All this is taken into account in the Density Measure (DEM). DEM is intended
to approximate the representational-density or information-content of classes and consequently the level of knowledge detail.

Density calculations are currently limited to numbers of relations, subclasses, superclasses, and siblings. We dropped the number of instances from this measure as this might skew the results unfairly towards populated ontologies which may not necessarily reflect the quality of the schema itself.

Definition. Let \( S^c = \langle S^c_1, S^c_2, S^c_3, S^c_4 \rangle = \langle \text{relations}[c], \text{superclasses}[c], \text{subclasses}[c], \text{siblings}[c] \rangle \)

\[
\text{dem}(c) = \sum_{i=1}^{4} w_i |S^c_i| \tag{11.14}
\]

\[
\text{DEM}(o, T) = \frac{1}{n} \sum_{i=1}^{n} \text{dem}(c_i) \tag{11.15}
\]

where \( w_i \) is a weight factor set to a default value of 1, and \( n = E(o, T) + P(o, T) \) which is the number of matched classes in ontology \( o \).

Semantic Similarity Measure

Similarity measures have often been used in information retrieval systems to provide better ranking for query results. Ontologies can be viewed as semantic graphs of concepts and relations, and hence similarity measures can be applied to explore these conceptual graphs. Resnik applied a similarity measure to WordNet to resolve ambiguities [Resnik, 1999]. The measure he used is based on the comparison of shared features, which was first proposed in Tversky [1977]. Another common-feature based similarity is the shortest-path measure, introduced by Rada et al. [1989]. They argue that the more relationships objects have in common, the closer they will be in an ontology. Rada et al. used this measure to help rank biomedical documents which were represented in a semantic knowledge-base. Variations of these techniques have been used to measure similarity between whole ontology structures [Maedche and Staab, 2002, Weinstein and Birmingham, 1999] (cf. above Section 11.4.4).

The Semantic Similarity Measure (SSM) calculates how close the classes that match the search terms are in an ontology. The motivation for this is that ontologies which position concepts further away from each other are less likely to represent the knowledge in a coherent and compact manner\(^{10}\). The SSM formula used here is based on the shortest path measure defined in Rada et al. [1989]. SSM is measured from the minimum number of links that connect a pair of concepts. These links can be isA relationships or object properties.

\(^{10}\) Further studies are required to find whether or not this assumption is dependent on certain ontology properties, such as size or level of detail.
Definition 3. Let $c_i, c_j \in C[o]$, and $c_i \xrightarrow{p} c_j$ is a path $p \in P$ of paths between classes $c_i$ and $c_j$. Let a path from $c_i$ to $c_j$ be defined as an alternating sequence of vertices and edges, beginning with $c_i$ and ending with $c_j$, such that each edge connects its preceding with its succeeding vertex.

\[
ssm(c_i, c_j) = \begin{cases} 
\frac{1}{\text{length}(\min_{p \in P(c_i \xrightarrow{p} c_j)})} & : \text{if } i \neq j \\
0 & : \text{if } i = j 
\end{cases} \quad (11.16)
\]

\[
SSM(o, T) = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ssm(c_i, c_j) 
\]

\[
n = E(o, T) + P(o, T) \text{ which is the number of matched classes in ontology } o.
\]

**Betweenness Measure**

One of the algorithms that JUNG provides is Betweenness [Freeman, 1977]. This algorithm calculates the number of shortest paths that pass through each node in the graph. Nodes that occur on many shortest paths between other nodes have higher betweenness value than others. The assumption is that if a class has a high betweenness value in an ontology then this class is central to that ontology.

The BEtweenness Measure (BEM) calculates the betweenness value of each queried concept in the given ontologies. Ontologies where those classes are more central will receive a higher score. In the following definition, an ontology is assumed to be a connected graph\(^{11}\).

Definition 4. Let $c_i, c_j \in \text{classes}[o]$, $c_i$ and $c_j$ are any two classes in ontology $o$, $C[o]$ is the set of class in ontology $o$. Paths are defined as in Definition 3. $bem(c)$ is the BEtweenness Measure for class $c$.

\[
bem(c) = \sum_{c_i \neq c_j \neq c \in C[o]} \frac{\sigma_{c_i c_j}(c)}{\sigma_{c_i c_j}} 
\]

where $\sigma_{c_i c_j}$ is the length of the shortest path from $c_i$ to $c_j$, and $\sigma_{c_i c_j}(c)$ is the number of maximally short paths from $c_i$ to $c_j$ that pass through $c$.

\[
BEM(o, T) = \frac{1}{n} \sum_{k=1}^{n} bem(c_k) 
\]

where $n = E(o, T) + P(o, T)$ is the number of matched classes in ontology $o$, and $BEM(o)$ is the average Betweenness value for ontology $o$.

**Total Score**

The total score of an ontology can be calculated once the four measures are applied to all the ontologies that the search engine returned. Total

\(^{11}\) The definition here closely follows those presented in [Brandes, 2001]
Ontology URL

- a http://www.csd.abdn.ac.uk/~cmckenzi/playpen/rdf/akt_ontology_LITE.owl
- b http://protege.stanford.edu/plugins/owl/owl-library/koala.owl
- c http://protege.stanford.edu/plugins/owl/owl-library/ka.owl
- d http://reliant.teknowledge.com/DAML/Mid-level-ontology.owl
- f http://www.mondeca.com/owl/moses/univ2.owl
- g http://www.mondeca.com/owl/moses/univ2
- h http://www.lehigh.edu/~yug2/Research/SemanticWeb/LUBM/Universityo_o.owl
- i http://www.lehigh.edu/~yug2/Research/SemanticWeb/LUBM/Universityo_o.owl
- j http://www.mondeca.com/owl/moses/ita.owl
- k http://www.mondeca.com/owl/moses/ita.owl
- l http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- m http://www.mondeca.com/owl/moses/ita.owl
- n http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- o http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- p http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- q http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- r http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- s http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- t http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- u http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- v http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- w http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- x http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- y http://www.ml.uni-trier.de/ontology/LUBM_00.owl
- z http://www.ml.uni-trier.de/ontology/LUBM_00.owl

Table 39. Order of search result for “student university” as returned by Swoogle. Duplicates were removed and are indicated by a dash.

The score is calculated by aggregating all the measures’ values, taking into account the weight of each measure, which can be used to determine the relative importance of each measure for ranking.

Definition 5. Let \( M = \{M_1, \ldots, M_4\} \) = \{CMM, BEM, SSM, DEM\}, \( w_i \) is a weight factor (defined below), \( o_j \) ranges over \( O \) which is the set of ontologies to rank (\( 1 \leq j \leq |O| \)), and \( T \) is the set of search terms.

\[
\text{Score}(o_j \in O, T) = \sum_{i=1}^{4} w_i \frac{M_i(o_j, T)}{\max(M_i(O, T))}
\] (11.20)

Values of each measure are normalised to be in the range (0–1) by dividing by the maximum measure value for all ontologies. The first rank will be given to the ontology with the highest overall score, the second rank to the second highest score, and so on.

11.5.3 Experiment

In this section we report the results of running AKTiveRank over an example query submitted to Swoogle\(^{12}\).

The weights for calculating total score (equation 11.20) for our experiment are set to 0.4, 0.3, 0.2, 0.1 for the CMM, BEM, SSM, DEM measures respectively. Further experiments will be required to identify the best mix of weights to reach optimum ranking results.

\(^{12}\) Note that we used Swoogle 2005 in our experiment. The results of the search may thus be somewhat different if using the latest version of Swoogle which became available near the completion of our experiment.
Now let’s assume that we need to find an OWL ontology that represents the concepts of “University” and “Student”. The list of ontologies returned by Swoogle at the time of the experiment as a result of the query “university student type:owl” is shown in table 39. Some of the ontologies returned by Swoogle were duplicates (i.e. the same ontology was available under two slightly different URLs). As expected, the same rank was produced by AKTiveRank for all duplicated ontologies, and therefore were removed from the table to save space. It is worth mentioning that Swoogle returned those duplicated ontologies in very different orders. For example the exact “koala” ontology was returned under 3 URLs in the 2nd, 9th, and 18th position.

Some ontologies were no longer online and hence were dropped from the ranking experiment (they are given index “–” in the table). When AKTiveRank was applied to the resulting list shown in table 39, it produced the values given in table 40, which are displayed in figure 29.

<table>
<thead>
<tr>
<th>Onto</th>
<th>CMM</th>
<th>DEM</th>
<th>SSM</th>
<th>BEM</th>
<th>Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.833</td>
<td>0.632</td>
<td>0.250</td>
<td>0.806</td>
<td>0.688</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>0.5</td>
<td>0.197</td>
<td>0</td>
<td>0</td>
<td>0.220</td>
<td>12</td>
</tr>
<tr>
<td>c</td>
<td>0.667</td>
<td>0.5</td>
<td>0.25</td>
<td>1</td>
<td>0.667</td>
<td>2</td>
</tr>
<tr>
<td>d</td>
<td>0.417</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.267</td>
<td>11</td>
</tr>
<tr>
<td>e</td>
<td>1</td>
<td>0.632</td>
<td>0.111</td>
<td>0.452</td>
<td>0.621</td>
<td>3</td>
</tr>
<tr>
<td>f</td>
<td>0.833</td>
<td>0.579</td>
<td>0</td>
<td>0</td>
<td>0.391</td>
<td>7.5</td>
</tr>
<tr>
<td>g</td>
<td>0.833</td>
<td>0.579</td>
<td>0.167</td>
<td>0.065</td>
<td>0.444</td>
<td>6</td>
</tr>
<tr>
<td>h</td>
<td>0.5</td>
<td>0.553</td>
<td>1</td>
<td>0.323</td>
<td>0.552</td>
<td>4</td>
</tr>
<tr>
<td>i</td>
<td>0.5</td>
<td>0.579</td>
<td>0.167</td>
<td>0</td>
<td>0.291</td>
<td>10</td>
</tr>
<tr>
<td>j</td>
<td>0.5</td>
<td>0.579</td>
<td>0.125</td>
<td>0.839</td>
<td>0.535</td>
<td>5</td>
</tr>
<tr>
<td>k</td>
<td>0.667</td>
<td>0.579</td>
<td>0</td>
<td>0.097</td>
<td>0.354</td>
<td>9</td>
</tr>
<tr>
<td>l</td>
<td>0.667</td>
<td>0.685</td>
<td>0</td>
<td>0.194</td>
<td>0.391</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Table 40. AKTiveRank results

Analysis of Results

From the results of this experiment, it can be seen that ontology a scored the highest value in AKTiveRank. The ontologies c and h where given the second and third rank respectively. The koala ontology, which was placed second in Swoogle’s results list, got the least AKTiveRank

---

13 http://protege.stanford.edu/plugins/owl/owl-library/koala.owl
   Protege/koala.owl
15 http://www.iro.umontreal.ca/ touzanim/Ontology/
   koala.owl
score, and thus was placed last in the ranked list. Even though this ontology contains classes labelled “Student” and “University”, but those classes are not closely associated (i.e. zero SSM\(^\text{16}\)) and not graphically central to the ontology structure (i.e. zero BEM). The *koala* ontology is not exactly about students or universities, and therefore deserves the last rank in this context.

Note that 5 of our ontologies received a SSM of 0.0. This indicates that AKTiveRank did not manage to find any paths connecting the two given queried classes. Semantic paths that cross via the imaginary *owl:Thing* class are ignored.

The ontology that scored the highest in the Class Match measure (CMM, section 11.5.2) was ontology *e*. This ontology had 2 classes with labels exactly matching our search terms, and 3 partially matching ones; *Phd-Student*, *University-Faculty* and *Distance-Teaching-University*.

The highest DEM value was calculated for ontology *d*. This ontology had a total of 5 subclasses and 10 siblings for the two classes matching our search terms. This added to its DEM value and made this ontology score best on this measure.

Ontology *h* received the maximum SSM value because it has the relation *enrolled_at* which directly connects the classes “Student” and “University”.

And finally, ontology *c* was found to have the highest average betweenness value for the two classes in question, which indicates that these classes are more structurally central in this ontology than in the other ontologies.

Ranking based on each measure separately is displayed in figures 30, 31, 32 and 33. When taken separately, none of the measures provided

---

16 Jena disagrees with Proégé OWL on its rendering of a restriction in the Koala ontology between the classes Student and University.
Figure 30. Ranks based on CMM.

Figure 31. Ranks based on DEM

Figure 32. Ranks based on SSM

the same ranking list as when the measures were combined.
11.5.4 **Ontology Ranking: Conclusions**

Even though the initial results are interesting, a great deal of research is required before making any conclusive remarks about AKTiveRank’s measures.

This work on the ranking of ontologies has been partly motivated by an awareness that ontologies are not artefacts like any other document on the web. They are crafted usually with considerable care where (for example) the importation of other ontologies usually has a certain significance. On the other hand, it is usual when constructing a domain specific ontology to import general ontologies like foaf which contain relatively little domain specific content. It is important to distinguish the function of an ontology from that of a web page. A web page is read by a human being and any links it may have may or may not be followed by the reader. In contrast, an ontology is designed to be read by a machine and any links it may have are by definition imports pointing to other ontologies which must be included. This poses a dilemma in ranking an ontology as to whether to include all imports or not. Because the imports tend to be high level general ontologies, they are relatively vacuous if the user is seeking a domain specific ontology. Further more if ontology $O_1$ is dependent on ontology $O_2$ to represent class $c$, then $O_2$ will be evaluated separately anyway assuming it is included in the set retrieved.

It is very difficult to pinpoint the right selection of parameters or structural properties to investigate when ranking ontologies. The selection can be dependent on personal preference as well as use requirements (i.e. the purpose for which the ontology is intended). One focus of future research should be to perform a user evaluation to include a significant number of human participants and a significant number of queries. Queries need to be posed to the system over a sufficient range of topics so as to allow confidence in the ranking methods we have used. Provisional experiments have shown it is difficult to present ontologies effectively to evaluators. Screen shots
often show only a partial picture of the whole ontology, and some individuals prefer to examine the native OWL in understanding the ontology and making judgements. This is highly dependent on the background and skills of the user. Users must be given the freedom to browse the ontologies in an ontology editing tool such as Protégé [Noy et al., 2001] or Swoop [Kalyanpur et al., 2006], rather than given screen dumps or schema descriptions. For this reason, extensive user-based experiments are required to at least find out what are the properties that users tend to look at when judging the general quality or suitability of an ontology. Unlike ordinary search engines, where the user can be safely assumed to be relatively naive, with ontologies the typical user is either a knowledge engineer or software developer who has preconceptions of a technical nature.

Other parameters can be taken into account, such as whether a class is defined or primitive (currently indirectly covered by the Density measure), or if the classes of interest are hubs or authoritative in a graph-network sense, which might increase their ontology’s ranking.

The most appropriate criteria for searching for ontologies are still unclear. Swoogle is mainly based on keyword search, but other searching techniques can be imagined, based for example on the structure of ontologies or based on whether the ontologies meet certain requirements [Lozano-Tello and Gómez-Pérez, 2004]. However, whatever the search mechanism is, there will always be a need for ranking. The ranking criteria will obviously have to be designed to fit the chosen search technique.

11.6 EVALUATION OF ONTOLOGY LEARNING SYSTEMS

The discussion in this chapter concerning evaluation has been motivated partly by the need to evaluate ontologies per se but also to achieve a greater understanding of the task of evaluating ontology learning systems. There is an inherent conundrum in trying to evaluate an ontology learning system. If the system is evaluated in terms of a gold standard, then all we know is that the system is able to replicate that gold standard. If the system generates an ontology significantly different from the gold standard, it is impossible to prove that this alternative ontology is not better or just as good as the other (gold standard) one. This is the nature of knowledge acquisition. If the knowledge of a domain is previously unknown, we are unable to judge whether it correct in an objective sense.

One way round this problem is to evaluate separately the different steps in an ontology learning system. There is a long and successful tradition in NLP to evaluate separate component elements of a complex systems. Such is the case in the evaluation of POS taggers, Named Entity Recognisers and Parsers, all of which may form components in a Machine Translation, Question Answering or Information Extraction systems. The only serious attempt to undertake such an approach in the context of ontology learning, has been the work of Velardi and Navigli in their OntoLearn system (discussed earlier in this chapter in
In their paper on evaluating Ontolearn, Velardi et al. [2005] note that the system is based on five main algorithms (ibid. p.98):

1. extraction of terms
2. extraction of natural language definitions
3. parsing of natural language definitions
4. semantic disambiguation (to identify correct sense of term components ...)
5. identification of semantic relations between term components

They argue that the overall quality of an ontology is a function of the quality of these steps. It is important to bear in mind that the OntoLearn system is conceived to work closely with domain experts and essentially speed up their work. So while steps 1-3 can be evaluated by human evaluators directly, steps 4 and 5 depend for their evaluation on an understanding of sense inventories, taxonomic organisation and formal specifications.

With respect to terminology extraction, their evaluation resulted in 80-85% precision and about 55% recall. This figure was obtained by “manually identifying truly relevant terms from a list of syntactically plausible multiword expressions, automatically extracted from the available documents” (ibid, p.99). Extraction of Definitions is achieved by using a set of lexico-syntactic patterns and they claim to be able to obtain definitions for about 80% of terms. In view of the purpose of their system being to help ontology engineers construct an ontology, they also quote the increase in speed achieved in the construction of definitions for terms in the ontology (down from 16 minutes per term to 7.5 minutes). However, the timing figures were not obtained very rigourously and represent estimates. Parsing of definitions to obtain the hyperonym achieves precision 0.94 - 0.97 and recall of 0.91 - 0.95. The task is essentially to identify the main NP of a definition and then extract that as the hyperonym. They note that on the one hand the task is made easier because only a very restricted subset of sentences are accepted as input (the definitions found on the web) and on the other hand even very vague hyperonyms are acceptable. They give the example of extraction the hyperonym object form the definition of component: “and object adhering to a component architecture”. The uninformative nature of the term object is unimportant because the system’s output is intended to be revised by human specialists. The Semantic Disambiguation task involves the assignment of the component terms of multi-word terms to the appropriate synset in WordNet. Over a number of evaluations [Navigli et al., 2003, Navigli and Velardi, 2004], they claim precision of 82-86% and recall 60-70%. The final type of evaluation they have attempted is the annotation with semantic relations which we described above in Section 11.4.2.

Velardi and Navigli’s approach is very insightful in their deconstruction of one approach to ontology learning and its evaluation. However, as an approach it does not tell us whether their methodology is a good one as a whole i.e. it is only really useful if one accepts their
architectural assumptions for the components of an ontology learning system and compares comparable component x with component y. Thus it is not an approach which recommends itself if one wishes to make overall comparisons of different ontology learning systems as a whole. Furthermore, there is no way in their approach to determine whether the output of their system is a ‘good’ ontology or not by some criterion.

It is unfortunately the case that there exists no other work (to the date of writing) which concerns itself with ontology learning systems evaluation. This is clearly a significant area for future development.

11.7 EVALUATING ABRAXAS

Abraxas is both a running system under development and concurrently a model of how a set of possible approaches or implementations of ontology learning systems could be undertaken. Consequently we will present two different kinds of evaluation in this section. First we consider how well this model fits in with both our requirements (as expressed for example in Chapter 5) and the observations we have made concerning the nature of knowledge and text in a number of places subsequently. Secondly, we present results from an implementation of Abraxas using a Gold Standard evaluation.

11.7.1 Abraxas as a Metamodel

Of a number of features we identified in Chapter 5, the most difficult to quantify or determine was the first, coherence. We defined this as a “a coherent, common sense organisation of concepts or terms” (Section 5.5.1) and clearly this is both highly subjective and context dependent. In a certain sense, the ontology ranking measures presented above (11.5) were an attempt indirectly to quantify or measure such a dimension. The success of this approach is disputable. However, the insistence of the Abraxas approach on evidence, primarily corpus based evidence, means that there is a strong likelihood that both the individual knowledge triples and the overall ontologies that result from this will reflect the convictions and assumptions of the social community represented by the corpus. Just as the use of corpus evidence has by common consent vastly improved the quality modern dictionaries, we believe that we would observe the same effects in the construction of ontologies. This is in contrast with such methodologies as protocol analysis etc. which involve arbitrary choices on the part of the ontology engineer concerning both the source of knowledge and its interpretation. It has been observed repeatedly how ontologies covering more or less the same domain can be so arbitrarily different. By establishing a more coherent and systematic corpus based methodology, given the same inputs an approximately similar ontology should emerge covering that domain.

The acceptability of placing the same term in multiple positions
or multiplicity, as we termed it, was another constraint we identified. We made this requirement because many existing approaches made exclusive decisions for a given term and its parent. Clearly by using textual evidence, there may well be contradictory evidence that term $< T_i \text{ ISA } T_\alpha >$ and also $< T_i \text{ ISA } T_\beta >$ and the system needs to be able to handle both cases. The collection of evidence for each knowledge triple in no way precludes the possibility of both being valid knowledge triples. A cat is a mammal and is a pet - this is the natural way human being organise their conceptual space. Furthermore, as the Abraxas models become more sophisticated, we expect that it will be possible to examine the evidence for each knowledge triple (much like examining the citations in lexicography) and thus determine what the differentiating context is.

The method and approach presented in Chapter 10 by definition will provide single labels which we also defined as a further requirement. Of course, there could algorithms which would respect other aspects of the approach, and yet not generate single labels. However, the emphasis on textual evidence and the use of knowledge triples extracted from text using IE methodologies makes the Abraxas approach almost inevitably result in unique labels for concepts. Furthermore, there is good linguistic evidence to treat each term separately as a unique concept and to avoid the conflation which ontology editors have been prone to.

A further requirement was the use of corpora as the main source of data and that any system should be able to use existing ontologies. The Abraxas approach clearly focusses on using corpora as a primary data source whether in the form of a seed corpus, an institutional repository or the Web. As we noted in Chapter 10, the Corpus set can be initially empty, partially filled with a seed corpus, or fully filled with a specified domain corpus. In the empty case, documents will be added as needed by the system. Equally the Ontology or knowledge triple set can be empty, partially of fully filled. Thus it is possible to start with a seed ontology or an existing ontology that needs to be adapted or augmented. This enables the Abraxas approach to be fully flexible and adapt to the needs of the user.

The one criterion for which we cannot give the Abraxas approach such a clean bill of health is in the ease of computation. The need for considerable computations and multiple iterations, the continuous updating of the system, all indicate a considerable computational complexity. This is offset by the incremental iterative design of the overall approach, but the use of recursive equations for the calculation of some of the measures mean that the approach requires considerable computational power and/or time.

From a completely different perspective, we need to consider how well the Abraxas approach is adapted to the nature of knowledge and text. It as been our repeated claim in this thesis that a successful approach to ontology learning must take into account more effectively (than has been done until now) the nature of human knowledge and its relationship to and expression in texts. At the beginning of Chapter
we summarised some of our conclusions about knowledge and texts (repeated here as in Table 41).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Knowledge is not monolithic, monotonic or universally agreed. It is uncertain, revisable, contradictory and differs from person to person.</td>
</tr>
<tr>
<td>2</td>
<td>Knowledge changes continuously over time and so what is believed to be true today will be revised and re-interpreted tomorrow. Changes can be more or less rapid due to social and technological factors.</td>
</tr>
<tr>
<td>3</td>
<td>Only a small part of what is generally considered knowledge can be represented in an ontology.</td>
</tr>
<tr>
<td>4</td>
<td>Ontologies are useful if incomplete models of domains.</td>
</tr>
<tr>
<td>5</td>
<td>Texts assume the reader has a certain amount of background knowledge. The great majority of what we would expect to be in an ontology is exactly in this assumed background knowledge.</td>
</tr>
<tr>
<td>6</td>
<td>While it is easy to establish that some relationship exists between two terms, explicit defining contexts are relatively rare in texts due to data sparsity.</td>
</tr>
<tr>
<td>7</td>
<td>Data sparsity necessitates the use of multiple sources of information concerning the ontological relationship between two terms, and furthermore the recognition that in certain cases textual evidence is too weak to say anything more than that there is some relationship.</td>
</tr>
</tbody>
</table>

Table 41. Key findings concerning knowledge and text.

The iterative approach of Abraxas, allowing full flexibility of a range of inputs takes into account of Item 1. The use of textual evidence to construct accounts based on degrees of certainty for each item of knowledge (knowledge triple) allows knowledge to be revisable, contradictory and differ from domain to domain. As we have noted before, the iterative process allows the diachronic process of knowledge change to be allowed for. If knowledge changes, this will be reflected in the evidence to the system i.e. further documents added to the corpus (Item 2).

Items 3 and 4 do not directly impact the design of the Abraxas model, although they should inform the users’ sense of the fallibility of all knowledge representations in general. A core aspect of the relationship between text and knowledge is noted in Item 5 and this has been a key design aspect in the Abraxas approach. Abraxas is designed to identify which items of knowledge are absent (given the data from the corpus) and search for a text which will provide that knowledge in an explicit form. By allowing for the fact that there are differing degrees of evidence for a specific knowledge triple, Abraxas allows for partial evidence (default level of some association) gradually building up to full explicit evidence. This allows it to accommodate Item 6 which acknowledges the realities of textual evidence and problems of data sparsity. In many cases, explicit contexts may never be found for a potential knowledge triple. Finally, although the use of textual evidence is fundamental to the Abraxas approach, it is not to be seen as an exclusive source. The use of multiple heterogeneous sources or is feasible and may in some cases be the most effective way to augment the knowledge of the system.
Abraxas is an approach designed to allow for the absence of evidence. Whether it is due to data sparsity, or the application of insufficient iterations for the system to achieve a high degree of reliability, Abraxas is able to attach to each knowledge triple (and in fact to each resource item) a degree of confidence. This means insufficient evidence can be highlighted if for example the user wishes to intervene to settle those issues. This also means that weak knowledge models can be generated from poor or insufficient data without the system failing. Our proposal of a weak or ‘default’ ontological relationship allow the system not to fall over in such situations.

11.7.2 Abraxas as a Running System

An implementation of one version of the approach described in Chapter 10 was undertaken in Java. As noted in Chapter 10, there are a large number of different potential parameters in the abstract model presented there. Each resource in the system could be empty, partially specified and thus extendable, or fully specified and thus not extended further (Section 10.2). In particular, we presented a table showing 27 different possible starting states (Table 25). In this partial evaluation of the Abraxas system, we evaluated Starting State 22 and 23 with a fully specified set of patterns, and partial or empty set of documents.

In order to evaluate the system, we chose to use a Gold Standard and thus we created a domain-specific hand-crafted ontology reflecting common sense knowledge about animals, containing 186 concepts up to 3 relations deep (included in Appendix C). In order to compare the GS ontology with the computer generated one, we chose to follow the methodology proposed by Dellschraft and Staab [2006] described above in Section 11.4.4, where we have cited the definitions for the metrics.

As a seed corpus we used a set of 40 texts from Wikipedia all entries concerning animals which were included in the GS ontology. All were entries for commonly known animals such as hedgehog, lion, kangaroo, ostrich, lizard, amounting to a little over 8,000 words. Note there is a substantial gap between the number of animals initially covered in the articles and the number present in the GS ontology. The articles were pre-processed to remove the markup present in the originals. The system added further texts from the web, in accordance with the process described in Chapter 10. The Yahoo API was used to issue queries, and the top 20 texts identified would be processed to calculate their RC scores and add the top scoring to the current corpus.

A series of experiments were conducted, each time varying the seed knowledge input to the Abraxas system. This was varied by the presence or absence of a starting set of texts, and the size of the starting set of knowledge triples. In all cases we used as a stopping

---

17 A complete evaluation of the Abraxas approach would need to evaluate all the other possible starting states or show why those were incoherent initial states cf. Future Work in Chap 12.
18 Publicly available also from http://nlp.shef.ac.uk/abraxas/
NP(pl) such as NP*  NP(sg) is a kind of NP(sg)
NP(sg) or other NP(pl)  NP(sg) is a type of NP(sg)
NP(pl) and other NP(pl)  NP(pl) or other NP(pl)

Table 42. Extraction patterns used: NP = noun phrase, sg = singular, pl = plural.

criterion the Explicit Knowledge Gap (EKG) measure described in Section 10.4.2.

We used the 6 extraction patterns, shown in Table 42, which our research presented in Section 7.3 had shown to have good precision. We intended to isolate the ontology learning process from the influence of pattern learning in these experiments, making results more comparable with those of the literature. For the same reasons, the system was tested in a completely unsupervised manner.

The parameters varied were a) whether the texts (described above) were included as part of the seed, and b) the number of seed RDF triples. When more than one triple was included in the seed set then it was a random selection from the GS. The following list shows the different cases:

Experiment 1 Corpus: 40 Wikipedia texts; Ontology: {dog ISA animal};

Experiment 2 Corpus: 0 texts; Ontology: {dog ISA animal};

Experiment 3 Corpus: 0 texts; Ontology: {worm ISA animal, bird ISA animal, reptile ISA animal, rat ISA mammal, dog ISA mammal};

Experiment 4 Corpus: 40 Wikipedia texts; Ontology: {worm ISA animal, bird ISA animal, reptile ISA animal, rat ISA mammal, dog ISA mammal};

Our initial experiment was with Case 1, running over approximately 500 iterations. The final results are shown in Table 43. Both the TF and TF’ obtained are significantly better than equivalent results in the literature, which often achieve maximum scores around [0.3] for both precision and recall [Cimiano et al., 2005].

<table>
<thead>
<tr>
<th></th>
<th>LP</th>
<th>LR</th>
<th>TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>0.40</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>0.70</td>
<td>0.60</td>
<td></td>
</tr>
</tbody>
</table>

Table 43. Results obtained for experiment 1.
Learning curves  Figure 34 shows how the results vary over the number of iterations. We can see here that LR steadily increases reflecting the growing size of the ontology and correspondingly its overlap with the GS. In contrast, LP is in constant flux but with a tendency to decrease. TP varies between set limits of [1.0 - 0.84] indicating that concepts are generally inserted correctly into the hierarchy. TR is also a measure in considerable flux and manual analysis of the different output ontologies show that sudden insertion of parent nodes (e.g. mammal at iteration 9) make a substantial difference which gradually stabilises over further iterations. Over long numbers of iterations, this flux in TR seems to become less likely. We also observe a steady increase $T'F$ in parallel with the increase in LR indicating that the system is doing better as it increases its coverage of the lexical layer of the GS ontology.

Variation of initial seeds  In order to evaluate the effect of varying the seed data, we then ran the same experiments with the three other cases listed above. The results are shown in Figure 35. By comparing case 2 with 3, we see that the number of triples used to start with has little impact over the final results probably because in this case all the triples were from quite a narrow domain. By comparing case 1 with 2, and case 3 with 4, we see that TR drops slightly when the starting seed contains texts.

The low LP and LR do not accurately reflect the real quality of the generated ontology. LP has a tendency to decrease because the system is using the Web as a corpus, so it will inevitably include items absent from the GS. On the other hand, manual inspection of the ontology showed that in 230 triples, there were 225 concepts of which only 14
could be clearly seen to belong to another domain (flour, book, farmer, plant etc.), and another 10 were borderline (predatory bird, domestic dog, wild rabbit, large mammal, small mammal, wild fowl, etc.). So a manual evaluation would suggest 201 correct terms or [0.89] precision. The gradually falling LP presents a challenge for ontology learning and may either need a different approach to evaluating this element or a need for techniques which focus the ontology more effectively.

The flux shown in the graph presented in Figure 34 in the early stages shows that in principle as more data is added to the system the output becomes more stable and consistent. The general tendency is for the measures to move upwards indicating a gradual but steady improvement over the progression of the iterations. These results are as was hoped and reflect the capacity of the system to adapt as the data added to the system changes the confidence values for individual items of knowledge. The high F measures for the system show that our approach has fundamental validity.

Given the high quality of the output of this approach the question arises whether this is really what is needed. Is this type of ontology too focussed and does it just succeed algorithmically to re-create the well-known tennis problem [Stevenson, 2002]? This can only be answered by further experimentation and evaluation, varying the parameters of the approach.
11.8 CONCLUSION

Ontology evaluation is a major challenge for ontology engineering in general and will always remain so because what is being evaluated is the congruence between a model of the world and our own internalised model. Each individual has a somewhat different representation and account of the world. Furthermore, our knowledge of the world is changing continuously from day to day, sometimes in imperceptible small increments, sometimes in larger leaps. Ontologies differ in their degree of granularity depending on whether they reflect what Velardi and Navigli call “foundational”, “core”, or “domain” ontologies. We can hypothesise a level below “domain” which could be termed “individual” in that the granularity reflected differences between individual people. In effect we can delineate three dimensions or axes along which ontologies differ: individual people, time, and granularity or specificity. It could be argued that along the granularity axis ontologies taper towards an apex of agreement, but this again can only be true within certain cultural boundaries.

We have discussed and summarised a range of approaches to ontology evaluation none of which obviously are individually sufficient for the task. There is a need for the judicious combination of the different approaches we have described. Factors such as time and effort involved also play an important part in determining which is the best approach for a particular task.

There is a major need for ontology evaluation competitions along the lines of MUC, TREC or SENSEVAL. We have noted the problems with gold standards and thus we believe an important element of such an effort should to create a standardised application where different ontologies can be slotted in and the changes in performance monitored. This will enable a standardised comparison of different ontology learning systems even if we need to remain cognisant of the associated problems.

We have argued that the Abraxas approach when viewed as a model of potential ontology learning systems captures the requirements we have identified and the realities of textual data and knowledge. The Abraxas approach provides the flexibility, the capacity for dynamic knowledge creation, the adaptability to the vagaries of actual text processing, which has been so far absent from the ontology learning literature. We have itemised the characteristics of knowledge and the idiosyncrasies of text which need to be taken into account.

Finally, we have shown that an implementation of one version of the Abraxas approach (i.e. a specific set of starting parameters) results in very good output, certainly superior to that currently available in existing systems.
The original objective of this work was to design a system which could automatically construct ontologies or at least taxonomies from a collection of texts. This appeared to be a feasible aim given the range of techniques available to Natural Language Processing, given the size of corpora, and given the sophistication of language processing methods. The reality was that a number of challenges arose in the process of exploring the problem.

The first of these challenges concerned the nature of knowledge. The ontologies we wanted to construct are considered representations of knowledge and this led to considering just what knowledge is and how much can we say we can capture in a structure such as an ontology. Knowledge is clearly a very complex psychological and cultural artefact and up to now only a small proportion of what we consider knowledge can in fact be represented in an ontology. We currently do not represent numerical knowledge in ontologies, nor knowledge that changes over time, nor can we represent a whole range of tacit knowledge. So ontologies seemingly only capture a fraction of what is ‘out there’ or rather ‘in here,’ in our minds. There are more important complications as well. One of these is that knowledge is continuously changing. Each experience a human being has, each text they read, changes the way they look at the world, changes part of their working ontology. Every step of technological change or scientific progress changes our understanding of the world. Also, human beings seem quite content to function with under specified knowledge, for which there is only partial or conflicting evidence. All this means that the static, linear, monotonic view of ontology building needs revision, especially if a successfully engineered system is to be designed.

The second challenge has been the limitations of Natural Language Processing as a provider of technology or a set of technological tools. NLP can only achieve a certain amount in its chosen set of tasks. While we can POS tag a text at 97% precision, most other tasks in NLP range in their performance from 50% to 95% precision. More fundamentally given a closed set of categories we can annotate or classify texts in a variety of ways reasonably successfully, but when it comes to open ended tasks involving a complex set of steps our absolute levels of success are limited. The classic example is machine translation which has never been able to challenge human translation in terms of overall quality. However, as Yorick Wilks has repeatedly noted, machine translation is a success commercially and practically when people need rapid access to foreign language texts and are kept aware of the limitations. Ontology learning can also be considered
a similarly complex task involving a number of intermediate steps all of which function imperfectly. Crucially, like machine translation, there is no one ‘correct’ ontology to be derived from a given set of texts. Just like a translation is an interpretation to some extent of the source text, so an ontology is an interpretation over the text collection. There is at least as much variability as in translation and probably more. We would argue that Ontology Learning is able to give usable, workable sketches of an ontology which will prove very useful to ontology developers.

The fundamental technical challenge, that we identified was that the ontology to a large extent was absent from the a given text or set of domain specific texts. This constituted our third challenge. A great deal of what an ontology engineer wishes to place in an ontology is typically absent from the text because it forms part of the background knowledge the reader brings to the act of reading and which the writer assumes in the process of writing. Some of this is common sense knowledge and could be assumed to be part of a general ‘upper-level’ ontology, but a great deal is also specific to the scientific or practical domain. That cars have four wheels is a common sense piece of knowledge but is not one that any mechanics handbook would state. The effect of this is that a great body of knowledge taken for granted by the writer is rarely made explicit, and thus it is computationally inaccessible to a significant degree.

The fourth major challenge is a corollary of the fact that we are dealing with psychological and subjective artefacts: it is very difficult to objectively evaluate ontologies. While we can establish guidelines for ontology development, and rules for the rigourous definition of concepts, in the final analysis, an ontology is an act of interpretation over a body of knowledge. It is a suggestion as to how the relevant concepts should be organised and what are the salient relationships. As such there is no absolute right or wrong, only degrees of correctness. This makes ontologies hard to evaluate, and it makes it even harder to evaluate ontology learning methodologies or systems. It is possible to evaluate elements of the output, or to evaluate the general approach but not the final total ontology. Using a gold standard approach, which is the unfortunate default method used in this field as in so much NLP research, means that one is in effect evaluating the success of the system in replicating the gold standard and not whether, in general, the system is good at learning ontologies. No perfect solution for this problem has been found and all that can be achieved is indirect evaluation.

This work has sought avenues to address these challenges largely by revising the objectives of ontology learning. We have argued that it is a task which should be made both weaker and more sophisticated. We propose to make the task weaker by acknowledging that perfect knowledge is impossible in many cases. We have suggested that there are degrees of precision in a piece of knowledge and degrees of confidence. Thus we may know that the terms motorcycle and vehicle have some connection to the domain i.e. belong to the ontology but nothing
more. Slightly more precisely we may know that the two terms are related i.e. have some specific ontological relationship, but be unable to specify what that is. Even more precisely we may be able to discover that motorcycle ISA vehicle, and finally and most precisely of all we may be able to place this knowledge triple in its correct position in an overall ontology concerning the domain (e.g. transport?). We may have more or less confidence in each of these steps, depending on the quantity and quality of the evidence available.

We have also viewed the task of ontology learning as more sophisticated by proposing mechanisms whereby the knowledge missing from a given domain corpus can be identified and the specific piece of knowledge needed can be sought, obtained and verified. We have also argued that effective knowledge acquisition needs to use the diachronic corpora to identify early uses of a term and large cross-discipline corpora to identify the borrowing of terms into other scientific disciplines so as extract explicit defining contexts.

Above all, we have taken the view that knowledge in itself is not black and white, that it is uncertain, varying in quality and continuously changing. Only a model of ontology learning which attempts to capture these aspects of knowledge, i.e. in the manner of knowledge production, can ultimately be successful. A more discriminating understanding of the relationship between knowledge, text and cognition will enable a more effective ontology learning model to be constructed. It is in this context that we have proposed Abraxas as a model of what ontology learning systems should be like and also as a specific instantiation of this model. Abraxas is a dynamic, data-driven model of ontology learning, which attempts to capture the degrees of certainty associated with items of knowledge. As we have described in Chapter 10, a key aspect of the Abraxas model is the concept of equilibrium between three resources (the corpus of documents, the set of extraction patterns, and the ontology or set of knowledge triples). Changes in this equilibrium result in a learning process which may add more knowledge triples, or add further documents or learn further extraction patterns. The model takes account of degrees of certainty in that each resource item (knowledge triple, document, extraction pattern) has an associated confidence level which continually changes as the system iterates. The system is dynamic in its ability to change its confidence level in any item over time or further iterations. It is a data-driven model because the confidence the system has in each resource depends on the its relationship to the other existing resources. So, for example, evidence from the corpora and the extraction patterns provide confidence in any given knowledge triple.

We have no illusions that the Abraxas model provides a complete solution to the problems we have identified. It is merely less bad than existing attempts at ontology learning for cognitive and theoretical reasons, and initial results from a computational perspective are promising. As we have noted elsewhere, potentially the Abraxas model could be implemented in a number of different ways while still keeping its fundamental characteristics.
12.2 FUTURE WORK

In this work that we have raised far more questions than we have been successful in answering. In this final section, we indicate some of the directions for future work that we hope will be undertaken either by ourselves or colleagues in the field.

There is a core issue in knowledge representation which is both philosophical and pragmatic. It concerns the limited extent to which knowledge can be communicated through language, and the even more limited extent to which it can be captured by knowledge representation mechanisms and ontologies in particular. While these limitations are obvious to researchers with experience in the field, there appears to be a need for a continuous re-education concerning the substantial gap between human knowledge and that which is representable for computers. A lack of awareness of this results in a great deal of re-invention of the wheel, but it is unclear to this author how this re-invention can be mitigated. So at a practical level it would be very important if future work on ontology learning and evaluation could start with an awareness of the limits and issues described in this thesis, rather than starting out to attempt the impossible with assumptions that writers over a long period have shown to be untenable.

With respect to the Abraxas approach, there are number of areas in which we would wish to undertake further investigation. The core issue addressed in this thesis - how to bridge the gap between what is explicitly stated in text and what we expect to go into an ontology - is open to a great deal of further work. On the one hand, the use of pattern learning extends the range of lexico-syntactic patterns and what can be identified in “explicit” contexts. On the other hand, for the general purpose of ontology learning it is important to extend the learning mechanisms to be able to mine structured as well as unstructured texts. One way of doing this would be use the approach of the Armadillo technology [Ciravegna et al., 2004, Glaser et al., 2004]. This bootstraps using very simple seeds to learn large quantities of knowledge using the inherent redundancy of the web.

In Section 11.7.2, we undertook some initial experiments varying the initial seeds, but as we noted in that Section and in Section 10.2, there are potentially 27 different starting states (25 of which lack equilibrium) and all these need to be investigated. Some of these starting states will result in “random” ontologies rather than domain specific ones. Other starting states will result in a system functioning like an IR system, where the ontology/set of knowledge triples act as a set of queries for which corresponding documents are sought. Another aspect worthy of further investigation is the role of the user. If we accept, as many researchers in the field do, that full automation of ontology learning is impossible, then the exact extent and nature of user intervention and/or input needs to be determined. For example, what is the most useful kind of input, and what are the returns of that effort? How does this vary across domains? These issues are important
in order to use Abraxas in an enterprise setting (in a company or organisation). Depending on the requirements, and the documents available, a thesaurus, a taxonomy, or a fully-fledged OWL compliant ontology may be required. For each of these scenarios (and there are many more fine-grained scenarios in between) it would be very useful to determine the needed starting states, and the required degree of user input to achieve a specified ontology of a certain formality. Recent work on getting users either unwittingly or willingly to perform tasks that are inherently hard for computer systems shows great promise [von Ahn and Dabbish, 2004, von Ahn et al., 2006, Good et al., 2006].

The approach has at its heart the concept of an equilibrium between resources. We have proposed in Chapter 10 a method to measure this equilibrium using what we termed the Knowledge Gap and the Explicit Knowledge Gap. However, we believe there may other, possibly better ways to evaluate the ‘fit’ between an ontology, a corpus of documents, and a set of extraction patterns. There is also closely related to this the problem that due to sparsity of data in text, there will tend to be an enormous number of terms that need to be added to the ontology but for which no ‘explicit context’ can be identified to determine the ontological relationship. This may lead in two directions. One is to consider whether the concept of equilibrium is the most fruitful in this kind of complex model of ontology learning, and two is to find other means to be able to calculate the ontological relationships.

For the later issue, we might either decide that a weak ontological relationship of ‘belonging to the domain’ is sufficient, or we may attempt to use the word class smoothing techniques used in speech recognition to identify the probable ontological relationships given a term’s distributional behaviour [Brown et al., 1992]. Closely related to the issue of whether the equilibrium measures are appropriate is the issue of stopping criteria. If there is always going to be a long tail of unaccounted for terms, then we need coherent stopping criteria in terms of information gain for each further iteration of the system.

The nature of the raw materials used in ontology learning (particularly corpora) need a great deal of further study because greater understanding will allow for the design of more sophisticated methods. For example, there arise complex issues with documents which objectively appear to reflect knowledge from multiple ontologies. One way of expressing this that the ontology currently under construction could get “contaminated” from documents which do not really reflect the domain being described. There could be various ways to address this (pre-clustering of documents, for example). We also discussed in Chapter 9 how having a proper diachronic access to corpora could greatly increase the likelihood of finding relevant texts with explicit context for knowledge acquisition. This would needs to be empirically proven to be a useful avenue. The Chronopath project [Catizone et al., 2006] attempts to tightly relate Internet texts with their time of writing/publication, and this kind of technology could be fruitfully

---

1 Consider the example given by [Sinclair and Ball, 1996] and quoted on Page 38
2 [http://www.nlp.shef.ac.uk/cronopath/](http://www.nlp.shef.ac.uk/cronopath/)
applied for the purposes of identifying the most useful texts for ontology learning. In a similar manner, a greater understanding of the role of text genre would also allow us to understand how knowledge is communicated more effectively or more concisely in certain contexts than in others. There is currently more and more research on automatically identifying genre and this could also be useful in the Abraxas context [Kessler et al., 1997, Boese and Howe, 2005, Santini et al., 2006].

Closely related to the issue of text genre is the degree of confidence associated with different knowledge sources, both textual and non-textual. In the area of biomedical ontologies and e-science, considerable work has been done on developing categories of “provenance” and “evidence” [Karp et al., 2004, Zhao et al., 2004, Couto et al., 2006] and given the importance of degrees of certainty in the Abraxas model this is an approach which should be integrated as well. One can envisage not just degrees of certainty but also types of certainty reflecting, for example, the type of textual or non-textual source for the particular knowledge item. Here a greater understanding of text genre and what that conveys as to the confidence a human reader has in the knowledge obtained from there would be a significant research area.

In Chapter 11, we discussed at some length the problem of ontology evaluation. We believe there are still a number of open issues. There is a great need to construct proper application based scenarios for the evaluation of ontologies. Most current research on ontology evaluation either uses a “methodological” approach (e.g. OntoClean) or uses metrics of one sort or another. Neither approach deals with the real issue of the quality of the knowledge in the ontology and its impact on the performance of a running system. Only if a standardised application is created where different ontologies can be “slotted in” can researchers evaluate the real difference in their design and construction of ontologies whether of a manual or automatic origin. Closely related to this, as noted in Section 11.6, is the need for clearly defined scenarios to evaluate ontology learning systems. This would mean to start with a standard set of texts from which to learn ontologies, a well-developed and consensually derived ontology or ontologies which could act as Gold Standards, and a clear set of parameters concerning the use of external knowledge, human intervention etc. This is a substantial undertaking which would need the kind of resources that have been given to TREC and MUC.

This is not to reduce the importance of the development of metrics for ontologies such as those we discussed for ranking ontologies. Our own work on ranking ontologies needs to be rigourously evaluated with human users of such metrics so as to establish the extent to which the metrics correspond to human instincts. A study here is needed with a reasonable number of subjects, all with some expertise in the use of ontologies, who can rank ontologies with respect to specific tasks or objectives. Care needs to taken in designing an appropriate

---

3 For example, consider the publications at the most recent Evaluation of Ontologies for the Web workshop: [http://km.aifb.uni-karlsruhe.de/ws/eon2006](http://km.aifb.uni-karlsruhe.de/ws/eon2006)
questionnaire which provides the necessary information and also asks appropriate questions.

There is a substantial overlap between the task of ontology learning and automated question answering (cf. the TREC QA series [http://trec.nist.gov/data/qamain.html]. Essentially, the QA tasks in NLP are seeking to identify isolated facts while ontology learning is seeking to collect a large body of facts, and organise these into a coherent whole. Both QA and ontology learning as we have approached the subject in this thesis involve extensive use of lexico-syntactic patterns as templates for identifying answers [Saggion and Gaizauskas, 2004]. Hence we believe there needs to be a collaboration between these two sub-fields in NLP so as to identify what each can learn from the other. This is especially important given that QA has a very rigourous evaluation paradigm which is absent from ontology learning.

There are two final directions for future research which we would like to mention. Firstly, one area for future work lies in the integration of Abraxas with dialogue systems (such as those currently being developed in the Companions project [http://www.companions-project.org]). Ontologies are likely to play a role in such systems in the future as part of the underlying knowledge representation system and need to know how to handle new pieces of knowledge provided by the interlocutor with the system. The confidence level approach integral to Abraxas is appropriate here because the type of interlocutor and the re-confirmation of pieces of knowledge can all influence the confidence the system has in the knowledge acquired. At the same time, the Abraxas approach provides a partial methodology for rapidly constructing an appropriate ontology for a new domain.

Finally, we would like to consider how ontology learning from texts can be integrated with social network software which is being seen as playing an ever more significant role in the Semantic Web and the internet in general. Could the behaviour of individuals in social networks be harnessed to provide input to the resource confidence levels? Could we design social networking components which would augment the ontologies learnt from texts? Folksonomies which are easily generated in social networks [Golder and Huberman, 2005, Schmitz et al., 2006] could be integrated with textually derived ontologies complementing them and re-enforcing elements where explicit knowledge is weak.
Part II

APPENDIX
Here we present the detailed results for the experiment described and analysed in Chapter 8 which used ten pairs of terms taken from the Gene Ontology [Ashburner et al., 2000]. The results are summarised in Tables 23 and 22 in that chapter.

The pairs of terms were derived from a version of the Gene Ontology of April 2003. Since then it has been revised several times so the exact terms may not be found.

<table>
<thead>
<tr>
<th>Terms:</th>
<th>histolysis isa tissue death</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td></td>
</tr>
<tr>
<td>Google citations</td>
<td>3, all from the ‘Gene Ontology’</td>
</tr>
<tr>
<td>Google glossary</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>1, under ‘lepidopteran’ “tissues of the larva undergo considerable histolysis (breaking down)”</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>histolysis: “the breakdown of bodily tissues”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms:</th>
<th>flocculation isa cell communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td></td>
</tr>
<tr>
<td>Google citations</td>
<td>31, of which about 10 from ‘GO’, but no defining contexts</td>
</tr>
<tr>
<td>Google glossary</td>
<td>9</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>7 references, none concerning cell communication</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>flocculate: to cause to aggregate into a flocculent mass, : to become flocculent</td>
</tr>
<tr>
<td>Terms:</td>
<td>vasoconstriction <em>isa</em> circulation (nb. the GO appears to be clearly wrong here)</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>0</td>
</tr>
<tr>
<td>Google citations</td>
<td>17k, many examples showing a close relationship but NOT the one specified in the ontology</td>
</tr>
<tr>
<td>Google glossary</td>
<td>9</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>11, Sub-article on vasoconstriction which implies it is a disease of the arteries,</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>vasoconstriction: : narrowing of the lumen of blood vessels especially as a result of vasomotor action</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms:</th>
<th>holin <em>isa</em> autolysin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>0</td>
</tr>
<tr>
<td>Google citations</td>
<td>30, including GO references, showing a close association but no ontologically clear relationship. Citations show “holing is a protein”</td>
</tr>
<tr>
<td>Google glossary</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>0</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms:</th>
<th>aminopeptidase <em>isa</em> peptidase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>0</td>
</tr>
<tr>
<td>Google citations</td>
<td>7k, which tell us aminopeptidase is an enzyme not that it is a peptidase</td>
</tr>
<tr>
<td>Google glossary</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>1, which can humanly be understood to convey that aminopeptidase is an enzyme</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>aminopeptidase: : an enzyme that hydrolyzes peptides by acting on the peptide bond next to a terminal amino acid containing a free amino group</td>
</tr>
<tr>
<td>Terms:</td>
<td>death isa biological process</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>0</td>
</tr>
<tr>
<td>Google citations</td>
<td>6k, including 4 specifying that “death is a biological process”, “biological processes”/death 350k, many for “biological processes such as cell death”</td>
</tr>
<tr>
<td>Google glossary</td>
<td>8</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>7, none helpful, however in article on ‘death’, it is humanly understandable that ‘death is a biological process’</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>death: a permanent cessation of all vital functions: the end of life</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms:</th>
<th>metallochaperone isa chaperone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>0</td>
</tr>
<tr>
<td>Google citations</td>
<td>257, including many references to GO, but none clearly specifying this ontological relationship</td>
</tr>
<tr>
<td>Google glossary</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>0 (no ‘metallochaperone’, and no biological reference to ‘chaperone’)</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terms:</th>
<th>hydrolase isa enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>2</td>
</tr>
<tr>
<td>Google citations</td>
<td>29k, many contexts where this is derivable: &quot;hydrolase is a ubiquitous cellular enzyme&quot; but also: “hydrolase is a protease” “hydrolase is a peroxisomal coenzyme”</td>
</tr>
<tr>
<td>Google glossary</td>
<td>0</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>1 article, with definition: “any one of a class of more than 200 enzymes that”</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>hydrolase: a hydrolytic enzyme</td>
</tr>
<tr>
<td>Terms: ligase</td>
<td>isa enzyme</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Textual Source</td>
<td></td>
</tr>
<tr>
<td>Original corpus</td>
<td>2</td>
</tr>
<tr>
<td>Google citations</td>
<td>31k, many contexts where this is derivable: &quot;DNA ligase: Enzyme involved in the replication and repair&quot; but also: &quot;ligase is a single polypeptide&quot; “ligase is a 600 kDa multisubunit protein”</td>
</tr>
<tr>
<td>Google glossary</td>
<td>9</td>
</tr>
<tr>
<td>Encyclopaedia (Britannica)</td>
<td>1 article with definition: “also called Synthetase any one of a class of about 50 enzymes that…”</td>
</tr>
<tr>
<td>Dictionary (Britannica)</td>
<td>ligase: “an enzyme that catalyzes the linking together of two molecules”</td>
</tr>
</tbody>
</table>
Here we present a fuller justification for the formulation of the formula for the Knowledge Gap and the Explicit Knowledge Gap used in Chapter 10.

**B.1 KNOWLEDGE GAP**

We defined the Knowledge Gap in Chapter 10 as follows:

The knowledge gap measure (KG) is the normalisation of the cardinality of the difference in the set of terms in the ontology ($\Omega$) vs. the set of terms in the corpus ($\Sigma$) as shown in Eq. (B.1) (this is the same as Equation (10.33)):

$$KG = \frac{|(\Omega \cup \Sigma) \setminus (\Omega \cap \Sigma)|}{|\Omega \cup \Sigma|} \quad (B.1)$$

We can exemplify this with the following cases:

<table>
<thead>
<tr>
<th>Case</th>
<th>Corpus ($\Sigma$)</th>
<th>Ontology ($\Omega$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - no gap</td>
<td>{cat}</td>
<td>{cat}</td>
</tr>
<tr>
<td>2 - small gap</td>
<td>{cat, dog, mouse}</td>
<td>{cat, mouse}</td>
</tr>
<tr>
<td>3 - large gap</td>
<td>{cat, dog, rat}</td>
<td>{dog, duck, bird}</td>
</tr>
<tr>
<td>4 - complete gap</td>
<td>{cat, mouse, bird}</td>
<td>{dog, duck, rabbit}</td>
</tr>
</tbody>
</table>

We then can show the stages of calculating the formula above (B.1) in the following table:
Table 4. Four hypothetical cases showing the calculation of the Knowledge Gap measure.

<table>
<thead>
<tr>
<th>Case Corpus (C)</th>
<th>Ontology (Ω)</th>
<th>Knowledge Gap (KG) (B.1)</th>
<th>Explicit Knowledge Gap (EGG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0 = φ/φ</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
B.2  EXPPLICIT KNOWLEDGE GAP

We defined the Explicit Knowledge Gap as follows (from page 161):

The difference between the set of pairs of terms whose ontological relationships are known (because they are in the ontology) and those which need to be added to the ontology (because they are key terms in the corpus) and whose ontological relationship is unknown, we will term the Explicit Knowledge Gap (EKG). The absence of explicit knowledge may be between two terms both absent from the ontology or between a new term and a term already assigned to the ontology set. Let $\Pi$ be the set of triples based on terms existing in the corpus, which are known to have some kind of ontological relationship on distributional grounds. Each triple consists of $t_i, r, t_k$ where $t_i, t_k \in T$ the set of all terms in the corpus, and $r \in R$ the set of ontological relations. The set $\Pi$ consists of triples where the ontological relationship is unknown (let us represent this by $r_0$). In contrast, $\Omega_3$ is the set of pairs of terms whose ontological relationship is explicit, and these are assumed to be the set of knowledge triples in the ontology. Again obviously these can be represented as triples consisting of $t_i, r_j, t_k$ where, in this case, $t \in \Omega$ i.e. the set of terms in the ontology.

Thus EKG is defined analogously to Eq. B.1 on page 235, where $\Omega$ is replaced by $\Omega_3$, and $\Sigma$ is replaced by $\Pi$.

$$EKG = \frac{|(\Omega_3 \cup \Pi) \setminus (\Omega_3 \cap \Pi)|}{|\Omega_3 \cup \Pi|} \quad (B.2)$$

NB. The EKG refers to sets consisting of triples not single terms as in the KG. However, the measure is not concerned with the nature of the relation value in these triples, only with the terms.

We can exemplify this with the following cases:

<table>
<thead>
<tr>
<th>Case</th>
<th>Corpus triples ($\Pi$)</th>
<th>Ontology triples ($\Omega_3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - no gap</td>
<td>{cat, $r_0$, mammal}</td>
<td>{cat, $r_i$, mammal}</td>
</tr>
<tr>
<td>2 - small gap</td>
<td>{cat, $r_0$, mammal}, {dog, $r_0$, mammal}, {mouse, $r_0$, mammal}</td>
<td>{cat, $r_i$, mammal}, {mouse, $r_i$, mammal}</td>
</tr>
<tr>
<td>3 - large gap</td>
<td>{cat, $r_0$, mammal}, {dog, $r_0$, mammal}, {rat, $r_0$, mammal}</td>
<td>{dog, $r_i$, mammal}, {duck, $r_i$, bird}, {bird, $r_i$, animal}</td>
</tr>
<tr>
<td>4 - complete gap</td>
<td>{cat, $r_0$, mammal}, {mouse, $r_0$, mammal}, {bird, $r_0$, animal}</td>
<td>{dog, $r_i$, mammal}, {duck, $r_i$, bird}, {rabbit, $r_i$, mammal}</td>
</tr>
</tbody>
</table>

We can then show the stages of calculating the formula above (B.2) in the following table:
Table 4.5. Four hypothetical cases showing the calculation of the Explicit Knowledge Gap measure

<table>
<thead>
<tr>
<th>Case</th>
<th>Corpus triples (C)</th>
<th>Ontology triples (Ω)</th>
<th>Case</th>
<th>Corpus triples (C)</th>
<th>Ontology triples (Ω)</th>
<th>Case</th>
<th>Corpus triples (C)</th>
<th>Ontology triples (Ω)</th>
<th>Case</th>
<th>Corpus triples (C)</th>
<th>Ontology triples (Ω)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = 0</td>
<td>{ {cat, r_{0}, mammal} }</td>
<td>{ {bird, r_{0}, mammal}, {dog, r_{0}, mammal}, {mouse, r_{0}, mammal} }</td>
<td>0 = 0</td>
<td>∅</td>
<td>{ {cat, r_{0}, mammal}, {mouse, r_{0}, mammal} }</td>
<td>1 = 0</td>
<td>{ {cat, r_{0}, mammal}, {mouse, r_{0}, mammal} }</td>
<td>{ {dog, r_{0}, mammal}, {rabbit, r_{0}, mammal} }</td>
<td>0 = 0</td>
<td>{ {cat, r_{0}, mammal}, {mouse, r_{0}, mammal} }</td>
<td>{ {dog, r_{0}, mammal}, {rabbit, r_{0}, mammal} }</td>
</tr>
</tbody>
</table>
APPENDIX C - THE ABRAXAS ANIMALS
ONTOMETRY

Below we enclose the Abraxas animal ontology (presented here using
the Manchester syntax for readability Horridge et al. [2006]) which was
used in the evaluation of the implemented version of Abraxas (cf. Sec-
tion 11.7). It is also available from the following website in OWL/XML
and rdf: http://nlp.shef.ac.uk/abraxas/resources.html.

Figure 36. Part of the ‘Animal’ ontology visualised using OWLviz

Figure 37. Another part of the ‘Animal’ ontology visualised using OWLviz
Imports(<http://purl.org/dc/elements/1.1/>)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Dragon
DisjointClasses(Dragon Bird)
DisjointClasses(Dragon Unicorn)
DisjointClasses(Dragon Mammals)
DisjointClasses(Dragon Reptile)
DisjointClasses(Dragon Amphibian)
SubClassOf(Dragon Animals)
DisjointClasses(Dragon Marsupial)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Blackbird
DisjointClasses(Blackbird Robin)
DisjointClasses(Woodpecker Blackbird)
DisjointClasses(Ostrich Blackbird)
DisjointClasses(Crow Blackbird)
SubClassOf(Blackbird Bird)
DisjointClasses(Owl Blackbird)
DisjointClasses(Nightingale Blackbird)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Fox
DisjointClasses(Cheetah Fox)
DisjointClasses(Fox Badger)
DisjointClasses(Mouse Fox)
DisjointClasses(Fox Mole)
DisjointClasses(Squirrel Fox)
DisjointClasses(Fox Weasel)
DisjointClasses(Fox Giraffe)
DisjointClasses(Cow Fox)
DisjointClasses(Hare Fox)
DisjointClasses(Fox Rabbit)
DisjointClasses(Orter Fox)
DisjointClasses(Lion Fox)
DisjointClasses(Rhinoceros Fox)
DisjointClasses(Fox Bear)
DisjointClasses(Sloth Fox)
DisjointClasses(Tiger Fox)
DisjointClasses(Camel Fox)
DisjointClasses(Fox Hippopotamus)
DisjointClasses(Hedgehog Fox)
DisjointClasses(Rat Fox)
DisjointClasses(Fox Antelope)
DisjointClasses(Fox Monkey)
DisjointClasses(Fox Dog)
SubClassOf(Fox Mammals)
DisjointClasses(Cat Fox)
DisjointClasses(Dog Sloth)
DisjointClasses(Dog Rabbit)
DisjointClasses(Dog Hedgehog)
DisjointClasses(Dog Hippopotamus)
DisjointClasses(Dog Bear)
DisjointClasses(Dog Rhinoceros)
DisjointClasses(Dog Tiger)
DisjointClasses(Cow Dog)
DisjointClasses(Monkey Dog)
DisjointClasses(Mole Dog)
DisjointClasses(Camel Dog)
DisjointClasses(Mouse Dog)
DisjointClasses(Cheetah Dog)
DisjointClasses(Badger Dog)
DisjointClasses(Rat Dog)
SubClassOf(Dog Mammals)
DisjointClasses(Dog Giraffe)
DisjointClasses(Squirrel Dog)
DisjointClasses(Dog Weasel)
DisjointClasses(Hare Dog)
DisjointClasses(Fox
Dog)
DisjointClasses(Lion
Dog)
DisjointClasses(Dog
Antelope)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Snake
SubClassOf(Snake Reptile)
DisjointClasses(Lizard
Snake)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Grasshopper
SubClassOf(Grasshopper Insect)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Tiger
DisjointClasses(Tiger
Bear)
DisjointClasses(Tiger
Lion)
SubClassOf(Tiger Mammals)
DisjointClasses(Tiger
Cheetah)
DisjointClasses(Tiger
Rabbit)
DisjointClasses(Tiger
Dog)
DisjointClasses(Tiger
Weasel)
DisjointClasses(Tiger
Cat)
DisjointClasses(Tiger
Badger)
DisjointClasses(Tiger
Camel)
DisjointClasses(Tiger
Hare)
DisjointClasses(Tiger
Mole)
DisjointClasses(Tiger
Squirrel)
DisjointClasses(Tiger
Otter)
DisjointClasses(Tiger
Mouse)
DisjointClasses(Tiger
Antelope)
DisjointClasses(Tiger
Hedgehog)
DisjointClasses(Tiger
Cow)
DisjointClasses(Tiger
Fox)
DisjointClasses(Tiger
Monkey)
DisjointClasses(Sloth
Tiger)
DisjointClasses(Tiger
Giraffe)
DisjointClasses(Tiger
Hippopotamus)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Cat
DisjointClasses(Cat
Lion)
DisjointClasses(Cat
Squirrel)
DisjointClasses(Cat
Badger)
DisjointClasses(Cat
Rhinoceros)
DisjointClasses(Cat
Giraffe)
DisjointClasses(Cat
Rabbit)
DisjointClasses(Cat
Mole)
SubClassOf(Cat Mammals)
DisjointClasses(Cat
Cheetah)
DisjointClasses(Cat
Tiger)
DisjointClasses(Cat
Bear)
DisjointClasses(Cat
Hippopotamus)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Cat
DisjointClasses(Cat
Dog)
DisjointClasses(Cat
Cat)
DisjointClasses(Cat
Horse)
DisjointClasses(Cat
Cow)
DisjointClasses(Cat
Hedgehog)
DisjointClasses(Cat
Weasel)
DisjointClasses(Cat
Hare)
DisjointClasses(Cat
Otter)
DisjointClasses(Cat
Fox)
DisjointClasses(Cat
Antelope)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Camel
DisjointClasses(Orter
Camel)
DisjointClasses(Camel
Lion)
DisjointClasses(Camel
Squirrel)
DisjointClasses(Camel)
DisjointClasses (Cheetah Camel)
DisjointClasses (Cow Camel)
DisjointClasses (Rabbit Camel)
DisjointClasses (Hippopotamus Camel)
DisjointClasses (Rat Camel)
DisjointClasses (Mole Camel)
DisjointClasses (Tiger Camel)
DisjointClasses (Hedgehog Camel)
DisjointClasses (Dog Camel)
DisjointClasses (Hare Camel)
DisjointClasses (Weasel Camel)
DisjointClasses (Monkey Camel)
DisjointClasses (Fox Camel)
DisjointClasses (Antelope Camel)
SubClassOf (Camel Mammals)
DisjointClasses (Cat Camel)
DisjointClasses (Mouse Camel)
DisjointClasses (Giraffe)
DisjointClasses (Rhinoceros Camel)
DisjointClasses (Badger)
DisjointClasses (Sloth Camel)
DisjointClasses (Bear)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Giraffe
DisjointClasses (Cat Giraffe)
DisjointClasses (Mole Giraffe)
SubClassOf (Giraffe Mammals)
DisjointClasses (Hippopotamus Giraffe)
DisjointClasses (Cow Giraffe)
DisjointClasses (Squirrel Giraffe)
DisjointClasses (Otter Giraffe)
DisjointClasses (Lion Giraffe)
DisjointClasses (Rabbit Giraffe)
DisjointClasses (Bear Giraffe)
DisjointClasses (Rhinoceros Giraffe)
DisjointClasses (Rat Giraffe)
DisjointClasses (Hare Giraffe)
DisjointClasses (Cheetah Giraffe)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Nightingale
DisjointClasses (Nightingale Robin)
DisjointClasses (Owl Nightingale)
DisjointClasses (Crow Nightingale)
DisjointClasses (Woodpecker Nightingale)
DisjointClasses (Ostrich Nightingale)
SubClassOf (Nightingale Bird)
DisjointClasses (Blackbird)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Bird
DisjointClasses (Dragon Bird)
DisjointClasses (Reptile Bird)
DisjointClasses (Sloth Cow)
Rabbit)
DisjointClasses (Otter
Lion)
DisjointClasses (Otter
Bear)
DisjointClasses (Hedgehog
Otter)
DisjointClasses (Otter
Mole)
DisjointClasses (Otter
Antelope)
DisjointClasses (Cat
Otter)
DisjointClasses (Otter
Hippopotamus)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Bee
SubClassOf (Bee Insect)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Weasel
DisjointClasses (Squirrel
Weasel)
DisjointClasses (Rabbit
Weasel)
DisjointClasses (Sloth
Weasel)
DisjointClasses (Cow
Weasel)
DisjointClasses (Otter
Weasel)
DisjointClasses (Tiger
Weasel)
DisjointClasses (Badger
Weasel)
DisjointClasses (Cheetah
Weasel)
DisjointClasses (Bear
Weasel)
DisjointClasses (Fox
Weasel)
SubClassOf (Weasel Mammals)
DisjointClasses (Giraffe
Weasel)
DisjointClasses (Hedgehog
Weasel)
DisjointClasses (Camel
Weasel)
DisjointClasses (Lion
Weasel)
DisjointClasses (Antelope
Weasel)
DisjointClasses (Rhinoceros
Weasel)
DisjointClasses (Mole
Weasel)
DisjointClasses (Hippopotamus
Weasel)
DisjointClasses (Rat
Weasel)
DisjointClasses (Dog
Weasel)
DisjointClasses (Cat
Weasel)
DisjointClasses (Mouse
Weasel)
DisjointClasses (Weasel
Monkey)
DisjointClasses (Hare
Weasel)
DisjointClasses (Kangaroo
Weasel)
SubClassOf (Kangaroo Marsupial)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Mammals
DisjointClasses (Mammals
Reptile)
DisjointClasses (Mammals
Unicorn)
DisjointClasses (Mammals
Amphibian)
DisjointClasses (Mammals
Marsupial)
DisjointClasses (Mammals
Bird)
SubClassOf (Mammals Animals)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Amphibian
DisjointClasses (Amphibian
Unicorn)
DisjointClasses (Amphibian
Reptile)
DisjointClasses (Mammals
Amphibian)
DisjointClasses (Dragon
Amphibian)
DisjointClasses (Amphibian
Marsupial)
SubClassOf (Amphibian Animals)
DisjointClasses (Amphibian
Bird)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Organism
DisjointClasses (Dog
Rabbit)
DisjointClasses (Rabbit
Weasel)
DisjointClasses (Rhinoceros
Weasel)
DisjointClasses (Cow
Rabbit)
DisjointClasses (Lion
Rabbit)
DisjointClasses (Tiger
Weasel)
DisjointClasses (Rabbit)
DisjointClasses (Cat Rabbit)
DisjointClasses (Mouse Rabbit)
DisjointClasses (Antelope Rabbit)
DisjointClasses (Sloth Rabbit)
DisjointClasses (Camel Rabbit)
DisjointClasses (Fox Rabbit)
DisjointClasses (Rabbit Giraffe)
DisjointClasses (Rabbit Mammals)
DisjointClasses (Cheetah Rabbit)
DisjointClasses (Squirrel Rabbit)
DisjointClasses (Hedgehog Rabbit)
DisjointClasses (Hare Rabbit)
DisjointClasses (Hippopotamus Rabbit)
DisjointClasses (Otter Rabbit)
DisjointClasses (Badger Rabbit)
DisjointClasses (Bear Rabbit)
DisjointClasses (Rat Monkey)
DisjointClasses (Squirrel Monkey)
DisjointClasses (Cheetah Monkey)
DisjointClasses (Dog Monkey)
DisjointClasses (Mouse Monkey)
DisjointClasses (Antelope Monkey)
DisjointClasses (Lion Monkey)
DisjointClasses (Cow Monkey)
DisjointClasses (Cheetah Giraffe)
DisjointClasses (Squirrel Giraffe)
DisjointClasses (Cheetah Mammals)
DisjointClasses (Mouse Mammals)
DisjointClasses (Antelope Mammals)
DisjointClasses (Lion Mammals)
DisjointClasses (Cow Mammals)
DisjointClasses (Rabbit Mammals)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Toad
SubClassOf (Toad Amphibian)
DisjointClasses (Frog Toad)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Unicorn
DisjointClasses (Amphibian Unicorn)
SubClassOf (Unicorn Animals)
DisjointClasses (Mammals Unicorn)
DisjointClasses (Dragon Unicorn)
DisjointClasses (Reptile Unicorn)
DisjointClasses (Marsupial Unicorn)
DisjointClasses (Bird Unicorn)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Monkey
DisjointClasses (Rat Monkey)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Reptile
DisjointClasses (Reptile Reptile)
DisjointClasses (Mammals Reptile)
DisjointClasses (Amphibian Reptile)
DisjointClasses (Reptile Bird)
DisjointClasses (Reptile Unicorn)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Unicorn
DisjointClasses (Dragon Reptile)
SubClassOf (Reptile Animals)
owl#Koala
DisjointClasses (Kangaroo Koala)
SubClassOf (Koala Marsupial)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Badger
DisjointClasses (Badger Antelope)
DisjointClasses (Hare Badger)
DisjointClasses (Cat Badger)
DisjointClasses (Fox Badger)
DisjointClasses (Otter Badger)
SubClassOf (Badger Mammals)
DisjointClasses (Badger Weasel)
DisjointClasses (Tiger Badger)
DisjointClasses (Cow Badger)
DisjointClasses (Sloth Badger)
DisjointClasses (Hippopotamus Badger)
DisjointClasses (Badger Mole)
DisjointClasses (Squirrel Badger)
DisjointClasses (Hedgehog Badger)
DisjointClasses (Rat Badger)
DisjointClasses (Badger Dog)
DisjointClasses (Badger Bear)
DisjointClasses (Mouse Badger)
DisjointClasses (Camel Badger)
DisjointClasses (Badger Monkey)
DisjointClasses (Lion Badger)
DisjointClasses (Badger Rabbit)
DisjointClasses (Rhinoceros Badger)
DisjointClasses (Badger Giraffe)
DisjointClasses (Cheetah Badger)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Ant
SubClassOf (Ant Insect)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Spider
SubClassOf (Spider Animals)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Hedgehog
DisjointClasses (Hedgehog Dog)
DisjointClasses (Hedgehog Mole)
DisjointClasses (Hedgehog Mouse)
DisjointClasses (Hedgehog Badger)
DisjointClasses (Hedgehog Camel)
DisjointClasses (Hedgehog Weasel)
DisjointClasses (Hedgehog Squirrel)
DisjointClasses (Tiger Hedgehog)
DisjointClasses (Hedgehog Bear)
DisjointClasses (Hedgehog Badger)
SubClassOf (Hedgehog Mammals)
DisjointClasses (Sloth Hedgehog)
DisjointClasses (Hedgehog Lion)
DisjointClasses (Hedgehog Rabbit)
DisjointClasses (Rhinoceros Hedgehog)
DisjointClasses (Hedgehog Fox)
DisjointClasses (Cat Hedgehog)
DisjointClasses (Hedgehog Antelope)
DisjointClasses (Hedgehog Hippopotamus)
DisjointClasses (Hedgehog Otter)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.
owl#Animals
SubClassOf (Animals Organism)
DisjointClasses(Hedgehog, Monkey)
DisjointClasses(Cheetah, Hedgehog)
DisjointClasses(Hedgehog, Giraffe)
DisjointClasses(Hedgehog, Hare)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Mouse
SubClassOf(Mouse, Mammals)
DisjointClasses(Cow, Mouse)
DisjointClasses(Mouse, Rabbit)
DisjointClasses(Mouse, Lion)
DisjointClasses(Rhinoceros, Mouse)
DisjointClasses(Mouse, Fox)
DisjointClasses(Mouse, Monkey)
DisjointClasses(Hedgehog, Mouse)
DisjointClasses(Mouse, Mole)
DisjointClasses(Mouse, Bear)
DisjointClasses(Squirrel, Mouse)
DisjointClasses(Mouse, Antelope)
DisjointClasses(Sloth, Mouse)
DisjointClasses(Tiger, Mouse)
DisjointClasses(Mouse, Hare)
DisjointClasses(Otter, Mouse)
DisjointClasses(Mouse, Dog)
DisjointClasses(Mouse, Hippopotamus)
DisjointClasses(Cat, Mouse)
DisjointClasses(Mouse, Camel)
DisjointClasses(Mouse, Badger)
DisjointClasses(Mouse, Giraffe)
DisjointClasses(Rat, Mouse)
DisjointClasses(Mouse, Weasel)
DisjointClasses(Cheetah, Mouse)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Insect
SubClassOf(Insect, Animals)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#I n s e c t
SubClassOf(Insect, Animals)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Antelope
SubClassOf(Antelope, Mammals)
DisjointClasses(Badger, Antelope)
DisjointClasses(Cheetah, Antelope)
DisjointClasses(Sloth, Antelope)
DisjointClasses(Lion, Antelope)
DisjointClasses(Rat, Antelope)
DisjointClasses(Antelope, Bear)
DisjointClasses(Antelope, Rabbit)
DisjointClasses(Antelope, Monkey)
DisjointClasses(Mouse, Antelope)
DisjointClasses(Hare, Antelope)
DisjointClasses(Squirrel, Antelope)
DisjointClasses(Tiger, Antelope)
DisjointClasses(Hippopotamus, Antelope)
SubClassOf(Antelope, Mammals)
DisjointClasses(Antelope, Weasel)
DisjointClasses(Camel, Antelope)
DisjointClasses(Hedgehog, Antelope)
DisjointClasses(Fox, Antelope)
DisjointClasses(Antelope, Giraffe)
DisjointClasses( Otter, Antelope)
DisjointClasses(Dog, Antelope)
DisjointClasses(Cow, Antelope)
DisjointClasses(Rhinoceros, Antelope)
DisjointClasses(Cat, Antelope)
DisjointClasses(Antelope, Mole)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Robin
SubClassOf(Robin, Bird)
DisjointClasses(Owl, Robin)
DisjointClasses( Woodpecker, Robin)
DisjointClasses (Owl
Woodpecker)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Crow
DisjointClasses (Crow
Ostrich)
DisjointClasses (Crow
Owl)
DisjointClasses (Owl
Coot)
DisjointClasses (Coot
Crow)
DisjointClasses (Coot
Ostrich)
DisjointClasses (Ostrich
Crow)
DisjointClasses (Crow
Nightingale)
DisjointClasses (Crow
Blackbird)
DisjointClasses (Woodpecker
Crow)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Mole
DisjointClasses (Rhinoceros
Mole)
DisjointClasses (Hedgehog
Mole)
DisjointClasses (Mole
Giraffe)
DisjointClasses (Mole
Sloth)
DisjointClasses (Mole
Bear)
DisjointClasses (Mole
Cat)
DisjointClasses (Mole
Fox)
DisjointClasses (Mole
Mouse)
DisjointClasses (Dog
Mole)
DisjointClasses (Rat
Mole)
DisjointClasses (Hare
Mole)
DisjointClasses (Camel
Mole)
DisjointClasses (Cow
Mole)
DisjointClasses (Squirrel
Mole)
DisjointClasses (Tiger
Mole)
DisjointClasses (Badger
Mole)
DisjointClasses (Lion
Mole)
DisjointClasses (Hippopotamus
Mole)
DisjointClasses (Cheetah
Mole)
DisjointClasses (Weasel
Mole)
DisjointClasses (Otter
Mole)
DisjointClasses (Mole
Monkey)
DisjointClasses (Rabbit
Mole)
SubClassOf (Mole Mammals)
DisjointClasses (Antelope
Mole)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Owl
DisjointClasses (Owl
Bird)
DisjointClasses (Owl
Crow)
DisjointClasses (Owl
Nightingale)
DisjointClasses (Owl
Robin)
DisjointClasses (Owl
Ostrich)
DisjointClasses (Owl
Blackbird)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Ostrich
DisjointClasses (Rhinoceros
Ostrich)
DisjointClasses (Cheetah
Ostrich)
DisjointClasses (Crow
Ostrich)
DisjointClasses (Ostrich
Crow)
DisjointClasses (Ostrich
Nightingale)
DisjointClasses (Ostrich
Blackbird)

// Class: http://nlp.shef.ac.uk/
abraxas/ontologies/animals.
owl#Giraffe
DisjointClasses (Rhinoceros
Giraffe)
DisjointClasses (Ostrich
Giraffe)
DisjointClasses (Crow
Giraffe)
DisjointClasses (Ostrich
Squirrel)
DisjointClasses (Ostrich
Hare)
DisjointClasses (Ostrich
Mouse)
DisjointClasses (Ostrich
Hedgehog)
<table>
<thead>
<tr>
<th>DisjointClasses (Cheetah, Badger)</th>
<th>DisjointClasses (Rhinoceros, Lion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DisjointClasses (Cheetah, Bear)</td>
<td>DisjointClasses (Rat, Lion)</td>
</tr>
<tr>
<td></td>
<td>DisjointClasses (Lion, Antelope)</td>
</tr>
<tr>
<td></td>
<td>DisjointClasses (Lion, Rabbit)</td>
</tr>
<tr>
<td></td>
<td>DisjointClasses (Lion, Hippopotamus)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Weasel)</td>
<td>DisjointClasses (Mouse, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Cat, Squirrel)</td>
<td>DisjointClasses (Lion, Monkey)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Monkey)</td>
<td>DisjointClasses (Cheetah, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Camel)</td>
<td>DisjointClasses (Lion, Bear)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Giraffe)</td>
<td>DisjointClasses (Lion, Giraffe)</td>
</tr>
<tr>
<td>DisjointClasses (Cow, Squirrel)</td>
<td>DisjointClasses (Lion, Fox)</td>
</tr>
<tr>
<td>DisjointClasses (Rhinoceros, Squirrel)</td>
<td>DisjointClasses (Lion, Mole)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Hippopotamus)</td>
<td>DisjointClasses (Lion, Weasel)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Fox)</td>
<td>DisjointClasses (Hedgehog, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Mouse)</td>
<td>DisjointClasses (Hare, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Antelope)</td>
<td>DisjointClasses (Otter, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Mole)</td>
<td>DisjointClasses (Lion, Badger)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Otter)</td>
<td>DisjointClasses (Sloth, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Hedgehog, Squirrel)</td>
<td>DisjointClasses (Lion, Dog)</td>
</tr>
<tr>
<td>DisjointClasses (Tiger, Squirrel)</td>
<td>DisjointClasses (Cow, Squirrel)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Badger)</td>
<td>DisjointClasses (Lion, Lion)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Mammals)</td>
<td>DisjointClasses (Lion, Bear)</td>
</tr>
<tr>
<td>SubClassOf (Squirrel, Dog)</td>
<td>SubClassOf (Frog, Amphibian)</td>
</tr>
<tr>
<td>DisjointClasses (Cheetah, Squirrel)</td>
<td>DisjointClasses (Frog, Toad)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Lion)</td>
<td>DisjointClasses (Cheetah, Hippopotamus)</td>
</tr>
<tr>
<td>DisjointClasses (Squirrel, Hare)</td>
<td>DisjointClasses (Hippopotamus, Dog)</td>
</tr>
<tr>
<td>SubClassOf (Squirrel, Mammals)</td>
<td>DisjointClasses (Cow, Hippopotamus)</td>
</tr>
<tr>
<td>DisjointClasses (Cat, Lion)</td>
<td>DisjointClasses (Lion, Hippopotamus)</td>
</tr>
<tr>
<td>DisjointClasses (Tiger, Lion)</td>
<td>DisjointClasses (Hippopotamus, Giraffe)</td>
</tr>
<tr>
<td>SubClassOf (Lion, Mammals)</td>
<td>DisjointClasses (Rat, Hippopotamus)</td>
</tr>
<tr>
<td>DisjointClasses (Camel, Lion)</td>
<td>DisjointClasses (Camel, Hippopotamus)</td>
</tr>
</tbody>
</table>

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Squirrel

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Frog

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Hippopotamus

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Lion
Hippopotamus)
DisjointClasses (Cat Hippopotamus)
DisjointClasses (Hippopotamus Badger)
DisjointClasses (Hippopotamus Mole)
DisjointClasses (Hippopotamus Antelope)
DisjointClasses (Mouse Hippopotamus)
DisjointClasses (Rhinoceros Hippopotamus)
SubClassOf (Hippopotamus Mammals)
DisjointClasses (Hippopotamus Rabbit)
DisjointClasses (Fox Hippopotamus)
DisjointClasses (Hare Hippopotamus)
DisjointClasses (Hippopotamus Bear)
DisjointClasses (Hippopotamus Weasel)
DisjointClasses (Hedgehog Hippopotamus)
DisjointClasses (Otter Hippopotamus)
// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Leopard
SubClassOf (Leopard Mammals)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Rat
DisjointClasses (Rat Monkey)
DisjointClasses (Rat Sloth)
DisjointClasses (Rat Lion)
DisjointClasses (Rat Rhinoceros)
DisjointClasses (Rat Cow)
DisjointClasses (Rat Antelope)
SubClassOf (Rat Mammals)
DisjointClasses (Rat Hippopotamus)
DisjointClasses (Rat Hedgehog)
DisjointClasses (Rat Mole)
DisjointClasses (Rat Camel)
DisjointClasses (Rat Cat)
DisjointClasses (Rat Rabbit)
DisjointClasses (Rat Cheetah)
DisjointClasses (Rat Giraffe)
DisjointClasses (Rat Otter)
DisjointClasses (Rat Badger)
DisjointClasses (Rat Bear)
DisjointClasses (Rat Dog)
DisjointClasses (Rat Fox)
DisjointClasses (Rat Squirrel)
DisjointClasses (Rat Hare)
DisjointClasses (Rat Mouse)
DisjointClasses (Rat Weasel)
DisjointClasses (Rat Tiger)

SubClassOf (Sloth Mammals)
DisjointClasses (Sloth Otter)
DisjointClasses (Sloth Dog)
DisjointClasses (Sloth Sloth)
DisjointClasses (Sloth Weasel)
DisjointClasses (Sloth Cow)
DisjointClasses (Sloth Mole)
DisjointClasses (Sloth Monkey)
DisjointClasses (Sloth Hare)
SubClassOf (Sloth Mammals)
DisjointClasses (Sloth Badger)
DisjointClasses (Rhinoceros)
DisjointClasses (Sloth Squirrel)
DisjointClasses (Sloth Fox)
DisjointClasses (Sloth Hedgehog)
DisjointClasses (Sloth Cheetah)
DisjointClasses (Sloth Cat)
DisjointClasses (Sloth Camel)
DisjointClasses (Sloth Lion)
DisjointClasses (Sloth Tiger)
DisjointClasses (Sloth Giraffe)
DisjointClasses (Sloth Hippopotamus)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Woodpecker
DisjointClasses (Woodpecker Blackbird)
DisjointClasses (Woodpecker Nightingale)
DisjointClasses (Woodpecker Robin)
DisjointClasses (Woodpecker Ostrich)
DisjointClasses (Woodpecker Crow)
DisjointClasses (Owl Woodpecker)
SubClassOf (Woodpecker Bird)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Lizard
DisjointClasses (Lizard Reptile)
DisjointClasses (Lizard Snake)

// Class: http://nlp.shef.ac.uk/abraxas/ontologies/animals.owl#Hare
DisjointClasses (Hare Badger)
DisjointClasses (Rhinoceros Hare)
DisjointClasses (Hare Mole)
DisjointClasses (Hare Antelope)
DisjointClasses (Tiger Hare)
DisjointClasses (Hare Fox)
DisjointClasses (Cow Hare)
SubClassOf (Hare Mammals)
DisjointClasses (Hare Camel)
DisjointClasses (Otter Hare)
DisjointClasses (Sloth Hare)
DisjointClasses (Mouse Hare)
DisjointClasses (Hare Giraffe)
DisjointClasses (Hare Bear)
DisjointClasses (Hare Lion)
DisjointClasses (Hare Rabbit)

DisjointClasses (Hare Monkey)
DisjointClasses (Rat Hare)
DisjointClasses (Hare Hippopotamus)
DisjointClasses (Cheetah Hare)
DisjointClasses (Hare Dog)
DisjointClasses (Cat Hare)
DisjointClasses (Squirrel Hare)
DisjointClasses (Hare Weasel)
DisjointClasses (Hare Hedgehog)

DisjointClasses (Ostrich Robin)
DisjointClasses (Crow Ostrich)
SubClassOf (Ostrich Bird)
DisjointClasses (Woodpecker Ostrich)
DisjointClasses (Ostrich Blackbird)
DisjointClasses (Owl Ostrich)
DisjointClasses (Nightingale Ostrich)

SubClassOf (Ostrich Bird)
DisjointClasses (Tiger Bear)
DisjointClasses (Sloth Bear)
DisjointClasses (Dog Bear)
DisjointClasses (Bear Mole)
DisjointClasses (Antelope Bear)
DisjointClasses (Mouse Bear)
DisjointClasses (Bear Weasel)
SubClassOf (Bear Mammals)
DisjointClasses (Cat Bear)
DisjointClasses (Bear Monkey)
DisjointClasses (Cow Bear)
DisjointClasses (Bear Giraffe)
DisjointClasses (Rhinoceros Bear)
DisjointClasses (Fox Bear)
DisjointClasses (Hare Hedgehog)
Bear)
DisjointClasses(Rat
Bear)
DisjointClasses(Badger
Bear)
DisjointClasses(Squirrel
Bear)
DisjointClasses(Hippopotamus
Bear)
SubClassOf(Bear Mammals)
DisjointClasses(Otter
Bear)

Bear)
DisjointClasses(Bear
Rabbit)
DisjointClasses(Cheetah
Bear)
DisjointClasses(Camel
Bear)
DisjointClasses(Lion
Bear)

)


Harith Alani, Stephen Harris, and Ben O'Neil. Winnowing ontologies based on application use. In Proceedings of 3rd European Semantic Web Conference (ESWC), Budva, Montenegro, 2006b. (Cited on page 195.)


Sue Atkins, Jeremy Clear, and Nicholas Ostler. Corpus Design Criteria. (Cited on page 72.)


Bonnie J. Dorr and Doug Jones. Role of word sense disambiguation in lexical acquisition predicting semantics from syntactic cues. In Proceeding of COLING 96, Copenhagen, 1996. COLING. (Cited on page 94.)


Mathematical and Ian R. Horrocks. Using an expressive description logic: FaCT or fiction ? In Anthony G. Cohn, Lenhart Yannis Kalfoglou, John Domingue, Enricco Motta, Maria Vargas-Vera, and Simon Buckingham Shum. My-


Margaret Masterman. The thesaurus in syntax and semantics. Mechanical Translation, 4(1-2), 1957. (Cited on page 36.)


Jose Braga de Vasconcelos, Chris Kimble, and Álvaro Rocha. Ontologies and the dynamics of organisational environments an example of a group memory system for the management of group competencies. In Proceedings I-KNOW’03, Graz, Austria, July 2003. (Cited on page 16.)


COLOPHON

This thesis was typeset using André Miede’s ClassicThesis style available from www.miede.de. It is typeset with \texttt{\LaTeX} using Hermann Zapf’s \textit{Palatino} and \textit{Euler} type faces (Type 1 PostScript fonts \textit{URW Palladio L} and \textit{FPL} were used). The listings are typeset in \textit{Bera Mono}, originally developed by Bitstream, Inc. as “Bitstream Vera”. (Type 1 PostScript fonts were made available by Malte Rosenau and Ulrich Dirr.)

The typographic style was inspired by Bringhurst’s genius as presented in \textit{The Elements of Typographic Style} [Bringhurst, 2002]. It is available for \texttt{\LaTeX} via CTAN as “\texttt{classicthesis}”. 

\textbf{NOTE:} The custom size of the textblock was calculated using the directions given by Mr. Bringhurst (pages 26–29 and 175/176). 10 pt Palatino needs 133.21 pt for the string “abcdefghijklmnopqrstuvwxyz”. This yields a good line length between 24–26 pc (288–312 pt). Using a “\textit{double square textblock}” with a 1:2 ratio this results in a textblock of 312:624 pt (which includes the headline in this design). A good alternative would be the “\textit{golden section textblock}” with a ratio of 1:1.62, here 312:505.44 pt. For comparison, DIV9 of the typearea package results in a line length of 389 pt (32.4 pc), which is by far too long. However, this information will only be of interest for hardcore pseudo-typographers like me.

To make your own calculations, use the following commands and look up the corresponding lengths in the book:

\begin{verbatim}
\settowidth{\abcd}{abcdefghijklmnopqrstuvwxyz} \\
\the\abcd \ % prints the value of the length
\end{verbatim}

Please see the file \texttt{classicthesis.sty} for some precalculated values for Palatino and Minion.

\textit{Final Version} as of July 11, 2008 at 10:42.