All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346
Systems of concepts and their extraction from text

Lee Gillam

Submitted for the degree of Doctor of Philosophy

August 2004

UniS

Department of Computing
School of Electronics and Physical Sciences
University of Surrey
Guildford,
Surrey
GU2 7XH,
United Kingdom

© Lee Gillam 2004
Abstract

A method that could be used to populate, or more accurately to seed, terminology collections, and subsequently to seed models of specific domains, latterly called ontologies, is proposed, demonstrated and evaluated based on analysis of text collections, and with reference to recent work in international standards for terminology. The activity of populating ontologies is referred to elsewhere as ontology learning. Ontologies are considered by some as vital to the development of the Semantic Web and its Grid counterpart, and to the development of the emerging, yet elusive, “Knowledge Grids”. Results of this work could be used to support activities of terminologists, document managers, developers of intelligent systems, and other language researchers.

The research investigates the population of knowledge bases with systems of concepts extracted from texts in arbitrary domains. Such population is normally undertaken manually by domain experts. The method relies on identifying evidence of key domain concepts, expressed through terms used in place of these concepts, in the definition of these concepts and to express relationships between concepts. The work presented may contribute to the Semantic Web and related initiatives by helping to overcome the well-documented and unsolved AI problem of producing an initial model of an arbitrary specialist domain from background resources without significant hand-crafting effort and involvement of a domain expert: the so-called "Knowledge Acquisition Bottleneck". This bottleneck is usually overcome through extensive interactions with domain experts, involving a number of expert interviews. The research explores issues of terminology extraction from domain texts, the need for and use of knowledge representation, and the means by which terminology extraction and knowledge representation can be combined with international standards for terminology to produce such an initial model of an arbitrary specialist domain. The result of applying the presented method, the initial domain model, can be validated by domain experts, reducing the need for expert involvement in the creation of this model.
Acknowledgments

Thanks are clearly due to Professor Khurshid Ahmad for providing the basis and guidance that has enabled this work to emerge. This has included activities in interesting areas – that includes sewage and criminal activities - with a variety of characters, from which ideas have emerged, systems have been developed, and eventually research papers have emanated. Discussions, debates, and similar at various volumes have always lead to interesting results, and I hope this collaboration will continue for some time.

To some of the characters over the years have included, in no particular order, Jannis Raftopoulos and Petra Ristau at JRC (ACE, GIDA), Laurent Romary, Allan Melby, Klaus-Dirk Schmitz, Gerhard Budin and Sue-Ellen Wright (SALT, ISO), members of Surrey and Kent Police (SOCIS) and a host of others.

To colleagues at Surrey including David Pitt, Gary Dear, Matthew Casey, Mariam Tariq, Lydia Kocur and Sophie Gautier O'Shea have helped to keep things on an even keel. I have worked with numerous others including Paul Holmes-Higgin and Caroline Jones. All have been a useful “sanity-check” at various times, which is certainly appreciated. Also, to the SSTL football team, with which I’ve managed to keep more than a tenuous link to physical fitness.

To those members of my family who helped me get this far – they know who they are. If I were to dedicate this work, it would be to John Arthur Gillam (1924-2000).

To “CAS” - she suffered my various attempts at explaining this and working on it late, amongst other work activities, including disappearing to conferences and other events. A wonderful distraction at all other times!

Finally, to those who read, digest and critique this work, including examiners – thanks for spending your time and effort on it. I hope you will think this time well spent.

To anybody I have unwittingly omitted from this I can only offer my apologies.
Contents

1 Introduction ....................................................................................................................1
  1.1 Scope .........................................................................................................................4
  1.2 Overview ....................................................................................................................5
  1.3 Research Contribution ...............................................................................................8
  1.4 Structure of Thesis .....................................................................................................13

2 Knowledge, Language and Computing .....................................................................16
  2.1 An elaboration of the ‘knowledge’ and language ......................................................18
  2.2 Ontology ....................................................................................................................24
  2.3 Terminology ..............................................................................................................39
  2.4 Intelligent Computing ...............................................................................................52
  2.5 Chapter Conclusion ...................................................................................................55

3 Constructing “Knowledge” Resources .......................................................................57
  3.1 Motivation .................................................................................................................58
  3.2 Terminology Collections ...........................................................................................59
  3.3 Terminology Extraction ............................................................................................76
  3.4 Modelling knowledge resources ..............................................................................113
  3.5 Conclusion ................................................................................................................122

4 System Description .....................................................................................................126
  4.1 Method Description ..................................................................................................129
  4.2 Combined algorithm ...............................................................................................135
  4.3 System Description ..................................................................................................141
  4.4 Testing and evaluation .............................................................................................145
  4.5 Conclusion ................................................................................................................153

5 Discussion and Outlook ..............................................................................................157
  5.1 Possible impacts ......................................................................................................160
  5.2 Contributions ..........................................................................................................161
  5.3 Questions for Future Research .................................................................................165
List of Figures

Figure 1: Research projects to which the author has contributed since 1995, denoted by the dotted horizontal line............................................................................................................................10
Figure 2: Software prototypes and their research (project) heritage ...............................................12
Figure 3: Meaning triangle (Ogden and Richards 1923)........................................................................20
Figure 4: Maedche's figure of interactions required for communication that can be supported by ontologies ..................................................................................................................................21
Figure 5: Tree of Porphyry (derived from Aristotle's categories).......................................................32
Figure 6: The Cyc Upper Ontology ....................................................................................................35
Figure 7: Oard's taxonomy of cross-language retrieval approaches ................................................39
Figure 8: ISO standards development workflow ..............................................................................40
Figure 9: Fragments of ISO 704, principles and methods of terminology work, showing the scope of the International Standard and an example concept-oriented terminology ..................42
Figure 10: Fragments of ISO 1087-1 showing part of the title page, and some of the interlinked definitions of "concept" types used in terminology (French and English definitions) ..........43
Figure 11: The principle differences between terminology and lexicography ...................................44
Figure 12: Ontology spectrum (Lassila and McGuinness 2001) ..........................................................51
Figure 13: Fragment of BS EN 12622 showing terms in English and French for gas welding equipment. Note the "concept" reference on the left, and that the English definitions do not follow the principle of substitutability. .................................................................63
Figure 14: Fragment of page 15 of BS 1000[8], 1993, showing codes for various languages, in which Biblical Hebrew is listed with code 811.411.16'02, and Biblical Aramaic is listed with code 811.411.171.1'02 .................................................................................................................................65
Figure 15: An example of ambiguity in the Lenoch classification system ........................................66
Figure 16: Ontological commitment to systems of classifiers and other descriptors needed to describe a standards-conformant term. Dotted arrows show interdependence in identification ........................................................................................................75
Figure 17: Plot of frequency x rank versus rank showing the region of agreement with the 1/10 corpus size estimation for BNC. This appears to approximate between ranks 500 and 3500 only.................................................................83
Figure 18: Plot of frequency times rank, versus rank, for the first 3000 most frequent words in the Brown corpus. Note the proximity to 1/10 (about 100,000) for these ranks in the corpus. .84
Figure 19: Graph of frequency multiplied by rank (Zipf's law) for all words in the 1 million word nanoscale science and design corpus........................................................................................86
Figure 20: Plotted trends from the above table of percentages of words occurring in the seven corpora.......................................................................................................................................88
Figure 21: Difference in percentage between predicted quantity of words at frequencies of 1 to 10 (x-axis) made using Zipf’s law, and the actual quantities in 6 corpora. Note values at frequencies around 3 and 4 tend to become positive for specialist corpora...........................................88
Figure 22: Graph of frequency multiplied by rank (Zipf’s law) for the first 1000 words in the 1 million word nanoscale science and design corpus..........................................................90
Figure 23: Graph of Z-Scores (x-axis) versus the percentage of types selected at this value for the five specialist corpora and the BNC, starting at z=1...........................................................................92
Figure 24: Log weirdness (x-axis) versus log rank (y-axis) for the "smoothed" values of weirdness for the Nanoscale Science and Design corpus.................................................................97
Figure 25: 1000 most weird words in the nanoscale corpus, using the modified version of weirdness...................................................................................................................................97
Figure 26: 100 most weird words in the nanoscale corpus, using the modified version of weirdness...................................................................................................................................98
Figure 27: Normalized counts of highly frequent-highly weird words characterised against the z-scores that produced them ..............................................................100
Figure 28: Compatibility of terminology collections and the specific example of the Language field..........................................................................................................................115
Figure 29: Terminology metamodel from ISO16642.................................................................116
Figure 30: User interface of the Protege system showing an example ontology for newspapers 119
Figure 31: User interface of the KAON workbench showing various items of information extracted about "terms" that have been coded into a hierarchy, and have some relations extracted. ..............................................................120
Figure 32: Refinement of Buchanan’s approach to converting from text to knowledge base via text understanding program, adapted from Waterman et al 1983, p132 .............................................128
Figure 33: (Following Maedche). An instantiasted ontology structure for a set of concepts from the DHydro Hydrographic Dictionary.................................................................138
Figure 34: System Architecture Diagram.......................................................................................142
Figure 35: The ColloQator research tool.......................................................................................143
Figure 36: Direct export of results to Excel for further manipulation...........................................144
Figure 37: Following export to RDFS, the seed ontology can be edited and visualised using the Protégé ontology editor .................................................................144
Figure 38: Tree representation of the candidate compound terms being extracted using this method.....................................................................................................................................146
Figure 39: A fragment of the index of Radovic's "Chemistry and Physics of Carbon" showing an indexed typology of carbon nanotubes ................................................................. 151
Figure 40: Fragment of text from Saito, Dresselhaus and Dresselhaus (Chapter 5, page 73) discussing synthesis of different types of carbon nanotubes ................................................ 152
Figure 41: Use of single-walled and multiwalled in relation to nanotubes in the Kolasinski text 152
Figure 42: Part-validation, through “in-text” and index analysis, of the extracted conceptual structure. Note that elements validated in this manner are shown here in red boxes .......... 153
1 Introduction

Questions relating to the nature of knowledge have often been asked, but a grand unifying theory of knowledge has yet to be proposed. According to a human-oriented view of the world, knowledge is applied in the process of carrying out some task: either existing knowledge is applied, or some form of new knowledge is created. How do we recognize this? When we are able to explain something, such as how something works or why something is the way it is, we could be said to “have the concept” of the thing. This having of concepts means that we can ask other questions of it, classify other things by it, and relate it in various ways to different audiences who may have extensive, or no, prior knowledge of it. We are said to learn by such a process. However, can we ask questions of something if we do not “have the concept”? A popular base for semantics is the link between some real world object, a mental concept, and some sign, whether symbolic or textual. If we do not know about something, how can we ask questions of it? How, then, can we learn new things?

Computing, and the so-called “representation of knowledge” has allowed us to appraise various questions regarding the shape, structure and nature of knowledge, although discussions in various communities tend actually to discuss information, its storage, transmission, compression and so on. This discussion of information as knowledge has a tendency to cloud the debate of what to represent and how, with taxonomic debates regarding what constitutes data, information, knowledge, wisdom, pragmatics and their ilk. We would like answers to questions such as: what is the knowledge of this subject? The answer to such a question depends on ontology and epistemology. The a priori classification of details of some computational representation as one of data, information or knowledge does not assist in the process of determining the nature or the existence of the known and its relation to other such things. The human processes of learning how to carry out certain tasks, of understanding how one thing relates to another, and of abstraction for the purposes of classification are embedded at an early age and only tested later: for example, grammar is taught after a first language has been learnt. If grammar were taught before any language, would we actually learn a language in the first place? Could the fact that England has significantly fewer multilingual people per capita than the rest of Europe be symptomatic of an attitude that English is more of a globally accepted language than, say, Danish? Or is it possible that exposure to more than one language prior to the imposition of grammar results in a process of both the learning and the abstraction of, say, grammar in general and not just that of the English language? Deeper questions relating to knowledge and language,
such as whether the human brain is predisposed to language, or whether language has developed the way it did because it is shaped according the human brain can also be asked. There must be some inherent need for communication, and the use of this communication for the purposes of development: evolution of the species perhaps? Darwinian arguments would seem valid in that where communities of animals do not have well-developed systems of communication, their societies do not change over long periods. Counter to this would appear to be the “dancing bees”: the notion that on finding food sources, honeybees perform elaborate dances and participate in a democratic process for finding the closest, most appropriate source. Perhaps the complexity of the communication in contrast to the end result is an inhibitor to evolution?

From the perspective of computing, although we can evolve highly elaborate theories regarding the nature of knowledge, it is only really when those theories are tested that show their ability to encompass reality: Newton’s notions of space and time lasted some 300 years before Einstein modified it: Einstein’s modifications, when tested, better predicted the orbit of Mercury than Newton had. Only with such a “benchmark” is it possible to say that one theory is superior to another. Computing is an activity, and computation requires that which is to be computed to be represented somehow. The internal binary representation of the computer underlies, say, the textual presentation seen on computer screen, while what is presented in text forms some kind of representation of what we wish to convey from, usually, one human brain to another. Predominantly, this computable form, and not the underlying binary, is referred to as information.

Where information becomes knowledge is highly debatable when we speak of what is inherently a set of wires, switches and other electronic devices. This treatment reduces grandiose notions of “knowledge processing” to that of “information processing”, which has overtones of Turing’s work, with a semantic counterpart in Minsky.

The computational treatment of knowledge, if we can be so bold, is treated variously in the context of the so-called Semantic Web (Berners-Lee, 2000), and its relationship with Knowledge Management (Schreiber et al, 2002). Generally, this consists of tagging information, organising the tags into hierarchies and writing code that can do something with information that has a certain tag. The continuing arrogance of practitioners of computing in appropriating ideas from other fields and then modifying what is understood of them for their own purpose, has resulted in this being called “ontology”. Ontology within computing is currently seen as the Holy Grail for system integration, for information organisation, for classification, and for a whole host of other tasks. The reality of the situation is that a grandiose tag – ontology – has been put on various resources such as dictionaries and glossaries: computer scientists probably perform the greatest abuses of the English language of all subject specialists, and yet it is the challenge of language
understanding that possibly poses the greatest challenge for these specialists. The irony is certainly not lost on this author.

Casting off the grandiose titles for a moment, what we are interested in is dealing with specific types of information. More specifically, we are interested in what Minsky referred to as the "thesaurus problem" (Minsky 1968, p27). This problem is of building and maintaining a thesaurus useful for a specific task or set of tasks, and indeed, of learning to build thesauri, and finding "new ways [to] make machines first to use them, then to modify them, and eventually to build for themselves new and better ones" (ibid). Since the majority of information now being produced is "born digital", and as the volume of digital content increases, a need for better management of this information becomes apparent. As more and more historic materials are brought into the digital environment, the quantity of available data year on year steadily increases. Each of these items of information has its own form and ways in which it can or might be described and subsequently used by a human or a computer. A number of approaches to organising and categorising these volumes of information into specific "views" exist. The common or different views provide elements of this thesaurus. The view of the world for a specific purpose reflects the view of the community that is likely to use it, and the description of this view, in natural language, provides an artefact that can be identified, stored and managed. Views of the world depend on the observer, in a manner reminiscent of Einstein's frames of reference. A particular person, in a particular place, at a particular time will perceive different features of a particular object to a person perceiving the same object but with any number of space-time locations different. Different features may be deemed of greater importance or significance due to a personal or professional outlook: a mountain may be perceived as an aesthetic feature of a landscape by an artist, as a series of challenges to be overcome by a mountaineer, and as an obstacle by a civil engineer. Two persons may be able to agree on similar features of the object, but their own interpretations, and specifically what that object means to each of them will differ. Consider, for example, animal taxonomy. A zookeeper does not necessarily require information about the genus/species differentia in the mammalian world, but is more likely to be concerned about whether two types of animals can co-exist in the same environment, or whether this bringing together would result in one less animal. The zookeeper therefore requires information about whether X eats Y, and if X and Y's environment is of a similar type, for example, aquatic. The X's and Y's can be referred to as the "things", for which these descriptions are needed in such a form that they can be reasoned over.

Reasoning over sets of related information is something young humans are very capable of dealing with, but machines require either data coded using consistent schemes, or highly structured reasoning systems (computer programs). Such reasoning systems have, historically,
been referred to as *expert systems*, although it has been argued that the labels *novice system* or *idiot savant* are more appropriate since the systems are not easily able to work with missing information, or infer new knowledge from new situations. The study of "things" and their features stems from work of philosophers such as Socrates, Plato and Aristotle and has undergone a recent resurgence in computing referred to as the subject of ontology, and in the insurgence of Grid computing; specifically in the combination of this with the Semantic Web into a so-called "Semantic Grid"; yet another grand title. Different groups, especially philosophers, but also cognitive scientists, psychologists and social engineers, over time, have variously labelled the items being described as "things", "kind", "types", "categories" and so forth; no commitment is made in this thesis to the correctness or otherwise of these labels, or to the use of one of these descriptions as superior to the others: indeed it would seem presumptuous to make such a commitment.

Descriptive systems of various hues are set up for the same purpose of describing things as they appear and interact in the world, or more specifically, the descriptive systems contain surrogates, such as images and text strings, for real world entities. The surrogate is a descriptive form that presents an abstracted view of the real world based on various perceptions. In modern computing, the surrogates themselves are called *metadata* (according to Foster and Kesselman, 1998, "users access data that has been turned into information through the addition of metadata that describes its origin and quality", so metadata is important to information processing), which can used to construct an *ontology*. The combination of metadata and ontology enables information discovery applications to perform such tasks as query expansion or refinement. The ability to "annotate" data, such as text data, with "high-value" markup, and to organise this data into specific structures allows for more powerful information processing applications.

It is with the "thesaurus problem", and specifically metadata and so-called ontology as can be automatically derived from text, that this thesis is concerned. Specifically, we are looking to the language of specialised subject fields to investigate how texts can be used to automatically annotate themselves and other texts. We consider that since terms are the lexicalisations of concepts, an organisation of the terms of a domain provides some indication, artefact, or surface form, of the conceptual organisation of that domain.

### 1.1 Scope

This research takes into account work in a variety of fields. Knowledge representation (a subfield of AI, itself a subfield of Computing), and Linguistics are considered as two general areas of study. The acknowledgement of philosophy through ontology is also to be made. Where
language is concerned, and especially where it is used for communicative purposes, there is both sociological and psychological interest. When systems are built to emulate and model communicative abilities of humans, which requires the encoding of things in the real world, these disciplines all provide some assistance and insight into the process, and deeper research questions within these fields may or may not benefit from the creation of such a system. A goal of this research is to improve the technology with which research into language understanding is carried out – and thus improve the understanding of how “things” are organised, classified and described. This thesis aims to determine some steps and processes for providing a framework for the (automatic) population of specific types of language-based resources for arbitrary domains, including knowledge bases (KBs). The results of this population should be usable by, for example, terminologists and lexicographers primarily, but should also have relevance to the information extraction and information retrieval communities. Resources produced that result from these processes should be usable in ontology engineering and related computing needs, validatable by humans, and machine processable. While these may appear ambitious goals, commonalities between the requirements for many of these needs are quite similar.

1.1.1 Thesis statement

Without reference to 70+ years of work in terminology science, current approaches to representing knowledge in “ontology learning” will falter. Terminology science can contribute to this form of knowledge representation, which longer term will be needed for the Semantic Web/Grid.

1.2 Overview

By exploring issues of extraction of knowledge, based in language and with specific reference to the terminology of specialisms, from different text types such as learned journals, glossaries and online terminology collections, and by applying techniques of knowledge representation and reasoning, it should be possible to demonstrate improvements in the process of knowledge acquisition. Here, the relationship between terminology and knowledge representation in knowledge bases needs to be understood, which can lead to the development of a terminological knowledge base. A goal of this research is to investigate how to populate an ontology representation that could be used for producing a knowledge base for an arbitrary domain. This population would occur from texts and related resources. The resulting system should be validated by an expert, and overall the expert involvement should be restricted to this final phase. The populating process will require identification of key domain concepts, the terms used to
define these concepts and the relationships between terms and concepts. Text understanding may require techniques such as anaphora resolution, word-sense disambiguation, co-reference resolution, and the treatment of metaphor and analogy. These topics are each a subject themselves, and as such are considered beyond the scope of this thesis.

A result of this extraction is the so-called “ontology”. In current parlance, this is a computer-readable representation of information such that it can be reasoned over by a machine. Knowledge bases can be produced by instantiating and creating rules over concepts in the ontology representation. For us, the concepts are referred to by terms, and here we see the importance of the consideration of terminology. The approach to be developed can be perceived as a terminology production line, which will automatically extract terms from collections of text. The terminology produced in this way has the potential to be mapped to a knowledge representation, latterly an ontology. The gap between, say, the definition of a concept in a terminology and its representation in, for example, a frame-based system, remains a substantial one. The information that is (partially) coded in language needs to be mapped to, for example, a predicate form, or a series of attributes and values in order to fully effect this transition. The complete coding required may not be available in the text due to shared common understanding. The challenge of producing ontology from terminology goes beyond the scope of the current work.

While automatic production of such resources is rather an ambitious task, various mechanisms exist which can be used for arranging terms in a terminology and, perhaps, an ontology, if not populating it to some extent. We propose that a standards-based approach is taken to interchange between systems and between interchange formats – addressing the need for an abstract representation to deal with not only data structures, but also varying granularities of data – and that this approach will increase the usability of the end result.

The research presented may help to look afresh on issues raised in knowledge management – how knowledge of an organisation is managed, starting from diffuse ideas formed from internal knowledge, passed on, socially, to others for comments/criticisms, made explicit and external to still others, and used in conjunction with other knowledge – combination. It is possible to consider that the process of a Knowledge Engineer interacting with an Expert relates to a mode of Knowledge conversion called Externalization, and the process of extracting salient information from a number of texts and related resources for populating an ontology is Combination “combining different bodies of explicit knowledge” (Nonaka and Takeuchi 1995, pp62-73). This form of information extraction would therefore appear relevant to the population of intelligent systems, to the development of computer systems and to the Knowledge Management community. Indeed, the need for a common ontology surfaces in literature on Knowledge
Management where there is talk about “a thesaurus of common terms” (Webb 1998, p34), “a common vocabulary .... ‘site vocabulary’, ‘metadata’, ‘lexicon’, ‘attributes’ or ‘bits about bits’” (Applehans et al 1999, pp77-79) and the modern knowledge management gurus, Nonaka and Takeuchi talk about the “common language” within an enterprise, and, citing a case study, observe that “Speaking a common language and having discussions can assemble the power of the group. This is a vital point, even though it takes time to develop a common language” (Nonaka and Takeuchi 1995, p99). The language used for communication is therefore of substantial importance to the builders and analysts of systems in the various fields of computing noted so far.

From the above discussion, the need to study the “language” being used within domains and enterprises to determine the common language that exists in such endeavours becomes apparent. For a computer to understand language, it should have an appreciation of the mechanisms used within the language. Within specific domains, the common language identified above can be considered as the so-called Languages for Special Purposes (LSP), which are heavily laden with terminology as opposed to Languages for General Purposes (LGP) or simply General Language (GL).

Language can be thought to exist on several levels. Luger and Stubblefield recognise 7 levels defined by linguists: prosody; phonology; morphology; syntax; semantics; pragmatics; world knowledge (1993, p339). Engels and Bremdal have phonology as their lowest level with a lexical level before morphology, combine pragmatics with common sense (world knowledge) and place discourse between this and semantics (2000, pp34-35). Fitting work on language into a single category within these schemes would be a difficult task as elements of more than one would be required for most language tasks. The computer needs some “knowledge” of these language levels in order to understand human language to some degree. Primarily, our work is focussed at the lexical level, but strongly heeds the need for semantics and world knowledge to be taken into account.

1.2.1 Research Questions

In relation to the foregoing, we propose the following research questions that will be answered to some extent in the course of the work presented:

1. How different are languages of specialisms from general language?

2. What is the link between language and knowledge? Does language use accurately reflect what is known?

3. Is it possible for a machine (computer) to develop an understanding of language?
4. Can the population of “domain ontologies” for use in the Semantic Web benefit from work carried out in the population of terminology collections on the extraction of information from text?

5. Can studying the link between language and knowledge contribute to the scientific community’s ability to exchange data?

6. Can studying the link between language and knowledge lead to improvements in the tools used in language research?

7. Is it possible to combine techniques in terminology extraction with those in knowledge representation?

8. Can we develop a computer system to assist in the semi-automatic construction of concept systems or ontologies: hence, can we bootstrap a domain-specific ontology from a text collection of consensus within a domain of discourse?

These questions will form a basis for the conclusion and discussion in Chapter 5.

1.3 Research Contribution

The work presented may be considered to make contributions to certain areas of research. These include:

- **language**: consideration of general features of language and investigations of systematic differences between specialist and general languages may provide a means to advance the research into LSP/LGP.

- **terminology**: in the consideration of LSP/LGP, we have contributed a method for automatic identification of terms. We also bridge the divide between prescriptive and descriptive approaches to terminology by consideration of how to construct collections of these terms in a standards-conformant way, while populating the collection in a descriptive manner through automatic extraction from text.

- **ontology**: consideration is made of conceptually organized collections of terms and how they can be presented in an ontology such that knowledge based systems could be developed using them as a basis.

- **knowledge management**: production of a “common vocabulary” is an important component of managing the knowledge of any kind of organization, which extends most
challengingly to the notion of transient virtual organizations (VOs) so popularly referred to in literature about Grid computing. Here, the notion of Knowledge Grids is interesting also.

- **computing**: with reference to Grid computing above, we can see a progression from terminology through to Grand Challenges in computing itself. One such challenge to which this work may contribute is “Memories for Life”, in which such memories may be searchable by others using their own terminology. So-called “blogs” are a current text-based example.

The overall contribution has been the *study of theories* of language that result in the proposal of a *method* by which to produce *candidate ontologies* via the construction of candidate terminology. A system that demonstrates this method has been *prototyped*, the results of which have been the subject of *evaluation*. The system is available for further evaluation, and may be useful for future investigations of a related nature: perhaps in answering some of the future research questions that will be proposed in Chapter 5.

There have been various contributions to the scientific community during the course of this work, including *publications*, *software prototypes* and *contributions to standards* that have been evaluated in a variety of ways. The majority of publications have been peer-reviewed; software prototypes have been evaluated within the remit of collaborative research projects funded by UK and EU agencies, and contributions to standards have been reviewed by panels of experts across more than 20 countries of the world.

### 1.3.1 Collaborative research

Specific contributions can be considered in reference to a number of research projects undertaken at the University of Surrey. These projects emerged from earlier work at Surrey on Knowledge-based systems and text/terminology management. The timeline for these projects is shown below in Figure 1. The dotted horizontal line approximately denotes the beginning of the author’s involvement with these projects. The lowermost items refer to projects to which contributions may be made subsequently: REVEAL will begin in September 2005, while TONES and LIRICS are submitted proposals.
Chapter 1

1.3.2 Software Prototypes

A variety of prototypes have been developed within the scope of this work to demonstrate ideas on a practical level. These have included applications for the semi-automatic extraction of terminology and other information, in a variety of structures, from text, both on file systems and on the Web. These applications have been used by other researchers, essentially validating the ideas in the course of their own works. Certain of these prototypes have become established components of the System Quirk software suite, which the author has been responsible for maintaining since 1996. A first release version of this software for PC was distributed at the start of 1998, and since then over 1500 copies have been distributed. The author has also installed Unix versions of this system at other international institutions and within a Grid computing infrastructure.

One prototype, Tracker, developed in conjunction with another researcher, has been previously submitted for both the European Union IT prize, and the British Computer Society (BCS) prize. Although neither submission won, the prototype has been evaluated at these levels.
Further prototypes, and the research projects they were developed in, include:

<table>
<thead>
<tr>
<th>Prototype</th>
<th>Research Project</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Corpus Manager (VCM)</td>
<td>TRANSTERM</td>
<td>re-engineered application for organisation of text corpora by pragmatic attributes</td>
</tr>
<tr>
<td>Plain English</td>
<td>SAFE-DIS</td>
<td>application for linguistic simplification of technical documents</td>
</tr>
<tr>
<td>Colloqator</td>
<td>INTERVAL</td>
<td>application for generating collocation statistics</td>
</tr>
<tr>
<td>CorpRand</td>
<td>INTERVAL</td>
<td>component application for enabling cross-corpus comparison</td>
</tr>
<tr>
<td>Percentile</td>
<td>ACE</td>
<td>component application to generate ranked clusters of words</td>
</tr>
<tr>
<td>Endings</td>
<td>ACE</td>
<td>component application to derive morphological variations</td>
</tr>
<tr>
<td>Tracker</td>
<td>ACE</td>
<td>application for mining information from free-text sources (websites)</td>
</tr>
<tr>
<td>News Analysis Module (NAM)</td>
<td>ACE</td>
<td>system for processing free text and providing information about key economic events; combination of other ACE technologies and components developed elsewhere</td>
</tr>
<tr>
<td>Validation Toolkit</td>
<td>INTERVAL</td>
<td>system for validating terminological resources</td>
</tr>
<tr>
<td>OntoMapper</td>
<td>SALT</td>
<td>application for mapping between subject field classification systems</td>
</tr>
<tr>
<td>DXLT Validation</td>
<td>SALT</td>
<td>re-engineered and improved functionality of Validation Toolkit; XML conformance</td>
</tr>
<tr>
<td>Fundamentals Extractor (FundEx)</td>
<td>GIDA</td>
<td>application for mining information from semi-structured sources (websites)</td>
</tr>
<tr>
<td>FormBuilder</td>
<td>SOCIS</td>
<td>component application for collection of images and textual information from Scenes of Crime</td>
</tr>
</tbody>
</table>

Other prototypes have been derived from those above, for example by reusing elements of the GIDA Fundamentals Extractor for mining X-ray photospectroscopy data and for harvesting Malaysian-English law translation information from the web. Other applications, including software for text summarisation and text alignment, have been re-engineered and offered for free usage on the Web as client and server applications.

The re-engineering or development and subsequent use of these prototypes, and the continued success of System Quirk have proved the importance of the production and distribution of such prototypes. These prototypes are considered, by the author at least, to be at least as important as the creation of conference or even journal papers, as they provide the means for others to evaluate the efficacy of the approaches outlined in a manner that a publication does not. The relationship between prototypes directly related to research projects may be more easily understood by reference to Figure 2 below.
According to the International Review of UK Research in Computer Science¹ (Fred B. Schneider and Mike Rodd, editors), the UK is a world leader in speech engineering and computational linguistics. The report further identifies the importance of building prototypes in order to explore and evaluate ideas. In the course of solving an existing problem, a computer scientist builds prototypes, as abstractions of reality, and it is in the building and evaluation of these prototypes that further specific technical problems are likely to become apparent which may require further approaches and abstractions to be applied. The prototypes then become the object of study, and the assumptions made in its creation may become apparent. To overcome such problems, further research into alternative techniques may be necessary. As the authors of this report note: “it is prototypes that launch paradigm shifts”.

1.3.3 Publications

The author has contributed to over 20 papers in Journals and Conference proceedings that are directly related to the research presented in this thesis. These are listed in Appendix A.

1.3.4 International Standards

Recognition of contributions to the scientific community resulted in the author being invited to participate in the process of international standardisation, representing the UK through the British Standards Institution as a Principal UK Expert to the International Organization for Standardization Technical Committee 37 (ISO TC 37) on “Terminology and Other Language Resources”. Work in International Standards has included contributions to the drafting of ISO 16642, the Terminological markup framework (TMF), including the content of Annex D, (published 15 August 2003). Other contributions have been made to revisions of ISO 12620 for data categories, ISO 12615 for bibliographic references, and to subsequent parts of ISO 639 for language codes that may be formed from work being done in the production of British Standard BS 8639 based on the Linguasphere Register of world languages. The author has participated in various ISO TC 37 meetings including plenary meetings, since 2000, and has attended numerous BSI meetings at BSI headquarters in London to discuss the UK perspective on the emerging standards. The author has also been Head of the UK delegation to both the setup meeting of ISO TC 37 subcommittee 4 on Language Resources that took place in Las Palmas, Gran Canaria, and a plenary session in Lisbon, Portugal in 2004. There is hope that some of the work presented in this thesis may constitute parts of revisions to existing and to possible future standards within this forum.

It is, perhaps, unfortunate that the principal measure of research at UK Universities, the Research Assessment Exercise, does not explicitly list standards as an available output type, and that explicit lists of authors are generally not contained in standards. Of the 18 available categories for the RAE, perhaps such contributions fit into “Other”.

1.4 Structure of Thesis

The thesis comprises 5 chapters including this chapter dealing with introductory material. Chapter 2 discusses the seemingly inextricable link between what we variously term “language” and “knowledge”. How we think about language, and how we think in language are two aspects of this discussion: in order to have knowledge of something, we should be able to express it in howsoever a form. Although modalities such as gestures can be used for certain types of expression, and communication can be made pictorially, it is using language that we are able to decide whether the communication has been effective. A principal difficulty of language is that we must use language to describe it.
Scientific texts may be useful as a basis for discovering the potential structure of the scientific ontology through consideration of the (specialist) lexicon. For a scientific subject to exist, it appears that it should have a concomitant set of texts, a quantity of which will have been peer-reviewed, existing within scientific journals and books. With reference to Ogden and Richards' meaning triangle, we consider the link between language and the knowledge being expressed in the language, and how this link is exemplified in the terminology of specialisms.

Chapter 3 discusses practical aspects of language use, with specific reference to terminology. We consider the very essence of terms and terminology collections with reference to the definition of a term as stated in an international standard, and other standards that are required to understand this essence. The need for subject field classifications, language identifiers and consideration of semantic shift is made with a view to understanding how to formulate terms. We then consider various metrics that may be useful for automating the process of getting terms: terminology extraction. Language is studied from a broad perspective, with reference made to George Kingsley Zipf and his principle of least effort. Zipf proposed that, essentially, people are lazy in their use of language, so are more likely to repeat words than to use them only once. Zipf's Law is considered in the context of both general language and specialist corpora to try to discover systematic differences. Subsequently, frequency values, weirdness and Smadja's statistics for patterns of collocation are discussed for identifying terms. These terms can be organised hierarchically by inclusion. In understanding the essence of a term, and identifying how terms exist in text, we then consider how this can be used to produce a) terminology collections and b) ontologies. For a), we look at the international standards once more: ISO 12620's data categories and ISO 16642's method for defining terminology markup languages, with the aim that a terminology management system should be able to import such a defined language. For b) we consider ontology markup languages, and how it may be possible to move between terminology and ontology. The purpose of Chapter 3 is to provide a basis for developing a system, based on the understanding we have of terminology, ontology, and the means by which we can automatically produce these.

Chapter 4 proposes and demonstrates a method based on the terminological analysis of text collections that uses an arbitrary collection of text in a specialist domain as its input and automatically extracts a candidate terminology / candidate ontology. The candidate ontology can be modified using freeware ontology editors such as Protégé, and subsequently this provides a means by which to instantiate knowledge bases. While the international standards for computer collections of terminology are available, we used an RDFS encoded export format for use with Protégé. The system is the result of considerations made in Chapter 3, and a resulting ontology is intended for evaluation by a subject expert.
Chapter 5 concludes the thesis, presenting possibilities for future work and evaluation that emanate from the work described in the previous chapters.
2 Knowledge, Language and Computing

In this chapter we discuss the seemingly inextricable link between what we variously term "language" and "knowledge". How we think about language, and how we think in language are two aspects of this discussion: in order to have knowledge of something, we should be able to express it in howsoever a form. Although modalities such as gestures can be used for certain types of expression, and communication can be made pictorially, it is using language that we are able to decide whether the communication has been effective. If we consider ancient cave drawings depicting animals being killed, were these used to communicate an effective way to kill such an animal, or as a means to celebrate the kill? Deciphering such visual scenes without further information such as could be provided using language is not as straightforward as one might expect, although imagery can be used in place of language in well-understood subject fields such as mathematics, or in chemical formulas that carry with them an understanding of the atomic properties of specific elements such that \( \text{H}_2\text{O} \) represents not only a number of elements, but a physical orientation of these elements dictated by their attractive/repulsive forces. The words, and other symbols, used would seem to represent some surface form of knowledge that is being used to communicate some idea: Ogden and Richards refer to the study of a tribe who appeared to have 5 words for the action of somebody tapping on a table (1960, p77-78). The caveat they offer is that the act had been interpreted in 5 different ways, each resulting in a particular word relating to "hardness", "tapping", "material", "covering" and, finally, the word for the table itself. Ogden and Richards' so-called "meaning triangle" is referred to by many as the essence of semantics. The authors insist that meaning cannot be considered by itself, it requires a theory of signs (semiotics): to consider the meaning and its relationship to language they require us to have some theory of language.

A principal difficulty of any theory of language is that we must use language to describe it. Dictionaries and similar resources that define the words of the language use the language in the definitions. It is not possible to use a dictionary without some prior understanding of language. Where we say, for example, that we are using the English language, we have used words of the English language to identify the language we are using, so we even need the language to describe the language. One almost wonders how we deal with language at all, and yet it is something that can be learnt by small children. Dictionaries being artefacts of the language, it must be the case that the language exists before the dictionary is created, and hence before we can have a theory of it: the 2300 pages of Samuel Johnson’s Dictionary of 1755 catalogued the use of language at that
time, recording an opinion of the meanings that people ascribed to words used then. Language had existed for some significant time previously: William Caxton printed a book in English in 1474, some 280 years before a dictionary was deemed necessary; written languages had been around for a significant time previously. How did people know what words of English meant before the dictionary? And, significantly, how did they know about meaning before they had developed a theory of language? Without the means to understand meaning, how could the Royal Society have published the first scientific journal (Philosophical Transactions) in 1665?

Before the consensus document, the dictionary of English, was created, scientists such as Newton referred to specific ideas by denoting that they had named it somehow:

“The Light whose Rays are all alike Refrangible, I call Simple, Homogeneous and Similar; and that whose Rays are some more Refrangible than others, I call Compound, Heterogeneous and Dissimilar”

This passage from Newton’s Opticks, heralded by Halliday and Martin as being at “the birth of scientific English” (1993, p57), shows how he declared words and how he was using them – what they meant to him - within the body of text he produced: there was something that needed to be named, and so a name was selected and documented. In the mind of the scientist, the vocabulary and what they meant by it, was being developed, and some of this development was being recorded: perhaps Newton could be perceived as an early lexicographer, compiling a dictionary within the text? 250 years on from Samuel Johnson, we contend that this process of naming in text, and of course in speech, continues: the Oxford English Dictionary make frequent announcements of words that are considered to warrant entry into this gargantuan resource and, indeed, scientists continue to make discoveries and provide names for these discoveries that people may agree or disagree with. People who do not necessarily have a theory of language or a theory of signs are able to mean something using them. The same arguments hold for children being taught language: only after they have learnt how to use it are they taught a theory (grammar) of it.

The description of the things that are named, then, occurs in part in written language, perhaps more so in specialised subjects than in general language usage. It is the texts of scientific subjects that form a principal resource for many aspects of sciences, especially pedagogic ones. For a scientific subject to exist, it appears that it should have a concomitant set of texts, a quantity of which will have been peer-reviewed, existing within scientific journals and books. A specific subject may evolve from an existing subject, in which case its heritage is the area from which it evolved, but it doesn’t truly exist until it has acquired a reasonable body of literature.
A single text within a scientific subject can only be considered meaningful if it is either complete in and of itself, or if we are able to refer to standards accepted within the subject, for example: measures such as length defined by reference to a specific item; chemical symbols explained through the periodic table, related defining works and so on. Within such a text, there exist granularities of *knowledge fragments*, from the full text itself, to other texts it may refer to, to the symbols that it is made up of, including terms, to the ways in which these symbols relate the text to other texts indirectly. As a number of authors have pointed out, the ‘natural’ habitat of terms is texts, where they are likely be in the company of other terms. Scientific texts will contain scientific terms to which the authors are (ontologically) committed, which will have reference to a system (ontology) of such terms which allows their meaning and relationship to other terms to be determined, whether or not the ontology exists in a formal sense. This ontology of the subject will reflect the current understanding of the scientific subject: the terms represent the surface form, or lexicon, of the subject ontology, while the subject ontology may itself be a surface form of the “one true” ontology (in the philosophical sense). Ogden and Richards’ meaning triangle would appear to be evident here. The treatment of scientific texts as a basis for discovering the potential structure of the scientific ontology, through consideration of its (specialist) lexicon, allows us to consider various possible results depending on the knowledge fragments under consideration.

It is often not easy to discuss open-ended and interconnected topics like knowledge and language, and then try to turn the discussion to a specific and/or practical application. But as ‘knowledge’ resources will be discussed (Chapter 3), and the realisation of a text/term analysis system (Chapter 4), it is important to outline what, in this thesis what has been understood by knowledge and language (2.1). This will lead to ontology (2.2) and terminology (2.3) and onto the application of these discussions: intelligent systems (2.4).

**2.1 An elaboration of the ‘knowledge’ and language**

**2.1.1 Knowledge**

The first, and often asked, question is: What is knowledge? Knowledge, and its related aspect of wisdom, form the basis for philosophy: “The love, study, or pursuit of wisdom, or of knowledge of things and their causes, whether theoretical or practical” (OED). Studying this “knowledge of things” is therefore a philosophical study, and so a study of how knowledge is conveyed presents itself as a philosophical question that involves explaining this knowledge. Philosophers such as Wittgenstein have been accused of reducing the study of existence to the study of language,
although this is perhaps a somewhat extreme interpretation: Wittgenstein’s “Philosophical Investigations” presents an understanding of how things exist as the way that we describe them, and discusses the problem of description, a problem that also concerned Socrates and Aristotle. That we have problems with questions such as “how good is good” evidences the difficulty we have in expression: perhaps the language we use for our communication is just not good enough? Although it is possible to describe knowledge, if we cannot do so without language we have no choice but to study how language relates to knowledge and more widely to existence.

The OED provides a variety of definitions for knowledge around a theme of “The fact or condition of knowing”. In following the various derivations of “know”, we discover knowledge to be contingent upon the having of facts, or being informed about some thing. Somebody who says “I know what the weather is like in China” we presume to be in possession of some facts regarding said weather, however if the same person says “I know what the weather will be like in China tomorrow”, we cannot be certain about the possession of facts. The full form of such a statement should be something like “I have some information from which I believe I can predict with some confidence what the weather will be like in China tomorrow”. Of the two initial statements, which constitutes knowledge? Can the answer be both? Certainly the processes involved appear to be similar: in the first case, possibly the person has been given information that it is currently raining in China, but is this then information and not knowledge? Perhaps the information was not complete and the person has reasoned over hearing some other person in China saying “I had to walk home in the cold and wet” and concluded that there must be rain and low temperatures. For the later statement, has this person seen a prediction elsewhere regarding tomorrow’s weather, or are they in possession of meteorological information from which they have reasoned towards this statement, or perhaps something intuitive? In both cases, we can only determine whether we are dealing with the relaying of information, or whether the information has been processed and suitably presented. Herein lies a problem in the definition of what knowledge is, since the boundary between information and knowledge is not as clear as many would have us believe, and this lack of clarity is evident in what has come to be known as “knowledge management” during the last decade.

The boundary between information and knowledge, with specific reference to the difference between information management and knowledge management has led to knowledge management being dismissed as “nonsense” (Wilson 2002). For Wilson, knowledge “involves the mental processes of comprehension, understanding and learning that go on in the mind and only in the mind”. He points out that when we convey such knowledge, through any medium or mode of communication, what is conveyed is then “information’, which a knowing mind may assimilate, understand, comprehend and incorporate into its own knowledge structures”. Wilson
additionally makes a distinction between data “simple facts”, and information “the data are embedded in a context of relevance to the recipient”. Proponents of knowledge management as a discipline also make a distinction between data, information and knowledge: data are uninterpreted signals; information is data equipped with meaning; knowledge is the whole body of data and information that people use to carry out tasks and create new information (Schreiber et al 2002 pp3-4). The contrast here is between knowledge being internally stored in the mind, and knowledge being information equipped with meaning that can be used for some purpose. In the latter case, does this mean that a dictionary is knowledge? Certainly it is information equipped with meaning that can be used for some purpose.

The position we take for this thesis is that this “knowledge” is merely information, for which humans require knowledge in order to correctly interpret it. The former descriptions of data, information and knowledge are more appropriate here, and the latter two can be more comfortably contrasted with the meaning triangle, a graphical interpretation of Frege’s distinction between concept and referent, introduced by Ogden and Richards. It shows the relationship between signs such as symbols (or words), mental concepts and real things. The dotted line denotes association rather than a direct link. Meaning triangles elsewhere use “representation”, by which terms and other designations are meant, in place of symbol and “object” in place of thing, although the description of object generally starts with the word “thing”.

![Meaning Triangle](image)

**Figure 3: Meaning triangle (Ogden and Richards 1923)**

The meaning triangle can be thought of as associating words (symbol) via mental concepts (thought or reference) to the thing that the symbol can be said to stand for (referent).
meaning triangle is, perhaps, difficult to demonstrate for a mythical creature such as a unicorn, since there is no real world thing the word can signify, although it does represent a mental concept. If we accept the idea that the study of language and the study of ontology can be considered intertwined in the majority of cases, perhaps we can reduce the problem of ontology to the previously unsolved problem of language, or rather the problem of creating an ontology becomes a problem of language understanding and analysis. The human importance of giving names to concepts can be taken to extremes by considering words for things for which words have not been invented (as seen Adams and Lloyd 1983): this completes the meaning triangle in a surprisingly satisfactory fashion.

Maedche, strongly influenced by the meaning triangle, provides an illustration of the use of ontologies in communication, repeated here in Figure 4 below, that shows several layers with humans and machines interacting with symbols and structures to share information about things in the real world (Maedche 2002, p15). This shows the effect of bringing together two meaning triangles with different senses of the word “Jaguar”, which evoke different “things”.

Maedche’s example shows two objects in the real world, one an animal, another a car, both referred to by the symbol “Jaguar”. The reference is correct in both cases, but may lead to problems of understanding. The difficulty can be overcome through consideration of related information (that could be contained within an ontology) about the two items: although both have four items that enable them to move, in one case these four items are wheels, in another they are legs. This example shows that the car and the animal may share particular characteristics when
considered from a certain perspective: for example, both require some form of energy to assist in moving, and both can make a kind of growling noise. Furthermore, if either are “broken”, they are taken to a specialist who is (hopefully) able to fix them. Both have a body, and characteristics of length, weight, and power, and although a brownish-yellow rosette-marked car may be unusual, a “Black Jaguar” could again refer to either car or animal. Both may be found in South America, and in both cases it is possible to consider the phrase “I intend to purchase a Jaguar”. To determine which concept is being referred to by a single word in combination with a few characteristics, it is therefore important to find the appropriate distinguishing characteristics. For each of these various characteristics, we have used language as the means by which to convey them, to provide some means by which both can, effectively, be distinguished by other humans during communication. This seemingly inherent link between language and knowledge will surface throughout this thesis. The forefathers of modern so-called “knowledge management” Nonaka and Takeuchi would doubtless refer to language as the externalization of knowledge (its explicit form). Their notion of internalization is perhaps difficult to quantify, since there is no physical trace available of this beyond experiments involving the tracing of electrical activity in the brain.

2.1.2 Language

Schank and Abelson recognise that language is an area of overlap between psychologists and computer scientists: “The two orientations intersect when the psychologist and the computer scientist agree that the best way to approach the problem of building an intelligent machine is to emulate the human conceptual mechanisms that deal with language” (Schank and Abelson 1977, p1). Schank and Abelson emphasise the importance of building computer models of human behaviour to better understand human behaviour, and to gain some insight into how both familiarity of situations, and the underlying intentions of the human actors, is important to the understanding of daily events. For Hebb, language is a purely human characteristic: although parrots can repeat words, they cannot manipulate them effectively (Hebb 1966 p297). Learning, for Hebb, is the process of setting up conceptual types by naming, for example, individuals (Hebb 1949 p116). Hebb refers to an “economy of language” along with an economy of thought, learning and understanding and that a familiar term is more easily recalled than a new one (ibid). This economy of language would seem to imply that language is dominated by common words or phrases, an implication that we will see to be the case later.

According to Winograd, “When a person sees or hears a sentence, he makes full use of his knowledge and intelligence to understand it” (Winograd 1972, p1). Schank and Abelson’s “scripts” are an attempt to develop a theory for how a computer can “fill in the gaps” between
surface descriptions of occurrences, and the reasons for those occurrences by applying encoded knowledge and intelligence, in a manner which a human does. The interactions between separate sentences are studied to identify the common theme in the sentences. To comprehend these utterances requires the agent – human or machine – to comprehend the situation implied by the words and phrases employed. Winograd explains that this relies on knowledge of grammar, words, the sentence context and the understanding of the subject. The author makes a number of assumptions about comprehension between both the “sender” and “receiver” of the message, and notes that a lack of such shared understanding can be problematic. Winograd argues that for the computer to understand language, it needs some understanding of grammar, semantics and reasoning. Language is, he notes, “one of the most complex and unique of human activities, and understanding its structure may lead to a better theory of how our minds work” (ibid, p6). In developing his language understanding systems for the blocks world, Winograd models the interaction between the speaker (writer) and the listener (reader) as an interaction between two knowledge-based systems. Both systems have a three-tier knowledge base: (i) a knowledge base of linguistic knowledge comprising lexical, grammatical and semantic facts and rules; (ii) a knowledge base of the context of interaction and (iii) an encyclopaedic knowledge base of world knowledge. For Winograd, these two knowledge-based systems help facilitate interaction between two intelligent agents.

There is much debate about how correct language is, for example, what constitutes English. Linguists consult with so-called native speakers of English – those brought up by carers, especially parents, whose first language is English - to determine the acceptability of a word, or the correct grammatical construction of a sentence or meaning of a certain word. Randolph Quirk argues that “assessments by native speakers of relative acceptability largely correlate with their assessments of relative frequency” (Quirk 1985, p33). As an historical parallel to this, Sharp cites the work of Edmundson and the borrowing of a principle from information theory where the significance of a word should be a function of rarity, rather than frequency of appearance; rarity means the infrequency of appearance of a word in general usage, not in a given document (Sharp 1965). For Sharp, “this principle of ‘relative frequency’ would give greater discrimination than Luhn’s method in that it would distinguish between common and rare words used with equal frequency within a document, [...] and would also eliminate general words normally rejected by the use of ‘stop lists’.” This combination of frequency and rarity is similar to the Weirdness coefficient (for example, Ahmad 1999b), and provides a contrast to the latter-day Information Retrieval literature where the so-called Term Frequency / Inverse Document Frequency (TFIDF) metric is prevalent.
Chapter 2

For some, “Language is a [...] major vehicle for the transmission of information [...] any nontrivial word that occurs sufficiently frequently must be a valid content indicator, or it would not be used so often” (Sparck-Jones and Kay 1973, p2). These authors try to understand how human beings communicate a presumably infinite variety of ideas using only the finite, albeit large, set of words of a language, and argue that “meaning is an essential part of the very notion of language, and the claim that one can study any part of language in isolation from meaning is incoherent” (p32). This is encouraging for us since terminology is concerned with how terms convey meaning in a text. Winograd notes that “The decision to consider syntax as a proper study devoid of semantics is a basic tenet of recent linguistic theories”, but argues that this is an example of asking the wrong question and that understanding meaning is essential. Wolterstorff provides an extensive treatment of ambiguity within simple statements depending on the intended meaning, discussed in relation to notions of predicates, properties and the so-called “Universals”, although he has difficulty drawing a distinction between universals and non-universals, and argues that his entities “cited as examples of universals are in fact kinds (types)” (Wolterstorff, 1970, p235). Wolterstorff makes a direct link between language and philosophical notions.

Chomskian theories of universal grammars are less of interest to us: although a conceptual organisation of knowledge is considered, the reduction of sentences of natural language to their canonical forms is not our focus. The conceptual organisation itself, however, certainly is. If there are means by which we organise both concepts and the language by which we refer to those concepts, we have the basis for some form of language processing that may have utility elsewhere. To do so, we need to consider conceptual systems, and the current flavour of “ontology”, including how these systems of concepts are built, populated, what they contain, and how this relates to language.

2.2 Ontology

The Encyclopaedia Britannica defines ontology as:

“the theory or study of being as such; i.e., of the basic characteristics of all reality. Though the term was first coined in the 17th century, ontology is synonymous with metaphysics or "first philosophy" as defined by Aristotle in the 4th century BC”.

And Webster’s Collegiate Dictionary defines ontology as:

1 : a branch of metaphysics concerned with the nature and relations of being

2 : a particular theory about the nature of being or the kinds of existents
The origins of Ontology are deeply embedded in philosophical studies. Ontology, as a branch of philosophy, deals mainly with the importance or otherwise of some concepts as compared to others. Such discussions are essentially a critique of extant theories in philosophy. The question for us is this: what is the relevance, and consequently utility, of ontology for building information systems? The key notion that we will consider is the link between language and ontology. Ontologies, from a computational perspective, represent artefacts of the world viewed in some manner for some purpose. They are built from key notions about the world that are required for that purpose. For the purposes of computing, we wish to determine what can be represented, and how, such that a machine can use it. We therefore seek an implementable form for the ontology, and an understanding of what the ontology should contain. We should first develop some understanding of what is meant by "ontology".

2.2.1 The use of the term/concept “Ontology”

The use of the term/concept “Ontology” is all-pervasive. This can be discerned from titles of publications selected from the University of Surrey library catalogue and the British Library Public Catalog. The origin is philosophy was apparent in Philosophical Essays from 1733 and 1734, through discussions of works of various philosophers – Kant, Buber, Heidegger and Hegel – to more modern appraisals of Whitehead’s Ontology (John W Lango 1972) and Sartres Ontology (K Hartmann 1966). The pervasiveness ranges from art and architecture to human sciences, and from politics to physical, biological and computing science, and can be seen in Table 1 below.
<table>
<thead>
<tr>
<th>Domain Title</th>
<th>Domain</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philanthropy of mathematics: structure and method</td>
<td>Mathematics</td>
<td>Philanthropy of mathematics: structure and method</td>
</tr>
<tr>
<td>Learning for the semantic web</td>
<td>Computer</td>
<td>Learning for the semantic web</td>
</tr>
<tr>
<td>Knowledge processing for global information systems</td>
<td>Computer</td>
<td>Knowledge processing for global information systems</td>
</tr>
<tr>
<td>of computer, philosophy, law and the shape of intellectual property</td>
<td>Computer</td>
<td>of computer, philosophy, law and the shape of intellectual property</td>
</tr>
<tr>
<td>of the official conceptions of disease</td>
<td>Computer</td>
<td>of the official conceptions of disease</td>
</tr>
<tr>
<td>of modern physics</td>
<td>Computer</td>
<td>of modern physics</td>
</tr>
<tr>
<td>The ontology of physics</td>
<td>Physical sciences</td>
<td>The ontology of physics</td>
</tr>
<tr>
<td>of physical and operational</td>
<td>Physical sciences</td>
<td>of physical and operational</td>
</tr>
<tr>
<td>of sociology</td>
<td>Social sciences</td>
<td>of sociology</td>
</tr>
<tr>
<td>The ontology of sociology</td>
<td>Social sciences</td>
<td>The ontology of sociology</td>
</tr>
<tr>
<td>and the semantic flexibility of temporal system in German and English</td>
<td>Social sciences</td>
<td>and the semantic flexibility of temporal system in German and English</td>
</tr>
<tr>
<td>and the logical analysis of language</td>
<td>Social sciences</td>
<td>and the logical analysis of language</td>
</tr>
<tr>
<td>in natural language</td>
<td>Social sciences</td>
<td>in natural language</td>
</tr>
<tr>
<td>The ontology of construction</td>
<td>Engineering</td>
<td>The ontology of construction</td>
</tr>
<tr>
<td>of English</td>
<td>Engineering</td>
<td>of English</td>
</tr>
<tr>
<td>Knight of the White Cross: William Blake and the English</td>
<td>Engineering</td>
<td>Knight of the White Cross: William Blake and the English</td>
</tr>
<tr>
<td>in art</td>
<td>Engineering</td>
<td>in art</td>
</tr>
<tr>
<td>Philosophy of the white expression: Chaucer</td>
<td>English Literature</td>
<td>Philosophy of the white expression: Chaucer</td>
</tr>
<tr>
<td>Essays on art and music</td>
<td>English Literature</td>
<td>Essays on art and music</td>
</tr>
<tr>
<td>The world viewed: Reflection on the ontological</td>
<td>Art</td>
<td>The world viewed: Reflection on the ontological</td>
</tr>
<tr>
<td>AM/Archaeology</td>
<td>Domains</td>
<td>AM/Archaeology</td>
</tr>
</tbody>
</table>

Table 2: Concordance analysis of "ontology" in titles of works in the Surrey and British Library catalogues as they relate to specific subjects.
Each domain appears to have its ontological description - e.g. *Ontology of* film; art; economics; socialism; modern physics; construction; cyberspace; the clinical conceptions of disease. Often the term is used together with its superordinate, i.e. philosophy, or with other branches of philosophy -- epistemology, phenomenology. “Ontology” becomes the head (noun) in compounds like temporal ontology or feminist ontology. In computing the term is used as a modified: ontology-based query processing or ontology learning.

Ontology, it appears, is the study of existence of entities, abstract and concrete. Ontology may be described as a way to formalise the world by describing the objects, features and relationships that occur in the world. The world can be described in a natural language or other semiotic systems, and may be represented using, for example, predicate calculus for computational purposes. Historically, ontology seems to deal with questions of existence - metaphysics - and with the boundary between reality and representation. Lango has argued that: “Ontology is the study of being or beings. But what is being? [...] The beings (entities) of an ontology are divisible into types (e.g., material bodies and minds)” (Lango 1972, p1). Whitehead’s ontology, according to Lango, contains actual entities, prehensions, nexus, subjective forms, eternal objects, propositions, multiplicities, and contrasts. Lango introduces a somewhat obscure relation between entities – *synonty* – such that things are defined as ‘being’ only in terms of other things (p16). The definition of “red” is used to exemplify this relation: *define red as* “related to blue via relative position on the color spectrum, [...] red is related to color by the relation of species to genus [...] red is related to straight line by the relation of color to shape [...] red is related to “any” by the relation of value to universally quantified variable (instantiation)” (p28). This is quite a peculiar argument since everything else has to exist before anything else can, which by extension means that nothing is able to exist. Wittgenstein’s discussions about the difficulty of understanding how language is processed into an understanding (mental concepts) identifies the difficulties with a computational understanding of natural language and the deeper need to study human aspects of communication. Wittgenstein’s “five red apples”, a small fragment of language, evokes for him a “looking up” of colour on a colour chart, numbers on a number chart, and apples on a chart of fruit, in order to derive comprehension. Even for this small fragment, the quantity of shared understanding required in its interpretation is significant.

The question of “existence” is not limited to the determination of “why are we here”, but can be applied to various other questions of what reality is. For example, Cavell addresses questions of existence related to photographs (Cavell 1971):

“A photograph does not present us with “likenesses” of things; it presents us, we may want to say, with the things themselves. But wanting to say that may well make us ontologically restless. [...] Obviously a photograph of an earthquake [...] is not an
earthquake happening. [...] We might say that we don’t know how to think of the connection between a photograph and what it is a photograph of. The image is not a likeness; it is not exactly a replica, or a relic, or a shadow ...” (p17)

and the distinction between photographs and artworks such as paintings:

“You can always ask, of an area photographed, what lies adjacent to that area, beyond the frame. This generally makes no sense asked of a painting. You can ask these questions of objects in photographs because they have answers in reality. The world of a painting is not continuous with the world of its frame; at its frame, a world finds its limits”. (Cavell, 1971, p23)

Similar arguments can be made regarding an audio recording – a recording comprises “sound”: here we have the possibility of identifying an item that was recorded by the “sound” it makes. Cavell’s example is that of an English Horn, which can be identified visually or audibly. Audible identification means it can be identified without actually being present, indeed it can be identified as being a component of the overall sound reproduction, without being able to associate it with the original item that produced it. Visual identification can be carried out without a sound being made from a picture of such an item. Here is an ontological problem, in that both the picture of the English Horn, and the audio recording can not be said to “be” the instrument itself. This is reminiscent of the “Ceci n’est pas un pipe” artwork of Magritte, which challenges the perception of the image itself. This challenge is further expanded by identifying that while we can say that a record reproduces a sound, we cannot say that a photograph reproduces a sight (or look or appearance). Cavell suggests that language is not expressive enough to cater for this. (pp17-19), and perhaps the relationship between language and ontology is not as simple as it may at first seem.

Lundeen discusses art and ontology in an appraisal of the works of William Blake (Lundeen, 2000). Here, the “difficulty” of ontology is identified in the manner in which Blake poses text in relation to objects within his artwork: “Blake challenges the ontology of word and image by allowing a reciprocal exchange of character and function to take place between them”. (pp25-26).

Blake created works in which text represented or took over a human body [Night thoughts, p73; Jerusalem, plate 62]. In Blake’s works, three modes of text-image combination are noted: the intertwining of text and image, and modes where image not allowed to interact with the centrality of text and vice-versa. This suggests the notion of primary and subsidiary focuses in appraising specific media – depending on the message, different interactions of the mediums provide a certain focus, and graphical presentation may convey certain messages better than others. As an example, Lundeen uses the plate of Blake’s Tyger: the text projects a fearsome creature, while the
accompanying picture of a tiger is somewhat tame looking. In Newton’s Opticks, the illustrations show the image as primary, with lettering added around it for reference purposes (Newton 1952). This would seem to be a general feature of scientific images. Lundeen notes that language is primary even when discussing art: “The dominance of language is, in fact, so insidious, verbal metaphors control the very way in which critics describe non-verbal expression; hence, they speak of the “grammar”, “syntax” and “vocabulary” of art” (Lundeen, 2000, p40).

When discussing ontology, researchers and especially philosophers revert consciously or subconsciously to the origins of the subject itself, proving or disproving the existence of a “God” (for example, Lango 1972 p92-93). While the existence or otherwise of a “God” is the subject of significant theological debate, philosophers have great difficulty in proving physical things, so the proof of abstract entities is certainly beyond our current treatment. Our treatment of ontology is not concerned with proving the existence of God, although such attempts at this proof are certainly of interest to many scholars beyond the theologians.

So, can we bridge the gap between the rarefied and ethereal notion of “existence” with its articulation in language and subsequently onto knowledge representation schema? And if so, how? Lukács provides an explanation of the difference between existence itself and that which can be dealt with in logic, which may assist in this effort (Lukács 1978, p50):

“The development of a pattern, an organism or a social formation, is ontologically a question of real genesis. The laws of its birth (and death) are in the first place a characteristic of the specific being in question. In logic, however, one concept is deduced from another, irrespective of whether this deduction proceeds from the general to the particular or from the particular to the general. As long as logic is employed methodologically as something that does not determine reality, but that is obtained from it by abstraction, there is no need for this difference to give rise to anything that distorts our knowledge of reality.”

Lukács identifies Hegel as the forefather of logic-based ontology: “it was only with Hegel that logic, as newly fashioned by him in a dialectal form, became the bearer of the new ontology” (p27), but warns of the nature of such logic-based systems: “when […] logic is conceived as the theoretical foundation of ontology, it is unavoidable that logical deductions come to be conceived as the proper forms of ontological genesis” (p51). Lukács’ perspective on ontology is interesting, and echoes our interest in the subject:

“The indispensable need for the ontological mode of treatment demonstrates itself [...] as a method that should in no way remain confined to philosophy as such, but should rather
emerge spontaneously in each scientific area; thus the requirement that mathematical formulas in physics, biology, economic, [sic.] etc. should always be interpreted physically, biologically, economically, and so on, if the concrete problems themselves are not to be distorted, has an ontological foundation, but is at the same time an indispensable postulate for true scientific concreteness and exactness" (p92).

This subject-specific treatment of ontology, as opposed to a grand unifying treatment, seems to make sense when considering the scientific language that will be used for describing the subject ontology. Künig makes an important connection between language and ontology: "Much that we find in Aristotle can be interpreted as language analysis, and there is an intimate connection between his ontology and the forms of language" (Künig 1967, p1). Künig further suggests that philosophical problems arise from a misuse of language, and can be resolved by careful and systematic analysis of the language; perhaps by treating each subject field separately? He takes the relationship between language and ontology even further: "So far as natural languages are concerned, philosophers have in the past frequently attempted to relate the structure of language to the structure of reality – grammatical categories to ontological categories" (p7) and discusses the relationship between language and ontology with reference to other philosophers including: Russell’s parallel between the ontological structure of reality and the logical structure of sentences (p14); Frege’s semantics of proper names (p39); Wittgenstein’s relation of sentences to facts and to actual reality (abbilden - ‘to depict’) and relation of sentences to possible states of affairs (darstellen - ‘to represent’) (p54); and the reduction of Wittgenstein’s ontology to mere syntax (p54) Kaminsky supports the relationship between language and ontology: “Generally, the study of ontology is the study of sentences to determine what exists and how, and how to organise or categorise it” (Kaminsky 1969, p19).

The attempted reduction presented here of the study of ontology to the study of language was further evident in Wittgenstein, whose “language game” forms the basis of his philosophy. The position that we express the things that exist using language is useful for our approach to concept systems and specifically to ontology.

2.2.2 Computer Systems and Ontology

The question of ontology, as discussed above, expands beyond the current (computing) flavour of composing relationships between things, to the very question of the existence of the things and how should be represented or described. That we accept the photograph of a scene as some form of representation of that scene, and we may have some knowledge beyond that scene is not necessarily something that can be represented in a “system”. The tacit nature of such knowledge
(which may include notions of emotion about a particular scene) is difficult to represent. Steward argues for a so-called "Ontology of Mind": "Different ways of conceiving of the mode of existence of such entities as beliefs, thoughts, and sensations have given rise to a new ontology of mind, in which the mental events, states, and processes have replaced modifications of the soul" (Steward 1997, p2).

The root of many such systems of representation generally has a taxonomic structure. Taxonomies specifically, and hierarchical organization in general, are widely used for classification and knowledge representation. The notion of inheritance of features, characteristics and values from supertypes has been studied and even formalised by some practitioners, especially Touretzky (1986). Early hierarchically arranged systems, which include Aristotle's Categories, and Porphyry's Isagoge made some attempt at classifying more abstract things in the world with a set of five "predicables" that could be used to describe the things in the world. These predicables were genus, species, difference (or differentia), property and accident. These Five Predicables can be described as follows:

<table>
<thead>
<tr>
<th>Genus</th>
<th>A class of things that can be divided into two or more sub-groups.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Groups resulting from the division of a genus that may each become a genus. ('definition' in Aristotle).</td>
</tr>
<tr>
<td>Difference</td>
<td>An attribute which produces a species by qualifying a genus.</td>
</tr>
<tr>
<td>Property</td>
<td>A quality possessed by each species of a genus but which is not peculiar to the members of that genus.</td>
</tr>
<tr>
<td>Accident</td>
<td>A quality which may or may not be possessed by any species of a genus that is not essential to the definition of a species.</td>
</tr>
</tbody>
</table>

A tree of Porphyry (Figure 5) was used to organise Aristotle's Categories based on genus/species such that human could be found as a rational, living, animate, material.
Hierarchical structures are easy to create, intuitive in terms of their inheritance of properties and make for simple diagrammatic representation of a domain. Hierarchies can be thought to consist of objects or classes, usually a prototypical representative, combined with generic features. Instances of these objects or classes represent, or correspond to, ‘real world’ entities. The notion of inheritance, albeit abstracted from taxonomic work, is used within hierarchies as a powerful inferencing technique. It has been shown by experiments with human subjects that it is easier to recall information that would be encoded at the level of the thing being discussed, than it is to recall more generic features, so the recollection that a canary sings is more readily retrieved than the fact that it does, indeed, breathe (Collins and Quillian 1969). Inheritance is the passing of properties from supertypes (the objects or classes) to the subtypes (other objects or classes, or instances). It has been posited that inheritance is most easily understood for names. “If you hear that your friend has a collie named Rex, you do not need to ask whether Rex is an animal [...] Rex inherits all the properties that define collie and implicitly inherit the properties of dog, canine, carnivore, placental mammal, vertebrate......entity” (Miller et al 1993). Whether Miller et al are focussing on the name of the instance, or the names given to each taxonomic level, is not clear.
Hierarchies need not be fixed systems: Sharp provides a treatment of military aircraft based on “facet analysis” to show how hierarchies can be developed depending on the choice of selections of features (Sharp 1965, pp30-37). According to Wolterstorff, a facet is what we pay attention to in what he calls “abstractive attention” – “we also frequently take note of the colors of colored things, the shapes of shaped things, the loudness of loud things” (Wolterstorff, 1970, p128). This is the attention to “some facet of a multifaceted thing”. A particular facet, he says, is abstracted from the other facets and furthermore, this appears to confirm Aristotle’s Categories: “the phenomenon of abstractive attention is good ground for the conclusion that there are predicables” (ibid). Wolterstorff notes the values of a particular characteristic, and this is echoed by Sharp: “one characteristic only be used at each step of the division and this must be exhausted before another is introduced. The next characteristic is then used to divide the species so derived to form sub-species, and all the resulting classes are then mutually exclusive”. Not following this rule produces what Sharp refers to as collateral classes, “classes of the same order, which are not mutually exclusive, resulting in the possibility of assigning a subject to any of several alternative places in the scheme”. This produces additional problems in sub-classification, and is likely to lead to lattice structures. Sharp refers to this problem as “cross-division”.

Sharp, by way of example of facet analysis, examines (by hand) an aeronautical library to discover terms, which, following Wolterstorff, would denote the shape of shaped things, the position of positioned things and the function of functional things. This analysis produces terms such as Delta, Fighters, Ailerons, Landing, Lift, Rectangular, Drag, Control, Wings, Elliptical, Bombers, Elevators, Lateral, […], Swept-Forward, High-Wing, […]. These terms are marshalled into the following arrangements:

---

2 Facet analysis: The literature which is to be classified is examined and all the significant terms which it uses are listed. The terms are then examined to determine how they might best be grouped according to their similarities, thus providing a number of ‘facets’ consisting of terms denoting similar kinds of concept. Each term in [such] a facet is called a ‘focus’.
In this organisation, it is possible to categorise, for example, “Low-wing, twin-engined military aeroplanes – Ac Cf Eg, Ad Cf, Eg and Ae Cf Eg, or generally Cf Eg”.

For this particular set of data, Sharp argues that “It is impossible to provide for a class of information on ‘the design of single-engined fighter aeroplanes’, because this is not a ‘species’ of single-engined fighter aeroplanes, and ‘design’ is not a legitimate ‘difference’ for the derivation of this spurious species”.

A similar principle has been applied to collections of texts to create so-called Virtual Corpora by forming hierarchies based on values of characteristics of texts (Ahmad et al 1996). Different hierarchies, or virtual corpora, are created by changing the order that attributes, such as language, domain, and text type, are encountered. In the cases of both facet analysis and virtual corpora, the construction of the hierarchy is made through the selection of characteristics and some commitment to the values of these characteristics. The zoologist’s selection of characteristics would differ from the taxonomist, however the “objects” (animals) could be quite common. This reflects on the notion of a view of the world for a specific purpose discussed in Chapter 1.

More recent activities, and here Cyc is a good example, have provided a hierarchical model of a more complex “reality” than that of Aristotle/Porphyry, as the basis for developing their representation of the world (Lenat 1995, Lenat and Guha 1990). We see, in the tree-structure of the “Cyc Upper Ontology” in Figure 6 that there is a degree of overlap between the five predicables of Aristotle and the labels that appear in the Cyc tree: items such as slot and attribute value have a definite relationship with difference, property and accident.
Hierarchically arranged systems are the basis for current treatments of ontology, and this can be traced through work such as that on Cyc where the term "ontology" is used for encoded information (descriptions) of objects and the relationships between objects, which can be used in combination with logic for computational purposes. As the developers of Cyc would probably agree, the only way in which to represent these aspects of reality is to either encode everything or provide powerful reasoning mechanisms that can infer everything. This "everything" is ontology. How this is encoded is dependent on purpose.

As authors have become confident in using the word ontology to describe their computing work we have seen the emergence of ontology as a sub-discipline of computing through the existence of texts since the year 2000. Texts in Artificial Intelligence have variously referred to and defined ontology, and such a definition of ontology that has been repeated in AI literature is that of an "explicit specification of a conceptualisation" (Gruber, 1993), which he qualifies as "A specification of a representational vocabulary for a shared domain of discourse - definitions of classes, relations, functions, and other objects". Gruber takes a utilitarian approach and for him ontology, or more specifically its proper use, will help in sharing (and reuse) of knowledge already available on a computer system. Kampa et al refer to ontology as the "study of 'things that exist' that began as a branch of philosophy and is now popular in the field of knowledge management" and note that ontologies model the real world through specifications of concepts, instances, relations, functions and axioms (Kampa et al 2001).

Ontology for Sowa is the result of conceptual analysis, the production of "a precise, formalizable, catalog of concepts, relations, facts and principals [...]. The ontology, then, is produced for a possible world - a catalogue of everything that makes up the world, how it is put together, and
how it works” (Sowa 1984, p294). This catalogue of concepts cannot be easily formulated for concrete concepts and the “more abstract concepts involve more complex philosophical issues” (p361). This has a slight contrast with his later definition of “the study of the categories of things that exist or may exist in some domain” (Sowa 2000, p492). These definitions suggest a tension between the catalog of things that make up the world, and the things that may exist but are not in the catalog: where, for example, are emotions? The latter definition includes the consideration of non-physical objects. Such a tension is evident when one considers the general top-level separations presented previously (Aristotle/Porphyry), and latterly in the top-level separation of the upper-level ontology of Cyc, which contains three abstract categories: Individual Object, Intangible and Represented Thing. The Cyc project has, for over 20 years, been capturing world knowledge. It could be that the goal of a complete system of human knowledge capable of keeping up with human knowledge is too ambitious: Schank and Abelson identified that “Perhaps there is no single set of rules and relations for constructing all potential knowledge bases at will”. (Schank and Abelson 1977, p3). The subsequent 27 years of research have yet to show otherwise.

Russell and Norvig’s work on Artificial Intelligence (Russell and Norvig 1995) and Sowa’s Knowledge Representation (Sowa 2000) provide descriptions of ontology and its relationship to knowledge representation. Russell and Norvig discuss ontological engineering as the result of deciding “what to talk about” and of deciding on the “vocabulary of predicates, functions and constants” (pp221-222). The resulting “informal list of the concepts in a domain” is the ontology about which logical sentences or axioms can be written. For Norvig and Russell, the result of a task analysis or problem analysis is a special purpose ontology. This ontology will only contain those items of information that are relevant to solving the problem at hand. It is possible, they suggest, to move from specialised ontologies to a more “general ontology” (p226 et seq.). The organisation of a general ontology is carried out under the headings of categories, in which they group notions of other authors of classes, collections, kinds, types and concepts, measures, composite objects, time, space and change, events and processes, physical objects, substances, and mental objects and beliefs. For Sowa, ontology forms a part of knowledge representation, a multidisciplinary subject that involves theories and techniques from the fields of logic (formal structure and rules of inference), ontology (definitions of things that exist in the domain and can be expressed in predicates) and computation (application of ontology and logic) (2000; p.xi). The computational application of logic and ontology is necessary for the construction of computable domain models and these three elements, logic, ontology and computation, forms the core definition of knowledge representation for Sowa. Sowa further describes a framework of distinctions (p75) that can be used to “discriminate and classify the things that exist and define the words that describe them”. Numerous ontologies have been created for different purposes, by which we mean the collections of labels and structures predominantly in flat ASCII-based files,
marked up using so-called ontology representation "languages". Since we have identified the potential multiplicity of views of the world, we can use this to explain why any number of ontologies intended for the same purpose have been developed by different individuals or groups.

For some, computer-oriented ontologies "provide a powerful tool for distributed agent-based information systems" since they represent a shared understanding of information, and promote interoperability (Weal et al 2001). These authors describe an ontology as a "designed artifact" that consists of a domain vocabulary and a set of definitions, and also as "a conceptualisation of a domain into a human understandable and machine readable format, characterised by the entities, attributes, relationships and axioms of the domain". This is a similar description to that provided by Kampa et al, whose ontological model of scholarly activities surrounding a standards committee (the OntoPortal effort) is very much a hand-crafted effort "constructed from knowledge and experience of the research community and recommendations from peers" (Kampa et al 2001). OntoPortal, like many systems, uses a structured representational system, the Simple HTML Ontology Extensions (SHOE). This system is used to annotate web resources (pages) and to infer new knowledge from these resources. Such systems are more generally referred to as ontology "languages", and represent a move towards structuring information in the Web community. This move indicates a shift from the ability to print information, through the simple presentation and sharing of information, to the need for intelligent access to this information. The manifestation of this shift is evident in the move from printing using Standard Generalised Markup Languages (SGML – Goldfarb 1990, ISO 8879) to the presentation of data over the World Wide Web using an application of SGML called the Hypertext Markup Language (HTML). The current stage of this shift can be seen in the emergence of the 'Semantic Web' (Berners-Lee 1999) where the focus is on data identification for information/knowledge content using the meta-markup language eXtensible Markup Language (XML – Bray et al 2000) such that autonomous systems may be able to infer new facts from these data. The creation of XML as a restricted form of its predecessor SGML has been the pivot for the development of a number of notations for the representation of information. Markup formats developed for various user communities exist, including, but not limited to, formats for: news texts (Reuters' NewsML); electronic business applications (ebXML); synchronized multimedia markup (SMIL); and the XML version of HTML, XHTML. These formats are sufficient for delimiting data, providing information, but communities have identified the need for the identification of semantic information. This has led to the development of formats for the purpose of "conceptual" markup, which include:
• **XML-based Ontology Exchange Language (XOL – Karp et al 1999)**

• **Ontology Inference Layer (OIL – Horrocks et al 2002)**

• **DARPA Agent Markup Language (DAML – Berners-Lee et al 2000)**

• **Resource Description Framework (RDF – Lassila and Swick 1999).**

These notations have been developed in various attempts to implement parts of or all of previous knowledge representation schemes such as the Knowledge Query and Manipulation Language (KQML – Chalupsky et al 1992, Finin et al 1994 and Finin et al 1994), the family of languages of KL-One (Brachman and Schmolze, 1985, Brachman et al 1991), and the Knowledge Interchange Format (KIF – Genesereth and Fikes 1992) and its extension to Ontolingua (Gruber 1992). What were “Knowledge Representation” languages, many of which had LISP-like syntaxes, are now being referred to as Ontology languages that are represented using XML. Indeed, it has been argued that the study of Ontology in AI is “becoming a research field called ‘Ontology Engineering’ like ‘Knowledge engineering’ in expert systems”.

Recall the so-called “thesaurus problem” discussed in introduction to Chapter 1. A thesaurus has been described as “one type of ontology, one specialized to information retrieval” (Oard 1997). Oard provides a taxonomy of Cross-Language Text Retrieval Approaches (reproduced, with emphasis added, in Figure 7 below) in which thesaurus-based retrieval is a subtype of ontology-based retrieval, itself a subtype of knowledge-based retrieval. The act of populating a thesaurus appears, therefore, to be generalisable to populating an ontology.

---

3 Furthermore, there is a movement for converging DAML and OIL into DAML-OIL. See, for example, [http://www.daml.org/2000/12/daml+oil-index](http://www.daml.org/2000/12/daml+oil-index), last visited 08 November 2004.

Oard, amongst many others, refers to the language resource (lexical knowledge base) EuroWordNet as an ontology (ibid) echoing others who have referred to Wordnet similarly, or as a “terminological ontology” (Sowa 2000, p497). The loosely defined synonym sets, purportedly arranged hierarchically, are more closely related to a thesaurus than a concept system. The lack of treatment of synonym sets as necessarily representing the same concept is evidence of this. Furthermore, Wordnet has, perhaps rightly, been identified as unsuitable for technical purposes since it is “overly general” and unlikely to contain domain specific terminology (Faure and Nédellec 1998). What, in reference to language resources, can be described as an ontology is therefore the subject of debate (Wilks5). We seek to understand this notion of ontology, and subsequently determine overlaps and potential relationships that may exist in the literature between knowledge representation, ontology and terminology.

2.3 Terminology

2.3.1 Standards

The International Organization for Standardization (ISO – not an abbreviation, but a capitalized version of the Greek word for equal) operates through various committees who interact through an exchange of documents and through a democratic balloting process. The process of creating an international standard takes around 3 years and may proceed through many iterations and

---

5 Wilks, Y. Presentation on "Ontotherapy: or how to stop worrying about what there is". LREC 2002, Las Palmas Gran Canaria.
The general process for this is shown below in Figure 8. The direct route shown does not tell the full story, since a deliverable at stage 1 may include a “Working Draft”, and there may be more than one “Committee Draft” in stage 2. Nor does this include workflow whereby commentary is considered, standards are revised or withdrawn, or voted against and so forth.

ISO has a specific committee which develops standards related to “Terminology and other language resources”, Technical Committee (TC) ISO/TC 37. The focus of the TC is the preparation of standards concerning

“principles and methods; vocabulary of terminology and other language resources; terminology work; preparation and management of language resources; preparation and layout of terminology standards; computerized terminography and lexicography; documentation of terminology and other language resources; coding and codes in the field

---

6 ISO 16642, the development of which the author participated during the course of this thesis, was able to bypass one stage, that of final draft international standard (FDIS), having received a 100% positive vote at draft international standard (DIS) stage.

of language resources; applications of terminology and other language resources in language engineering.

ISO/TC37 has been an important committee of ISO for 50 years and has added to work carried out by Eugen Wüster (1898-1977) who is regarded as one of the founders of terminology as a scientific discipline. Subcommittees of ISO TC37 are concerned with issues relating to three aspects of terminology:

- standards for terminology that describe the information that should be captured relating to a term, and how it should be presented;

- standards of terminology that present terminological information as terms in a number of languages, plus the definition in these languages (which is supposed to adhere to the principle of substitutability) and other usage information;

- standards for the management of terminology that include standards for the convergence of systems of terms, and for interchange between systems of terminology.

There are two particular ISO standards arising from TC 37 that are initially of interest, ISO 704 for “Terminology work – Principles and methods” that outlines structural aspects of terminologies, and ISO 1087 part 1, “Terminology work – Vocabulary – Part 1: Theory and application” that sets out to define the terminology of terminology. ISO 704 describes two types of conceptual relations: hierarchical (generic and partitive) and associative. Figure 9, below, shows two fragments from this standard, the upper describing the scope of the standard, the lower presenting an example of “generic” concept relations.

---

8 According to: http://linux.infoterm.org/iso-e/about-intro.htm, correct at 9 July 2002
Chapter 2

INTERNATIONAL STANDARD

ISO 704:2000(E)

Terminology work — Principles and methods

1 Scope

This International Standard establishes and harmonizes the basic principles and methods for preparing and compiling terminologies both inside and outside the framework of standardization.

This International Standard describes the links between objects, concepts, and their representations through the use of terminologies. It also establishes general principles governing the formation of designations and the formulation of definitions. Full and complete understanding of these principles requires some background knowledge of terminology. The principles are general in nature and this International Standard is applicable to terminology work in scientific, technological, industrial, administrative, and other fields of knowledge.

In the concept diagram below, 'pencil' is a specific concept in relation to the generic concept 'writing instrument'. Similarly, the concepts 'lead pencil' and 'mechanical pencil' are each a specific concept in relation to the generic concept 'pencil'. Each of the coordinate concepts 'lead pencil' and 'mechanical pencil' has a generic relation with the generic concept 'pencil'. The criterion used to increase the specificity of the concept is the nature of the outer casing and graphite core.

Comparing the essential characteristics of a concept and its related concepts (i.e., generic, coordinate and specific) may require an adjustment and refinement of the intension.

Figure 9: Fragments of ISO 704, principles and methods of terminology work, showing the scope of the International Standard and an example concept-oriented terminology

Although the current version of ISO 704 was published by ISO in 2000, it is the result of revision to the original version published in 1987. Perhaps there is a case for arguing that this was an early standard concerned with ontology?

ISO 1087 part 1, referred to as ISO 1087-1, is concerned with the vocabulary of terminology work, and provides a set of definitions for various aspects of terminology.
Table 2.1: Types of Concept

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervenial concept</td>
<td>A concept that is a specific concept relative to another concept.</td>
</tr>
<tr>
<td>Broader concept</td>
<td>A concept that is more general than another concept.</td>
</tr>
<tr>
<td>Concept</td>
<td>A specific concept in a specific subject field.</td>
</tr>
<tr>
<td>Conceptual relation</td>
<td>A relation between concepts.</td>
</tr>
<tr>
<td>Comprehensive concept</td>
<td>A concept that is a comprehensive concept relative to another concept.</td>
</tr>
<tr>
<td>Partial concept</td>
<td>A concept that is a part of a more comprehensive concept.</td>
</tr>
</tbody>
</table>

Although the current version of ISO 1087-1 was published by ISO in 2000, and its title page refers to it as a first edition, it is the result of revision to an original ISO 1087 published in 1990 that was subsequently split into two parts.

In ISO 1087-1, from which the descriptions of data categories in (computer-based) terminology resources were adapted (ISO12620, published in 1999), a term is defined as a "verbal designation of a general concept in a specific subject field" (item numbered 3.4.3 in ISO 1087-1, and A.1 in ISO 12620). The "verbal" element of this description is perhaps misleading in the context of computer-organised terminology collections, having some of seven adjectival senses in the Oxford English Dictionary including those referring to being "Expressed or conveyed by speech instead of writing; stated or delivered by word of mouth; oral". As with many definitions, the definition of what constitutes terms, concepts and so on is not easily made – and use of phrases such as unit of knowledge and specific field of special knowledge in related definitions require further understanding. Despite apparent difficulties in describing terms and terminology, it is agreed that a terminology collection comprises terms related to concepts, definitions, domain information, conceptual relations, and so on. Here we see the relation noted at the end of the previous section to vocabulary, syntax, definition, concepts and relations. We can argue that these sets of related identifiers are coded knowledge fragments. Such fragments may use proprietary or public identifiers, differing data structures and proprietary or public classification.
systems. These coded knowledge fragments represent complex elements of the domain in question; information about the terms and concepts that are used within a particular domain and within the languages of that domain. The terms themselves are the surface form for the deeper form of knowledge – the concepts. Terminologists refer to, and many terminology systems are based on, Ogden and Richards' triangular relationship, seen previously as used by “ontologists”, between the actual object, the mental concept, and the surface form, whether that is a term, or some other sign.

The purpose of a terminology collection is to enable understanding of a specific subject field, and in the case of a multilingual terminology collection, to enable translation between languages in that domain. There is a traditional tension in this notion of translation between terminologists and lexicographers. Some lexicographers such as Krista Varantola argue against notions of true ‘translatability’, and hence are arguing against the possibility of concepts shared across languages, due to a lack of a parallel in certain languages (in Corréard 2002, p36):

"a bilingual dictionary is a contradiction in terms. No such equivalence exists between two languages that would mandate a bilingual word list. However, as we know that there are a number of excellent bilingual dictionaries on the market, I will have to reconsider my statement”.

The notion of equivalence, even when considering supposed synonyms in the same language, is criticised by such authors who argue that these synonyms represent only approximations. Since definitions are only approximations anyway, unless they are complete and exhaustive, such claims are somewhat pointless and irrelevant. Provided there is enough equivalence between synonyms, or across languages, this would seem to suffice generally. The principal differences between terminology and lexicography have been described as follows (Melby et al 1993):

<table>
<thead>
<tr>
<th>A Lexicographical Entry</th>
<th>A Terminological Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treats a word (frequently called a headword) and the multiple senses of the word</td>
<td>Treats a concept and the terms used for that concept</td>
</tr>
<tr>
<td>Treats homonyms in separate entries</td>
<td>Usually treats other senses of the same word in separate entries</td>
</tr>
<tr>
<td>Provides grammatical and etymological information pertaining to the word and includes a full set of word classes</td>
<td>Includes only the grammatical information necessary for correct term usage and is comprised mainly of nouns, verbs and sometimes adjectives</td>
</tr>
<tr>
<td>Is arranged in strict alphabetical order for easy access</td>
<td>Frequently, but not always, presents entries in a systematic order, with alphabetical cross-listing</td>
</tr>
<tr>
<td>Usually treats a universal set of words taken from general language</td>
<td>Treats a systematically defined subset of a domain-specific special language</td>
</tr>
</tbody>
</table>

Figure 11: The principle differences between terminology and lexicography
Of these differences, the final difference between general language and domain-specific special language is perhaps the most interesting. The treatment of the domain-specific special language in terminology reduces the potential for ambiguity caused by word senses as may exist in a general lexicon. Within special languages, it is more likely that people can agree on approximate similarities between terms used for the same concept, which is significant enough to warrant their treatment as synonyms. Such a problem is recognised (ibid):

"Although one may be able to assume that L1-TermA = L2-TermB and L2-TermB = L3-TermC, it is not a foregone conclusion that L3-TermC will always be equal to L1-TermA. For this reason, careful terminologists have learned to document detailed contexts in which language pairs can be considered fully reversible".

The existence of a deprecated term within a concept is such an example of where full substitutability of terms for each other needs to be controlled.

Building a terminology collection is a process of selecting the characteristics required within the system along with their possible values, and using this as a model to populate and manage the system. According to ISO 704, managing a terminology includes the following tasks:

- identifying concepts and concept relations;
- analysing and modelling concept systems on the basis of identified concepts and concept relations;
- establishing representations of concept systems through concept diagrams;
- defining concepts;
- attributing designations (predominantly terms) to each concept in one or more languages;
- recording and presenting terminological data, principally in print and electronic media (terminography).

One question is, how to identify the concepts themselves and, perhaps determine whether they are individual or general concepts? The International Standards for terminology do not contain such information, such as criteria for acceptance of a term, which would enable a novice to make such a decision. According to ISO 860, "Concepts and terms develop differently in individual languages and language communities, depending on social, economic, cultural and linguistic factors. [...] similarity at the term level does not necessarily mean that the concepts behind the terms are identical". This tends to indicate that the identification and acceptance of a term varies from language to language and that the relationships between concepts, and by implication knowledge and language, is not a straightforward one. As such, it would be a non-trivial task to identify the terminology, but once collections of terms are readily available, such standards are applicable. ISO 860 deals with the means for the convergence of concept systems by analysis of
the similarities and differences of the characteristics of their concepts. The comparison of
corcepts is made by the comparison of their definitions, not their terms. This relies on the
definitions themselves being comparable, which for a single language may well be the case,
however for a resource where definitions may exist in other languages, this is not so simple.
Assuming comparability, definitions are collected, compared and where differences are
significant, different concepts are created. It can be argued that this standard forms the basis for
the convergence of terminology collections, and perhaps for ontologies, and the potential, if not
the complete implementation, for this analysis to be carried out. To determine the type of
concept, it is first necessary to collect whatever is available. This collecting process and the task
of establishing some thing as a concept relies on evidence of usage. If this thing can be
pluralized, it would tend to indicate belongingness to the general concept – hence talk is of
planets but not Saturns.

Terms and their associated information such as contextual references, sources and definitions can
be organised into terminological entries of various kinds and may be associated to concepts.
Following Wüster and other positivists, Part one of ISO 1087 (referred to as ISO 1087-1:2000)
states that a concept is 'not necessarily bound to particular languages'; ISO 1087 (parts one and
two) provides a terminology for terminology – a meta-terminology. A concept-oriented
terminology collection consists of concepts which have terms of various types, such as preferred,
deprecated, abbreviations and so on, associated to them. Synonymity is catered for by the choice
of one (preferred) term to serve for all synonyms, with reference from all the non-preferred or
deprecated terms. This is reinforced elsewhere: “a class of keywords may be regarded as a
mapping of several entry words onto a concept, or as a set of related terms” (Sparck-Jones and
Kay 1973, p150). Definitions within ISO1087 include those for a general concept, defined as a
"concept which corresponds to two or more objects which form a group by reason of common
properties". A term is a “verbal designation of an individual concept”. An individual concept
is a “concept which corresponds to only one object”. Examples are given such that planet is a
general concept and Saturn is an individual concept. However, planet is a label that is bound to
particular languages so the distinction between a concept and a term is not an easy one to make.
This seems to parallel the philosophy of W.V.O. Quine, as described in Wolterstorff (1970, pp42-
43):

"Quine defines a "singular term" as one which purports to be true of just one object, and a
"general term" as one which does not purport to be true of just one object (Word and
Object, pp96-98). He holds that the word “purports” in these definitions is essential.
"Pegasus" for example is to be counted as a singular term; for though, in Quine's
terminology, it is not true of anything at all, still it purports to be true of just one object.
The binding of a concept to a language can be a cause of problems: consider the apocryphal story of how the kangaroo got its name from an Australian Aborigine who “did not understand the question” of a European settler asking what the strange jumping marsupial was. Perhaps part of the awkwardness of the distinction between a term and a concept is the definition of an object as “anything perceivable or conceivable”. This does not indicate whether a physical thing, or even a thought or an emotion, should be defined by a term or a concept. Should an emotion such as anger be perceivable, in its manifestation, conceivable, in its ability to occur, or is possible that something as non-concrete as an emotion should not be considered an object at all? Indeed, it may be argued that the abstraction model is perhaps not as straightforward as that in Object Oriented (OO) Programming where an Object is defined by state and behaviour and represents an instance of a Class. In order to build models containing terminology, pragmatic and convenient decisions over design may be more important than accurate and complete description. Indeed, at some point in a concept system there is a notional supertype that subsumes all others – in Description Logics this is referred to as top; dictionary definitions and other systems refer to object or thing. The programming and terminological paradigms appear to contain some similarities, despite terminology differences.

2.3.2 The link between Terminology and Ontology

The link between language, specifically terminology, and ontology has been hinted at above (c.f. Wittgenstein, Lukács, Künig, Kaminsky, Sowa, and so on). Here, we re-emphasise the conceptual orientation of terminology, and note that in discovering the terminology of a domain, the end result represents the concepts (or types, kinds, collections etc) that we expect to populate the ontology; a term can be considered as a lexicalisation of a concept in a certain language. The construction of terminology databases and knowledge bases for expert systems initially involves identifying the concepts of the domain for which the database or knowledge base is being constructed. A terminology database for, for example, high-speed switching circuits, requires the terminology related to the concepts in the domain: currents, voltages, memories and cycle times. However, high-speed switching is made possible by the novel quantum mechanical phenomenon of tunnelling - being able to cross barriers with insufficient energy. The inclusion of concepts related to uncertainty in measurement and the probabilistic positions of other more tangible particles does not sit easily with the rather staid science and engineering of switching devices. Quantum mechanics is one of the open philosophical conundrums if ever one were needed.

An ontology should represent concepts that exist (although the principle of existence is debatable where thought is concerned) within a specific domain, and lexicons contain linguistic descriptors which map to the concepts within a subject field and the terms of the subject field. This
simplification tends to preclude the use of other semiotic systems for the identification of concepts, which is not the intention. Population of a concept-oriented terminology collection therefore appears to parallel many of the requirements of populating an ontology and a lexicon. It is possible to consider that a concept hierarchy within a terminology collection could be mapped to a concept hierarchy in an ontology. In this way, populating an ontology could be derived from the population of a terminology collection, and valid concepts in the ontology map, somehow, to concepts held in the ontology. As terminology collections tend to exist for specialised subject fields, the notion of a domain-restricted ontology could in some ways be derived directly from the terminology. If we can use logical statements about terminological relationships, we can build powerful terminology systems; reasoning about the relationships between words in an ontology requires deriving information about how these words interact.

An ontology that helps to formulate the underlying structure of a specific domain therefore appears to be inextricably linked to terminology; perhaps they are synonymous? A clear, transparent relationship between concepts and terms reflects an ontological clarity and transparency. An ontology has to be asserted in some form and a trace of ontology is in the term of a domain. Sowa discusses a so-called terminological ontology as “An ontology whose categories need not be fully specified by axioms and definitions” and he uses Wordnet as an example of this (Sowa 2000, p497). He further notes that “subsets of the terminology can be used as starting points for formalization” [for axiomatizing concepts], and that this is a valid endeavour since “most fields of science, engineering, business, and law have evolved systems of terminology or nomenclature for naming, classifying, and standardizing their concepts” (our emphasis). That terminology standards provide a starting point for formalizing concepts is part of my thesis.

Sowa has discussed the notion of formal lexicography: the basis of this notion is that a catalog of concept and relation types forms a basis for formally studying the vocabulary of a language. For Maedche, this catalog of concept and relation types forms a component of ontologies: “models that are used to communicate meaning between machines and human beings” (Maedche 2002, p17). Maedche has defined and used an ontology structure together with its lexicon for formally outlining the architecture of an ontology 'engine' and for outlining a more ambitious ontology learning framework. Maedche’s work suggests greater potential for association between terminology and ontology since he describes an ontology structure by exploring, amongst other things, ISO 704, “Principles and methods of terminology” (ibid, p17). Since ISO 704 does not explicitly define an ontology structure, the starting point is perhaps more focussed on working towards the description of how “things” are being described, in this case the description of what a terminology is and how to arrive at a collection of it. This reference to terminology is not
subsequently revived until related work is explored (Maedche 2002, p204). The description of an ontology structure is of interest if it can be related directly to collections of terminology, in which case such large-scale validated collections will be of value to ontology developers as seeds for the domain. Furthermore, the existence of standardized collections of terminology that have been developed along the lines of ISO 704 provide a ready-to-use agreed upon resource for such activities. Important to the development of this conversion is a means by which to undertake a mapping between terminology and ontology. To enable this, we need to define and establish this mapping such that part of the challenge of ontology acquisition can be accomplished through the population of a terminology collection. The result of this effort will be the production of a seed ontology for a given domain which can subsequently be modified by human experts. The result of providing a mapping between terminology and ontology is that of reducing the problem to one that is previously unsolved: the automatic acquisition of terminological knowledge from domain texts. The ability to provide such a well-developed terminology collection that could form the basis for a terminology standard becomes valuable to the subject field itself, whose standards will remain in development for some time.

Maedche's definition of the ontology structure is:

\begin{quote}
An ontology structure is a 5-tuple \( O := (C, R, H^C, \text{rel}, A^O) \), consisting of:

- two disjoint sets \( C \) and \( R \) whose elements are called concepts and relations, respectively.
- a concept hierarchy \( H^C \): \( H^C \) is a directed relation \( H^C \subseteq C \times C \) which is called concept hierarchy or taxonomy. \( H^C(C_1, C_2) \) means that \( C_1 \) is a subconcept of \( C_2 \).
- a function \( \text{rel}: R \rightarrow C \times C \), that relates concepts non-taxonomically. The function \( \text{dom}: R \rightarrow C \) with \( \text{dom}(R) := \Pi_1(\text{rel}(R)) \) gives the domain of \( R \), and \( \text{range}: R \rightarrow C \) with \( \text{range}(R) := \Pi_2(\text{rel}(R)) \) give its range. For \( \text{rel}(R) = (C_1, C_2) \) one may also write \( R(C_1, C_2) \).
- A set of ontology axioms \( A^O \), expressed in an appropriate logical language, e.g. first order logic.
\end{quote}

(p18)

Maedche further defines a lexicon for the ontology structure as a 4-tuple of lexical entries for concepts, \( L^C \), lexical entries for relations, \( L^R \), and what he calls “references” for concepts, \( F \), and relations, \( G \), as \( L = (L^C, L^R, F, G) \). For \( L \in L^C \), \( F(L) = \{ C \in C \mid (L, C) \in F \} \) and analogous definitions are suggested for \( F^{-1}, G, G^{-1} \).
A knowledge base structure, then, is constructed of an ontology, a set of element instances, a concept instantiation function and a relation instantiation function. A lexicon for the knowledge base structure comprises a set of lexical entries for instances and a relation reference for these instances. In his model, "one lexical entry may refer to several concepts or relations and one concept or relation may be referred to by several lexical entries. An ontology structure with lexicon is a pair \((O, L)\) where \(O\) is an ontology structure and \(L\) is a lexicon" (Maedche 2002, p20).

This represents a separation between structure and content for a knowledge base. It also suggests that items may exist in the lexicon for which there is no concept or relation in the ontology, and concepts and relations may exist in the ontology for which there is no item in the lexicon. The implications of this are, as yet, unclear.

Using this set of definitions, Maedche tries to define what he calls an *ontology wrapper* to Wordnet and GermaNet (the German counterpart of Wordnet) (p85). The following table shows the relations that have been considered in creation of an ontology wrapper for Wordnet:

<table>
<thead>
<tr>
<th>WordNet/GermaNet</th>
<th>Ontology O</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synset</td>
<td>(C, L^c)</td>
<td>A synset corresponds to a concept (C). Words contained in the synset are stored in the lexicon (L) and mapped to the specific concept (C).</td>
</tr>
<tr>
<td>Hyperonym, Hyponym</td>
<td>(H^c)</td>
<td>Hypernym relations were evaluated between two synsets and directly mapped to (H^c).</td>
</tr>
<tr>
<td>Meronym, Holonym</td>
<td>(S)</td>
<td>Meronym relations are named &quot;has-part&quot;, holonym relations are named &quot;part-of&quot;.</td>
</tr>
<tr>
<td>Antonym</td>
<td>(S)</td>
<td>Antonym relations are named &quot;opposite-of&quot;.</td>
</tr>
</tbody>
</table>

Table 2: Maedche's definition of the creation of an ontology wrapper for Wordnet and GermaNet

The appearance of \(S\) is not explained in his table, and this is suspected to be a printing error, since the last two rows represent relations, \(\mathcal{R}\). The use of \(\mathcal{R}\) is consistent with other Maedche papers (Maedche and Volz 2001, Maedche and Staab, 2001a). \(S\) appears to be highly similar to \(\mathcal{R}\), except \(S\) "specifies" the same relationship that the function \(rel\) assumes. Furthermore, the ROOT element appears in this earlier paper as the common supertype (universal supertype) such that that \(C \times \{\text{ROOT}\} \subseteq H\)^9.

While Oard's taxonomy produces a means for categorizing resources, and Maedche provides a formal means by which a resource such as Wordnet can be considered ontologically, a "spectrum"

---

^9 The majority of the paper on ontology comparison appears again in a further paper on ontology similarly (Maedche and Staab, 2002).
of types of ontology has also been proposed, reproduced in Figure 11 below (Lassila and McGuinness 2001), which emphasises the many views that can be held about what an ontology is.

![Ontology spectrum (Lassila and McGuinness 2001)](image)

Some elements of this ontology spectrum can be summarised as follows:

- catalogs: a finite list of terms
- glossary: list of terms and natural language meanings
- thesauri: relating terms by synonymy, typically non-hierarchical but hierarchy perhaps deducible from broader/narrower description
- informal is-a: a hierarchically arranged scheme without strict subclassing – the example provided is of Yahoo’s “categories”
- formal is-a: strict control of inheritance
- frames: including property information, with inheritance of properties
- value-restrictions: constraints on properties

Terminology, perhaps, sits on a boundary between informal and formal is-a. Certainly thesaural relationships can be constructed in terminologies, and hierarchical associations made, however the associations are not necessarily strict. We could consider value-restrictions related to use of terms (e.g. deprecation, valid grammatical constructs - contexts) within a terminology. Terminology, it appears, can be considered somewhere along this spectrum depending on how it has been constructed.

Lassila and McGuinness provide three properties which they say should hold in order to consider something an ontology:

1. Finite controlled (extensible) vocabulary
2. Unambiguous interpretation of classes and term relationships

3. Strict hierarchical subclass relationships between classes.

Strict adherence of terminology to the third property is perhaps the most difficult: the first two can certainly be argued for.

Having established a theoretically motivated association between terminology and ontology, what is the purpose of making the association? Having considered language, and how terminology relates to ontology, the question for computing is, realistically, what is the nature of the problem being solved, and what will a solution to the problem provide? Here, we discuss how to build intelligent computer systems, and how the population of terminology or ontology resources that inherently contain or refer to items of language relate to intelligence such that systems can process natural language and, ideally, learn to process language.

2.4 Intelligent Computing

If we take the position that what we wish to develop are computer programs that mimic the behaviours of humans in carrying out specified tasks, and ascribe intelligence to these behaviours, we are attempting to construct intelligent machines. Thinking, learning and adapting are three things we would like these intelligent machines to be able to do; within thinking we include reasoning. Successes at implementing “intelligence” have been fairly limited. Perhaps this is a problem with how we define what intelligent behaviour should be? Expecting a chess “machine” to beat a grandmaster is, perhaps, ambitious—would we suggest that a child learning to play was not intelligent because they could not do this? A further question is: has the machine learnt to play a game, from which it may be able to abstract the means to play other games? Early chess-playing machines reduced the “intelligence” of chess players to a variety of search algorithms, and machines were able to “participate” in the game by outputting specific moves in response to certain input stimuli, but did they “understand” the game?

The question of intelligent behaviour is a fraught one, and perhaps our judgements are clouded in that we wish to retain intelligence as a human facet? We are able to say that both birds and aeroplanes fly, although the characteristics of aerodynamic lift and flapping wings provide different means by which this occurs (Armer in Feigenbaum and McCorduck 1963, p 396). The difference in means of mechanical flight and biological flight does not concern us. Perhaps because we are unable to fly otherwise? To emphasise this point, we go so far as to say that the people in the plane are actually flying. Although a machine, correctly programmed (engineered?) can produce certain results that if a human did the same we would consider the human intelligent,
we criticise the machine's consciousness and ability to adapt its behaviour. A machine that has algorithms for playing chess will not be able to use those algorithms for backgammon since the means by which pieces move, and winning the game, differ in a variety of ways between the two, but if both games are "appropriately" explained to either party, might such gameplay be possible? These examples represent highly specialised tasks, where the input representations and their combination with the algorithms become significant. Minsky suggests that suitable representation is key: "Only to the extent that this "knowledge" is suitably represented can the program's use of it be intelligent, in the sense of appearing to understand the information". (Minsky 1968, p17).

Building intelligent systems requires this suitably represented knowledge, which, in turn, involves collation of concepts. This activity of building intelligent systems has been referred to variously as "knowledge engineering". According to Luger and Stubblefield, "The knowledge engineer is the AI language and representation expert" (Luger and Stubblefield 1993, p314). Knowledge engineers specify, design, build and test knowledge based systems. Knowledge engineering involves a process of acquiring knowledge: extracting and formalizing knowledge of a particular domain for the purpose of problem solving: "the transfer and transformation of problem-solving expertise from some knowledge source to a program" (Buchanan and Shortliffe, 1984, p149). Successful knowledge acquisition work is expected to lead to a conceptual model of the domain created in an AI representation language: such languages can be based on frames, conceptual schemas, object-relations, lambda calculus, description logics, or some hybrid of these.

Knowledge engineering can be undertaken in many and various domains, for example, sewerage design, scenes of crime, financial risk management, medical diagnosis – all requiring different approaches, analysis of various different types of data and production of different conclusions or recommendations. The knowledge engineer attempts to determine the key things within the domain and how they behave in relation to one-another. In each of these cases, the things within the domain will be described using words or, more accurately, terms that will be conveyed from expert to engineer via text or speech. The words themselves will be further differentiated using characteristics associated to these words: here we note again the importance of understanding the terminology. In order to operate in a domain concerned with design of sewerage systems, it is important to understand the possible different combinations of pipes and tanks, their physical characteristics such as length, diameter and the material they are composed of, and how these will assist or inhibit certain types of fluid flows. For a scene of crime domain, it is important to describe the relationships between specific objects, such as the distance of a fingerprint from a given point of reference, or the pattern of blood spattering. In the financial trading domain, the trader wishes to understand whether a given pattern of buying and selling activity within a specific financial market such as the FTSE or the DAX, or for a specific financial instrument such
as Reuters stock, or Dollar/Euro futures, can indicate a specific technical pattern that is bearish, bullish or a turning point, or whether it conforms to a frypan bottom indicating a reversal. In medical diagnosis, the identification of bradychardic or tachycardic heart activity in relation to, for example, a specific type of contusion can assist in the administration of the appropriate drugs and treatment regimen. In each case outlined, in order to produce a computer which could emulate the behaviour of a human expert, it would need to form some representation of pipes, blood spattering, reversal or tachychardia so as to determine how, given such a piece of information, it could associate other pieces of information to draw some conclusion about a given situation, or how to act in relation to it. The knowledge engineer is therefore expected to unravel to some degree the tacitness of tasks performed and honed over many years by skilled exponents such that the performance of these tasks becomes a possibility for a machine. This is not a straightforward and simple process since, as outlined above, it is first necessary to work out what the expert is talking about, which is important in further discerning important things, and the characteristics of those things, in the given domain.

The knowledge engineer therefore needs to understand the terminology of the domain, and other relevant principles including applicable technologies and developments: “The knowledge engineer has to talk to the expert using the expert’s terminology. It would be advantageous, therefore, for the knowledge engineer to have some background knowledge of the domain, for example, the types of problems encountered, the terminology, accepted methods and tools” (Hart 1992, p46). Since the inception of knowledge engineering, the principal obstacle has been the difficulty in acquiring knowledge accurately and quickly – the so-called knowledge acquisition bottleneck (Buchanan and Shortliffe 1984, Luger and Stubblefield 1993). The knowledge engineers to this day manually craft the knowledge of the applications domain; it is extremely rare to find reference to terminology databases or standards - ‘ontology engineering’ is the nearest reference to domain terminology one finds. Key figures in knowledge engineering like Jackson argue that attempts at automatic knowledge acquisition will ultimately lead to effective tools for dealing with the knowledge-acquisition bottleneck. (1990, p464). Automatic knowledge acquisition will depend, in our view, on automatic analysis of the applications domain.

The development of a conceptual model is relevant in terminology work, and so we see parallels in both the need to have some understanding of the terminology of a domain, and in the creation of a conceptual model. The need to update the knowledge base as the world changes, and to appraise the rules present in this knowledge base, is paralleled in the need to maintain current and valid terminology collections. The tasks of knowledge engineering and terminology maintenance both appear to be iterative.
2.5 Chapter Conclusion

In this chapter, we have discussed various aspects of knowledge and language, specifically in consideration of terminology in relation to the current notion(s) of ontology, and the relationship of terminology to the task of knowledge engineering. The purpose of this discussion has been to set the scene for the development of intelligent systems, for which a computational representation of natural language "concepts" is a significant component. To develop such systems, we need to be able to discover patterns of use of language which can be represented and computed, and these patterns should provide evidence of knowledge within given subject fields, and be repeated. In short, we need theories of language use that can be treated mechanically. If we can discover similar patterns of language use across different specialisms, perhaps we can produce a means for automatic knowledge acquisition that will work for an arbitrary domain? Construction of these computer ontologies from collections of text in arbitrary domain requires a means to treat the trace of knowledge that exists in the terms of the domain. The challenge is of identifying the ontology from terminology used in texts to convey knowledge. It should subsequently prove possible to reason over such ontological knowledge and here there are challenges in semantics and pragmatics.

Computing researchers, especially those working in knowledge engineering/representation and lately in the Semantic Web, link three levels of descriptions, syntax, semantics and pragmatics, in a representation scheme for describing the world. Our work is not specifically syntactic, but is informed by, and perhaps can inform, syntax. The Semantic Web community refers to these representation schemes as ontologies, and more specifically as ontology interchange languages. These representations may allow computers to analyse information in ways we had not originally conceived of, and perhaps help us to answer questions such as: "When a machine can use up all the knowledge we have given it, and use it systematically in ways that we cannot ... what will happen?" (Feigenbaum and McCorduck 1984).

Throughout this discussion of ontology, focussed around the "computable" form (a representation), words such as vocabulary, syntax, definition, concepts and relation have arisen. These words are commonly used also within the terminology science, and have been for over 50 years. These so-called ontologies seem to be usurping the ground covered by terminology, but without paying heed to lessons from it. Ontologies and terminologies appear to be interchangeable at first glance. The taxonomic classification of thesauri as ontology suggests that terminology contained within a thesaurus provides an ontology. Aspects of the surface form of knowledge, as evidenced in natural language, including the characteristics of terminology and terminology collections, how we can extract common patterns of language from texts and use
these to populate terminology resources and, subsequently, ontologies, is the focus of the next chapter.
3 Constructing “Knowledge” Resources

Chapter 2 discussed the inextricable links between language and knowledge, and how intelligent computing might be supported by providing representations (ontologies) of “things” that are described using the language of specific subject fields (terminology). Briefly we discussed the associations made by philosophers between knowledge, ontology and language, and how computer specialists currently perceive the use of ontology within intelligent systems. The description of “things” occurs in large part in written language, and it is the texts of scientific subjects that form a principle resource for many aspects of sciences, particularly pedagogic ones. For a scientific subject to exist, it must have a concomitant set of texts, a quantity of which will have been peer-reviewed; scientific journals and books. A specific subject may evolve from an existing subject, in which case its heritage is the area from which it evolved. A single text within a scientific subject can only be considered meaningful if it is either complete in and of itself, or if we are able to refer to standards accepted within the subject, for example: measures such as length defined by reference to a specific item; chemical symbols explained through the periodic table, related defining works and so on. Within such a text, there exist granularities of knowledge fragments, from the full text itself, to other texts it may refer to, to the symbols that it is made up of, including terms, to the ways in which these symbols relate the text to other texts indirectly. As a number of authors have pointed out, the ‘natural’ habitat of terms is texts. Scientific texts will contain scientific terms to which the authors are (ontologically) committed, which will have reference to a system (ontology) of such terms which allows their meaning and relationship to other terms to be determined, whether or not the ontology exists in a formal sense. This ontology of the subject will reflect the current understanding of the scientific subject: the terms represent the surface form, or lexicon, of the subject ontology, while the subject ontology may itself be a surface form of the “one true” ontology (in the philosophical sense). The treatment of scientific texts as a basis for discovering the potential structure of the scientific ontology, through consideration of its (specialist) lexicon, allows us to consider various possible results depending on the knowledge fragments under consideration.

In this chapter, we consider how such terminologies can be populated and, perhaps, ontological inferences made. We review terminology collections, their constituents, standards that they should conform to by which they can be more widely utilised, and the means by which it is possible to get towards such collections using primary resources (text collections). Creation and maintenance of a terminology data base for an arbitrary domain is a time-consuming and human-
resource intensive business. To reduce this resourcing, we consider various techniques for extracting specific words and phrases from collections of text, focussed around statistical, rather than linguistic, mechanisms. A significant problem here is that statistical methods produce limited evidence for valid term formation: as terms increase in length (measured by the number of contained words), the frequency (likelihood) of occurrence tends to decrease. One issue here is the use of statistical methods and techniques that largely work for language fragments (terms, collocations), which are not rare: typically in corpus linguistics and computational lexicography frequency of occurrence is correlated with that of its acceptability amongst (a group of) language users. We believe that ontological commitment is often expressed using multiword terms, which may occur infrequently since they are being generalized or specialized to other concepts in the text. This has been hinted at elsewhere as an economy of speech (to imply written language also) in which there are both more and less economical ways to convey meanings using words (Zipf 1949 p20 et seq). In producing a machine-processable representation that forms the basis of a terminology, it may also be possible to consider how such a representation can be used within an ontology system.

3.1 Motivation

A collection of terminology can be used to support a variety of computational tasks that manipulate (natural) language. Part of the challenge for supporting these activities is the production, or population, of the terminology collections. A subsequent challenge is the continued maintenance of such resources, and whether deriving a completely new resource, or adapting an existing one best provides this. The building of such resources requires the consideration of commonalities of language use: if language is systematically used, and new systems of use appear after major changes in a discipline, then, perhaps, we will be able to populate a terminology collection (partially) through computer programs. What we need to discover are patterns of use of language which can be represented and computed. These patterns may provide evidence of knowledge within given subject fields, and be repeated. If we can discover similar patterns of language use across different specialisms, perhaps we can produce a means for automatic knowledge acquisition that will work for an arbitrary domain. The challenge is of identifying the ontology from terminology used in texts to convey knowledge. It should subsequently prove possible to reason over such ontological knowledge and here there are challenges in semantics and pragmatics.

Our aspiration could be conceived of as the specification and development of a terminology production line which requires tools for the processing of text, for the storage of terms and related information, for the collection of texts from online resources, for the storage of texts in a
searchable collection, and for the management of workflow in the stages of validity of the terms. The needs of the production process differ from the needs of the users of terms, where the focus is on the ability to accurately translate and comprehend the terminology of a document, to ensure consistency in documentation and to search quickly for the most appropriate term. Our focus is on specific aspects of the terminology production line, rather than the development of the user-friendly mechanisms for accessing the terms.

Halliday and Martin recognized a gulf in linguistics between the study of language and the study of text, making a plea for a more 'scientific' treatment of language (1993, pp17-18). We intend to consider how we can scientifically treat text, and what our observations tell us about the use of language.

3.2 Terminology Collections

Terminology collections, stored as data files, databases and prototypical knowledge bases, are an essential component of a range of enterprises and are sponsored by governmental and non-governmental organisations. A terminology collection comprises terms and collections of identifiers and resources that relate to those terms. A computer-based terminology system is designed much like any other data- or knowledge-base by abstracting key characteristics through data modelling, for building database systems, or conceptual modelling, in the case of knowledge based systems. These key characteristics are mapped onto data structures, and increasingly, onto knowledge representation schemata.

Terminology collections exist in a variety of formats - those in medicine are a good example: the US-based Unified Medical Languages System (UMLS) (represented as a large ‘semantic network’) is organised differently to the terminology database of the World Health Organisation, although they may share a certain quantity of information. A number of private enterprises give away or sell terminology collections in finance, commerce and a host of other subject fields. Terminology collections play an important role in enabling multi-lingual and often linguistically divided communities to communicate amongst themselves, in less favoured or minority languages, and with each other. Examples include South African efforts in creating terminology collections in 11 of its languages; the Canadian government’s Anglo-French terminology collections; and the rise in such collections in Eastern and Central Europe. Well known, well-populated multilingual terminology resources include the following collections related to the infrastructure of Europe:
• **Eurodicautom**, the European Commission's multilingual term bank containing around 5 million terms in up to 11 languages

• **TIS**, the terminological database of the European Council containing around 600,000 terms in up to 11 languages

• **Euterpe**, the terminology of the European Parliament containing over 1 million terms

These collections have been developed over many years by many terminologists. In other enterprises, terminology collections may exist on index cards, or in a variety of software applications that have been used over time to organize them. Systems such as Trados' Multiterm and System Quirk's Browser/Refiner provide mechanisms for storing such collections. The need for computer-mediated terminology is apparent in the potential coverage of such resources. Knowledge of specialised terminology in 20 or more different European languages would be an enviable boast. To understand how such resources are created, it is important to study the information encoded in them. A terminology collection may exist for a number of purposes. It can act as an interface between different departments within a company that may be spread geographically. It may underpin other resources such as knowledge management systems or document workflow systems. It may underpin translation activities or provide a useful resource for any individual beginning work within a company who rapidly needs to understand the jargon that their colleagues talk. In the above, terms, definitions and other information about these items are stored in some underlying computer-based format. The consistency of the structures of these formats and the mechanisms they employ to describe these items are many and variable. In some cases, the underlying data model is no longer expressive enough to cater for the information being held, and it becomes common, albeit bad, practice to insert human-interpreted separators into data fields, such as semi-colons, which can provide for the necessary storage, but which would require further machine processing for interpretative purposes. Such human circumvention of data models introduces difficulties into the reusability of this data, however this may be a more appropriate solution than redesigning the data base to cope with a small number of additional needs.

other conventions, states that the definitions of terms in International Standards should follow the principle of substitutability such that the definition could replace the term within the text, without loss of information, subject to minor variations only. These standards have prescribed methods for their creation, for example the three parts of the British Standards Institutions standard for standards (BS 0), and the International Organization for Standardization's ISO/IEC directives; although in some cases they are not followed stringently enough. While a prescriptive approach is possible for a limited application, the combination of terminologies with other resources, and the need for an audit trail to manage at least the validity of terms within a computer system, leads to the need for further identifiers in the description of these terms.

Standards that define how terminology is treated are produced under the auspices of ISO TC 37 (see section 2.3.1). This committee comprises experts drawn from various backgrounds, both academic and industrial, developing terminology standards in a public, transparent, and consensual manner. These standards progress along three main themes; there is a fourth theme concerned with "other language resources". The three main themes are:

**Philosophy:** What terminology is, its principles and methods

**Language:** Identifying language, written forms of language (terminography, lexicography); management of terminology projects including preparation and layout of terminology standards

**Application:** storage and retrieval of terms in computer-based collections.

For us, interaction between these themes is a precursor to, and continues to be a forerunner of current work in ontology.

Those standards specifically relevant to the construction of terminology resources are shown in the table below, along with the thematic aspect (ISO subcommittee) to which the standard has been associated.
Table 3: Standards produced by ISO TC37 relevant to the work described in this thesis. Note that the author has been specifically involved with the production of ISO 16642, revision to ISO 12620 and extension to ISO 639 as a UK nominated expert to ISO TC 37.

If the ISO standards for building a terminology resource are followed, the end result should, in principle be a standardized terminology, with one preferred term for one concept. Other terms for the same concept may be deprecated or otherwise dismissed. This prescriptive view of terminology is quite an idealised one, which may be at odds in various ways with the creative use of language. The result of such an approach becomes a controlled vocabulary. Conforming to this vocabulary most likely results in changes in writing styles and, perhaps, constrains the creation of names for new ideas. ISO conformant terminology requires ISO conformant definitions, which are also constructed artefacts that adequately characterize the “concept” in contrast to other concepts.

The result of a standardization of terminology presents terms and their definitions in a form like that in BS EN 13622:2002, “Gas welding equipment – Terminology – Terms used for gas welding equipment”, a fragment of which is show for English and French in Figure 13 below.
Chapter 3

2.3.3 blowing off the flame
the detachment of the flame from the blowpipe nozzle. This may cause the flame to be extinguished
décollement de flamme
détachement de la flamme de la buse du chalumeau, pouvant conduire à l'extinction de la flamme

2.3.4 blowpipe with multiple flow rates
a blowpipe giving a range of flow rates corresponding to a series of nozzles
chalumeau à débits multiples
chalumeau donnant une gamme de débits correspondant à une série de buses

2.3.5 blowpipe with multiple flow rates adjusted by means of gas control valves
a blowpipe with multiple flow rates which are varied by means of the adjustment valves
chalumeau à débits multiples réglés par des robinets d'admission
chalumeau à débits multiples dans lesquels la variation du débit est obtenue par réglage des robinets d'admission

2.3.6 blowpipe with a variable injector
a blowpipe with multiple flow rates which are varied by means of a device for adjustment of the injector cross-section
chalumeau à débits multiples par réglage de l'injecteur chalumeau à aiguille
chalumeau à débits multiples dans lesquels la variation du débit est obtenue par un dispositif de réglage de la section de l'injecteur

2.3.7 blowpipe with a fixed mixer
a blowpipe with multiple flow rates which are varied by adjusting the feed pressures
chalumeau à mélangeur fixe
chalumeau à débits multiples dans lesquels la variation du débit est obtenue par réglage des pressions d'alimentation

Figure 13: Fragment of BS EN 12622 showing terms in English and French for gas welding equipment. Note the “concept” reference on the left, and that the English definitions do not follow the principle of substitutability.

The purpose of standardized terminology is to avoid ambiguity – despite the inconsistencies we may perceive in the description of key notions used in the standards. For instance, the term subject field is preferred to the term domain although the latter may be accepted. For a published standard, the subject field is identified in the title of the standard; in a terminology database, the subject field of each term is generally associated to that term by some identifier or set of identifiers that denote the subject field and, perhaps, the system of subject fields to which it belongs. The subject-field-specific nature of terminology collections tends to avoid many of the ambiguities caused by different senses of the same word - homonymy. This homonymy is caused, perhaps, by different etymological derivations – for example, bank in the sense of a river derived from Old English or Anglo Saxon and the sense of a financial institution derived from either banca – Latin - or banque - French. In a collection of terminology, instances of the same term in different subject fields are treated as instances of different concepts. This underlines the reason for associating the subject field directly to the term.

ISO 12620 describes categories of term information that include the term itself, term-related information, equivalence, subject field information, conceptual descriptors and relations, notes and documentation, and administrative data. These classifications of term information in ISO 12620 cover more than 170 descriptions of data categories related to terms, which includes such obscure identifiers as grey literature and vulgar register, which appear peculiar within standardised systems of terms. This comprehensive listing has its own data model and a series of explicit associations with other elements of data, for example some data categories act as filler values for others, for example, masculine and feminine being filler values for gender. ISO 12620
Chapter 3

describes a terminology for terminology – a meta-terminology – in addition to ISO 1087. Many of the items described can be considered as the *characteristics* referred to by ISO 1087. There are certain overlaps between descriptions that appear in both these standards, the most notable being the addition of a determiner to the descriptor in ISO 12620. Concepts can be described using attribute-value pairs, indeed characteristics of these concepts can be broken down to triples of an object-attribute-value or subject-predicate-object nature, reducing queries over concepts to the form “X such that the value of Y is Z”.

The properties of terminology, a kind of meta-terminology, are addressed at an abstract level within ISO 12620. There are a variety of ‘simple’ values for specific attributes/characteristics of a concept – indeed some attributes have a closed set of possible values; the example of *gender* once more. However, there are attributes whose values are more complex than may first appear, and here the subject field information and, indeed, the language identifier may be considered; there are properties whose values are potentially infinite (in so far as, for example, the value may be textual and an upper limit on language has yet to be set). Here, particular attributes are chosen where the possible values contain their own semantics: *subject field; language identifier; date; term*.

The standardized approach to terminology is at odds with text-based terminology wherein written texts, and particularly randomly-sampled texts in a domain may be used to establish and elaborate a term. Both these tasks are the domain of the specialists. As we show later, perhaps this conflict can be arranged for the mutual good of the two parties.

3.2.1 The *subject field* attribute

Recall from ISO 1087-1 / ISO 12620 that a term is defined as a “*verbal designation of a general concept in a specific subject field*”. We consider what the “subject field” is. ISO12620 describes a subject field as “A field of special knowledge”. By analysing collections of texts within a subject field, howsoever the subject field is defined, perhaps we can consider the “verbal designation(s)” of the “general concept”?

Aside from more explicit relationships between terms such as synonymy, hyperonymy and so on, the subject field represents a broad, yet complex, relationship between terms. A subject field exists in relation to other subject fields in various classification systems. One subject field may be derived from another, as Nuclear Physics is from Physics. There may be overlaps between subject fields, for example chemistry and biology in biochemistry.
ISO 12620 is not explicit about the “names” of subject fields or the structuring of subject field classification systems, perhaps due to the existence of well-known and well-extended catalogues of subject field names. Here, possibly the most known and widely used classification system, especially in the library sciences community is the Dewey Decimal Classification system\(^{10}\) (DDC), used in more than 135 countries and translated into over 30 languages. DDC was superceded by the Universal Decimal Classification (UDC) system, and is continuously revised. A list of the UDC provides some hierarchically arranged subgrouping (consider section 2.2.2), but the characteristics that form the groupings are not explicit. Various parts of UDC are published in the various parts of British Standard BS 1000. BS1000[8], 1993, contains the English Full Edition of UDC 8, “Language. Linguistics. Literature”. It comprises a 22 page, double-column, listing of codes (which also includes a variety of language codes).

Figure 14: Fragment of page 15 of BS 1000[8], 1993, showing codes for various languages, in which Biblical Hebrew is listed with code 811.411.16'02, and Biblical Aramaic is listed with code 811.411.171.1'02.

The systematic nature of this system of identifiers is not particularly clear. In the above example, both Biblical Aramaic and Biblical Hebrew appear with suffix '02. '01 seems in both cases not to exist. Furthermore, there are gaps between '04 and '08 under Hebrew (8.411.16). With 22 pages for this 1 section of 10, determination of the position of a reference work within UDC may not be straightforward. For example, should a work that deals with the subject of Computer Science appear in classifications under 000 or 500? Where would a text dealing with computer technology be classified? As the granularity of such systems increases, so does the possibility that some compromise is necessary to allow a classification to be made.

\(^{10}\) DDC was conceived by Melvil Dewey in 1873 and first published in 1876.
Some terminology collections use subject field classifications that originated in information and library sciences. For example, Eurodicautom claims to conform to the Lenoch Universal Classification (LUC), which is itself a hierarchically organised collection of identifiers. One version (version information unknown) of the LUC has 3631 Subject Fields within which terms are to be classified. The abbreviations for these 3631 subject fields appear to be systematically organised; subjects are usually given the first two letters as the basic code, EC for economics, PO for politics, RS for risk management; legal aspects generally appear to have the code J based perhaps on the term jurisprudence. Each subdiscipline is either given a letter or numeric value depending on its status within the subject; legal aspects of politics has code POJ, and economics is ECJ; in risk management we have the code RS9 for “disasters of human origin”, which has possibly evolved from “manmade disasters”. However, in certain areas of human activity, one ends up with a rather confusing set of abbreviations for subject fields. The Lenoch Universal Classification has two codes for political crime (RS953 and POJ1) (see Figure 15 below):

![Diagram of Lenoch classification system]

Figure 15: An example of ambiguity in the Lenoch classification system

Assigning a specific subject field to a term/concept associates that term/concept to other terms/concepts in the same subject field. The granularity of the classification system will determine whether one or more subject fields is appropriate for a particular term or concept, and the data model used by that collection will determine whether more than one subject field can be assigned to the term – the two considerations need to be made together. The appropriate position of any item within an existing classification system, where human judgement is concerned, may differ from person to person and, where as above there are ambiguous cases, the arguments for
positioning items within one or the other classifications would cause extensive debate. The compromise would be to employ both classifications, however if both are used in each case, why have two classifications?

ISO 12620:1999 contains three classification-oriented attributes: subject field (A.4), classification system (A.4.1) and classification number (A.4.2). For classification system, UDC is used as an example, and for classification number, UDC 621.3 ("electrical engineering") is referred to. According to the standard, more than one subject field can be indicated for a given concept and, for some reason, hierarchical arrangement can be indicated. By referring to the system and the identifier within the system, as with language identifiers, we would expect this to be a given. The result of commitment to a classification system, and selection of (a) classifier(s) from this system provides a slight ambiguity in presenting subject fields. There appear to be three possibilities:

\[\text{subject field} = \text{electronic engineering}\]

\[\text{subject field} = \text{UDC 621.3}\]

\[\text{classification system} = \text{UDC} \& \text{classification number} = 621.3\]

In the first possibility, no information about the system - "electronic engineering" could occur in any classification system - or relative position of the identifier within the system is retained. The label "electronic engineering" is a language-based (English) label for the code. The second possibility suggests perhaps a single code of "UDC 621.3". Without prior knowledge, this could be a code without a system, or the code "UDC" in the system "621.3"; here, the convention becomes relevant. In the third combination, the separation has been made to system and number, however what "UDC" refers to also has to be understood.

In each of the possibilities presented above, the underlying semantics of the simple labels need further explanation. For humans who understand what UDC is, and how to reference an item in it, this is not an insurmountable problem. For computer representations, we do not have such luxuries. Without associating a language to a code, we have no label that makes sense. Without understanding what both UDC and 621.3 "stand for", we cannot have useful communication. Supposing these values were filled instead by numbers from a system we were not familiar with? The communication would be ineffective. There is more work required than simple selection of a value, and such work is beyond our current considerations. This discussion does, however, relate in an important manner to that of ontology: "what is a subject"? This existential question is itself ontological. Since it is something that we cannot get beyond at present, in our treatment the "subject field" is a human selection.
3.2.2 The language identifier attribute

The above discussion on subject field raises a question of identifying (a) language. The definition of "term" (ISO 1087/12620) uses the word "verbal", by which we should, surely, interpret "in a given language". Where subject fields have their own systems, we also have systems and standards for use in identifying languages. Here, we have a further ontological problem in what a language is. If we refer to something as "English", does this limit it us to the English only of native speakers, or does it also include English as a second language or otherwise? Do we mean either spoken or written forms, or both? Does it cover dialectal variations such as "cockney" or "scouse"? These are questions we cannot easily answer, but we can accept that for certain purposes we can make use of language labels to communicate information between people, where imprecision is acceptable. This may present problems to computer use, where imprecision is difficult to deal with.

A (highly granular) system of identifiers that can enable us to distinguish resources, and here we mean terms, between languages is important. There is an inherent regression here, in that the label "English" is the label for the concept of the language English; the label for the concept of the language English in French is "Anglais". We need to identify the language to identify the language!

ISO 639 is an extant standard for labelling resources with "language". The author has been involved with extending the scope of ISO 639 to provide a means by which language identifiers become defined better in relation to other languages: existing parts of ISO 639 are presented simply as lists (Dalby and Gillam 2004, Dalby, Gillam et al 2004). Such a system for relating languages can then be considered as a kind of classification system.

If we accept that we can describe how to label or tag a language using an item of that language, and we accept the "baggage" that such a tag carries with it, for example, geographical information, then we should be able to use systems of language codes for identification, and indeed standardization of terminology. This identification should enable computers to process terminology, in multiple languages, and to make use of it in related tasks. The language of the term is generally implicit in monolingual terminology – why mention the language if there is only one language? – but although this may be suitable for human interpretation – as long as we don’t ask about regional variants, speech, period of time - computers will be able to store the information since they do not possess a need for identification, but this implicitness will be difficult to retrieve. For bilingual terminology collections the implied languages should be specified – as they are for, for example, French-English dictionaries. For multilingual terminology it is essential to make this identification explicitly. For computational resources,
where terms may not be organised into a physical “half” of a database, as say in a dictionary, the
language of the term needs to “live with” the term.

There are three ISO standards of relevance here: (i) ISO 12620, (ii) ISO 639 (currently being
expanded from 2 to 6 parts, with which the author is involved), and (iii) ISO 3166. In ISO 12620,
the code A10.7 (language symbol) can be used to indicate the name of a language. As such, a set
of appropriate values could be taken from existing parts of ISO639 (ISO639-1 containing codes
for 136 languages and ISO639-2 containing codes for 460 languages)\(^{11}\). Use of these standards
for this purpose is not enforced. Current work is expanding ISO 639 to include the Summer
Institute of Linguistics’ (SIL) Ethnologue (Grimes 2000a, 2000b), and Linguasphere
Observatory’s Registry (Dalby 2000) codes. Some of these standards are directly relevant to the
development of the Internet itself: the Internet Engineering Task Force (IETF) document “Tags
for the Identification of Languages” (referred to as RFC 3066) describes how to interpret
language tags (a basis for the XML attribute known as xml:lang). The author has been involved
in discussions regarding the evolution of ISO 639, and the potential evolution of RFC 3066.

The difficulty in describing a language, noted above, can be seen by investigating these resources.
ISO639 currently lists English with the codes en (part 1) and eng (part 2), alongside its name in
both English and French, shown in Table 4 below (as well as alternative codes for bibliographic
and terminological uses of part 2 codes).

<table>
<thead>
<tr>
<th>Language Name (English)</th>
<th>Language Name (French)</th>
<th>639-2</th>
<th>639-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Anglais</td>
<td>eng</td>
<td>en</td>
</tr>
<tr>
<td>German</td>
<td>Allemand</td>
<td>ger/deu</td>
<td>de</td>
</tr>
<tr>
<td>French</td>
<td>Francais</td>
<td>fre/fra</td>
<td>fr</td>
</tr>
</tbody>
</table>

Table 4: Examples of Language codes taken from ISO639

To cater for regional use of languages, ISO 3166 for country codes needs to be used in
combination with ISO 639, whereby a possible result is a code such as en-GB. This combination
is thought by many to represent British English, however taking the listed combination, we clearly
have United Kingdom English\(^{12}\) (see Table 5 below containing English-dependent, Numeric and
Alpha-2/3 examples of codes).

\(^{11}\) They could alternatively be taken from a system such as the Linguasphere Register (covering
13,000 so-called inner languages plus 8,800 constituent dialects) or the Summer Institute of
Linguistics’ Ethnologue (covering more than 6800 languages). Each of these systems has been
developed from a different point of view and therefore has different coverage.

\(^{12}\) The difference between “Great Britain”, and the “United Kingdom of Great Britain and Northern
Ireland” has clearly been lost in the human interpretation of such codes, and one wonders how
computer representations can fix such failings.
SIL’s Ethnologue lists English as a language spoken in a variety of locations including Eritrea and Ethiopia. The Linguasphere Register has regional aspects as an important distinction. Linguasphere and Ethnologue both have information regarding dialectal variations such as *Geordie* or *Scouse*. Although this may only be relevant in terminology collections for highly localised products, where speech resources are being collected, they may be valuable. Typically, dialects of a language are of little interest to a terminologist except for deprecated terms.

The different “views” about language, and what it is, will determine which code set or combination of code sets is more appropriate. What appears on the surface to be a convenient choice of label hides a great deal of ontological commitment to the identification of one language compared to another. Currently there is no consensus on how to harmonize these systems. If we cannot define what a language *is*, how can we analyse it effectively? Again, we would be going beyond the scope of the current thesis.

The result of a commitment to a specific system, selection of a value from this system, and combination with an attribute of terminology, results in the provision of an attribute-value combination such as for ISO 639-1:

\[
\text{language} = \text{en}
\]

or as according to the Linguasphere system:

\[
\text{language} = \text{engl}
\]

However, the selected system is implicit in this reference. Suppose we consider a case-insensitive label of “fr”. Does this come from ISO 639-1 and imply that we are talking about the French language, or does this come from ISO 3166 and imply that we are talking about the languages of France? Perhaps as with the subject field classification we need to introduce the “classification system” again to ensure that the semantics of the attribute-value combination is clear.

The mapping between systems of language identifiers are not well understood, and yet identification of language is fundamental to, for example, archival of language materials. Converting between systems of language identifiers means, ideally, that the value represents a pointer to both the system AND the code. The system of language codes should handle the
mappings to other systems. Exchanging such information is currently based on assumptions regarding language codes. This is not an optimum solution, and is the subject of extensive and ongoing discussion within the language codes community oriented around ISO 639 (through ISO TC 37).

Systems of language codes present a means, additional to subject field classifications, by which information can be classified, related, stored and retrieved. Differently constructed systems of language codes provide a perspective on different views held by individuals or groups have of language at large. The question of what a language is, and how it relates to other languages remains a challenge that requires further philosophical consideration, which will be of import to computational representation. Again, this is beyond the scope of the current thesis, but again it relates to ontology: “what is a language”? This question is itself ontological, although it is something that we cannot get beyond at present, so for us the “language identifier” is a human selection.

3.2.3 The date attribute(s)

Our discussions so far have led us through hierarchically arranged systems of languages and subject fields. These systems provide some means to interpret terminology: it is in a given subject and in a given language. The ISO definition of a term, a “verbal designation of a general concept in a specific subject field”, leads us to ask questions about the “general concept”. Term-'external' data, by which we mean data rooted in the origination and use of the term cannot easily be deduced from linguistic constructs. The date on which a term stands in place of a given meaning, or undergoes a particular semantic shift, is data that is currently not recorded in terminology resources. Once again, this information is ontological: what is the thing that the term is used in place of?

Here, we consider the issue of “semantic shift”: while the term remains the same, the concept beneath it changes. Such semantic shift has been dealt with in relation to the atom (Ahmad 1998, 1996). The atom has shifted from being indivisible to spontaneously-breaking; from being composed of a nucleus with protons and neutrons, and orbiting electrons, to these elements having other elementary particles, and in modern science to looking for sub-atomic particles that exist for billionths of a second after reactions requiring massive quantities of energy to produce. These views of matter moved through various centuries, with scientists including Max Planck, Ernest Rutherford, Niels Bohr, Albert Einstein and Enrico Fermi developing the 20th century view of protons, neutrons, electrons, neutrinos, mesons and photons in the first half of this century, and quarks, gluons, partons more recently. The ideas have been refined through Rutherford's
experiments, Planck’s Quantum Theory, Einstein’s Relativity Theory, Bohr’s nuclear model of the atom and Fermi’s refinement of the model.

Each change in understanding of the atom did not result in changing the word “atom”, but in the underlying concept — scientists changed what they meant when using the word, but not the word itself. This seems almost counter-intuitive: how can it be easier to change an entire definition that to use a new label? We have to know what the “atom” is that we are referring to. Are scientists in 1900 referring to the same atom as those in 2000? Has just the understanding of an atom changed, or has the referent changed also? This is analogous to asking whether “Rome” represents the same concept and geography throughout history.

Date information in ISO 12620 is restricted to administrative function, recording when a term was entered into a terminology resource, rejected from use, validated by an expert, used last, and so on (see Table 6).

<table>
<thead>
<tr>
<th>Data Category Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
</tr>
<tr>
<td>origination date</td>
</tr>
<tr>
<td>input date</td>
</tr>
<tr>
<td>modification date</td>
</tr>
<tr>
<td>check date</td>
</tr>
<tr>
<td>approval date</td>
</tr>
<tr>
<td>withdrawal date</td>
</tr>
<tr>
<td>standardization date</td>
</tr>
<tr>
<td>exportation date</td>
</tr>
<tr>
<td>importation date</td>
</tr>
<tr>
<td>last usage date</td>
</tr>
</tbody>
</table>

Table 6: Data Categories in ISO12620 for recording date information. Note the primarily administrative function of these categories.

The author has proposed (to ISO TC37 SC3) that most of these date identifiers are redundant since a) they should have the same form (that of a date) and b) since the various collocations relate to terminology management events, which exist autonomously: it is not necessary to maintain a date typology when an event typology can be associated to a date. At time of writing, the removal of these redundant items is being debated.

Although date information here relates to management of terminologies, rather than existence of terms and concepts as we would like, it is important in terms of reusability and consistency to ensure that where a date is employed, the format of this date is clear. Here we have a single ISO standard for dates and times: ISO 8601. The representation of date and time still needs to be decided since ISO 8601 assumes the use of Coordinated Universal Time (UTC), and divergence from this can be represented using the number of hours or hours and minutes (e.g. 05 or 0500).
Such accuracy is, perhaps unnecessary for the purposes of terminology, and an ISO 8601 conformant date composed of 4 characters for year, 2 for month and 2 for day, in the string form YYYY-MM-DD could provide a consistent representation, although 8601 allows a number of alternative possible forms (variations) for this. American English data may use the non-8601 form YYYY-DD-MM, which may pose difficulties in mapping undocumented data. ISO 8601 does not support the use of language-dependent date forms (for example use of Juli / Juillet / July for German, French and English respectively) for which look-up tables and, again, language identification would be needed. Internal consistency of a terminology collection therefore requires some element of date processing.

Supposing, at some future point, we decide to measure semantic shift, and that all dates can be represented using a single common form, the result of commitment to a date representation system, the selection of a representation from this system, and its filling with the actual data provides, simply, an attribute-value combination such as:

\[ \text{date} = 2004-04-12 \]

Fortunately, unlike language codes and subject field classifications, there are not competing sets of identifiers that require further treatment, only variations according to meridian and selection.

The question of date has identified the ontological question of existence, including existence in relation to other things. We have not, and will not, address the underlying ontological problem of time itself.

### 3.2.4 The term attribute

Recall the standards-based definition that a term is a “verbal designation of a general concept in a specific subject field” (ISO 12620). We have considered the “specific subject field” with reference to systems of subject field classifiers and their identifiers; we considered “verbal designation” in relation to the identification of the language in which the term exists; and we considered semantic shift as an aspect of “a general concept”. A term, then, exists in relation to a system of languages, a system of subjects, and some position on a time axis when it was defined in a given way. Presenting information with regard to each of these items requires some agreement on the systems of identifiers being used: indeed it requires an ontological commitment to the structure and representation of such systems.

Consider an example: the University of Surrey library contains a text called “Semiconductor spintronics and quantum computation”. This is classified under UDC 621.381. BS 1000 for
section 621.3 (1989) contains no .381 sub-element. In working through the outlines of UDC, we find the following trace:

- 6 - Applied sciences. Medicine. Technology
- 62 - Engineering. Technology in general
- 621.3 - Electrical Engineering.

From BS 1000, we discover:


But without direct access to current versions of UDC, is it not possible to find out what 621.38 "means". It is believed to be "Solid State Electronics".

From this information, assuming we take "semiconductor spintronics" as a term, the publication date (2000) as the date of this term, and assume the text is in American English, since it was published in New York, we would have an ontological commitment for this term as shown in Figure 16 below.

13 In DDC online, 621 is referred to as Applied Physics (see: http://www.oclc.org/dewev/resources/summaries/deweysummaries.pdf)
Figure 16: Ontological commitment to systems of classifiers and other descriptors needed to describe a standards-conformant term. Dotted arrows show interdependence in identification.

These various interdependencies, along with the structural and notational conventions, show to some extent the underlying complexity in the various systems, and that use of each of them is not necessarily a simple process. To identify a term minimally (according to ISO, but with our additional date constraint) requires all of these systems to be considered together, and we instantiate values in given languages for human readability – a problem of self-reference for the language codes. This is a non-trivial task, and perhaps the cost of terminology production is the result of this complexity.

Within this, the ontological question “What is a term?” still remains. A term has been described as a means to condense information “on the content ‘plane’” (Halliday and Martin 1993, p29), but here it seems to be condensing information on other planes in relation to subject, language, and time. The “verbal designation” is hiding a large amount of semantic (concept) information, and in text it stands in place of this large amount of information and needs to be “unpacked” in order to be understood (ibid, p31), but here the unpacking becomes more complex since to understand this term, we have to unpack language and subject fields also. The semantic unpacking relates to taxonomic relationships as well as other relations, and requires further terms in its understanding. Differentiations in language and subjects contribute here also. It would be difficult to make such
an unpacking in relation to other terms without having some means of presenting those other
terms in a related way and, perhaps, of unpacking them also.

It can be argued that a term is a place-holder for its definition, which in turn is a place-holder for a
concept: in standardized terminology the definition should follow the principle of substitutability
for exactly this reason. If “semantic unpacking” were to be formalized along the various axes
discussed, including that of related terms, subject fields and languages, this would result in further
progression of a terminology along the ontology spectrum discussed in section 2.3.2 (Lassila and
McGuinness 2001): whether this helps us to consider a move to a frame-based system, perhaps
with some value restrictions, is not yet obvious.

We discussed formulating subject field classifiers in section 3.2.1, and since the term seems to be
predicated on the specification of a subject field, designating the subject field, and the
classification system from which it comes. We have discussed language identifiers and dates in
related ways. What we have not yet addressed is what constitutes a term, and what does not. In
English, whitespace is designated as a word separator. A term can be constituted by one or more
whitespace-delimited words, with additional constraints on the words allowed for this
construction. The “value” for the term attribute therefore seems to be open, subject field specific,
language dependent (not in the definition of term according to the standard) and managed
according to the dates upon which specific events occurred, and perhaps eventually with reference
to etymology if it has been adopted from elsewhere. A term should have a definition (at the
conceptual level) that could be “unpacked”, and perhaps a number of example contexts or even a
defining context, and may be related to other terms and/or concepts in various ways. The
selection of values for these few characteristics shows some of the difficulties in developing
computer-processable systems of terms. In the next section, the extraction of terms from
collections of text will be discussed. The purpose of this extraction is to provide values that fill
the term attribute. The number of mechanisms employed to do so merits its own section.
Subsequently, we will discuss the impact that selection of various characteristics and values has
upon the possibility of exchanging data with others, specifically in reference to the burgeoning
ontology formats.

3.3 Terminology Extraction

In Section 3.2, terminology collections were considered with reference to the constituent
identifiers used for them, with reference to extant international standards. In this section we will
discuss how to produce candidate terms which, in turn, might populate the term attribute: they
may also contribute in some ways to providing some of the conceptual attributes
(subtype/supertype) that also exist in terminologies, however it is necessary to evaluate this provision.

We have discussed aspects of how terminology has been described, and various characteristics associated with terminology, and yet a challenge of terminology hinted at above is the filling of such descriptions. Traditionally, committees of terminologists and experts have been required to produce and validate terms. If experts consider a fragment of text to represent some concept, this information can be stored, perhaps using an interchange format as an intermediate such that it could subsequently be stored within any terminology database. What we intend to argue is that experts may need some assistance in finding potential use, and nuances, of terminology. Given the deluge of text and some automation, perhaps one can be more ambitious and talk about automating the whole process?

To automate a process of populating terminologies, texts can have information extracted from them to generate some form of knowledge fragment (computer understanding), which could subsequently be presented as an argument for the existence of a term by terminologists to experts. For Schank and Riesbeck, understanding is based on a “theory of causal connectedness of text” which evokes a “script” for filling in the “gaps in the causal inference chain between events in a story” (Schank and Riesbeck 1981). The act of filling in the “causal chain” shows the connectedness of the text (pp22-35). If a causal chain can be created, the text is “well-connected”. Their understanding is coded by eleven primitive actions using Conceptual Dependency (CD). Such understanding is exemplified best by the Script Applier Mechanism (SAM) (pp75-119) and the Pattern Applier Mechanism (PAM) (pp136-179) programs. The connectedness of a text is also a facet of work in lexical cohesion (Hoey 1991). In this approach, repetition of forms of words within text provides a measure of how cohesive that text is. In contrast with Schank and Abelson, this approach implies that all the information required is contained within the text being treated, or can be determined through a variety of syntactic and lexical relationships such as ellipsis, reference, substitution, repetition and so on (for further details, see Halliday and Hasan 1976, Benbrahim and Ahmad 1995). The use of thesaural relations from a general language resource such as Roget, Macquarie or the ubiquitous Wordnet enables words within the text to be related to other words within the text. To enable better computer understanding of text, and with reference to Schank and Riesbeck, and Hoey, we can consider the cohesion between texts. These texts may assist in filling in the “causal chain”, providing a better understanding of each text, as well as of the subject field in general.

We begin with a premise that the texts contain terms and exemplify a number of relationships between these terms, containing such devices as hyponymy/hyperonymy, meronymy, lists, causal links, analogies, exemplars and so on. For our present treatment, we ignore possible variations in
text due to genre, register, and so on. By analysing lexical structures within texts, and with reference to the international standards for terminology, our method may be more likely to facilitate the construction of terminological resources that, in conformity with standards, have a potentially wider scope of use than simply as collections of terms.

An automatic identification procedure for terms, and the analysis of relationships, appears possible, and here we follow Sparck-Jones and Kay who have noted that "there must be ways of reducing to some relatively well-disciplined form the great variety of ways in which a given thing can be said in everyday language" (Sparck-Jones and Kay 1973, p43). An analysis of texts in the so-called controlled languages supports the above claim. Controlled languages, with limited syntax, grammar and semantics, simplify the task of translation between languages and have been highly successful for Boeing and the Canadian Weather Forecast Offices. Church and Rau refer to these as "sublanguages" and point to critics such as Schank who claim that these often turn out to be far richer and more complicated than anticipated (Church and Rau, 1995). Perhaps they are implying that the terminology of these sublanguages makes them richer and more complicated? Church and Rau, and even Schank, perhaps did not consider that as the sublanguage contributes to the general language, some of the solutions to difficulties in sublanguages might become applicable more generally.

If extraction of terms from text can be assisted by a number of algorithms and computer programs, it will be possible to populate term bases, and then perhaps to infer the conceptual structure of a domain. The analysis of text has been variously broken down in the natural language processing (NLP) community into a number of key processes that include segmentation, tokenization, counting, tagging and parsing (e.g. as shown in the architecture model for the Information Extraction system “GATE”, Cunningham 2000, p93). For our purposes, tokenization has importance, although predominantly we are concerned with 'whitespace' separation since English is our source/target language. Counting is particularly important, and here we consider word frequency counts: we are interested in the notion of frequency correlating with acceptability (Quirk et al 1985, p33). Although we are carrying out some form of tagging, it is not part-of-speech tagging, as the NLP community intend.

Frequency and concordance analysis techniques, originally used to partly validate religious texts, are well established. Frequency analysis relates to counting words and letters within a given text and comparing them with a reference distribution of words and texts. These techniques were intended to authenticate that copies were the faithful reproductions of original works. The objectivity of authentication was assured using a mechanically reproducible method that worked independently of the counter - a mechanical check on change. This method has been applied to tasks such as determining authorship, analysing court transcripts, text categorisation, and more.
diversely, analysis of microbial genomes (Bruccoleri et al 1998). In corpus linguistics, it is suggested that text corpora can be used to validate theories about language at a fixed point in time (synchronously) and over time (diachronically). Methods and techniques in corpus linguistics complement those in information extraction and information retrieval. Such complementarity may allow us to advance the claim that perhaps one can understand documents synchronically and diachronically, and hence entire subjects, by harvesting terminology from collections of text. These texts can themselves be, increasingly, harvested from the Web. Existing collections of terms could be used as seeds for searching and retrieving texts from which further terms can be extracted; terminologies could be bootstrapped from existing terminologies. Once collected, however this is done, experts should validate the terms and their relationships with other terms. The validated (new) terms are then seeds for further information collection, and the process continues.

In this section, we investigate a selection of text analysis methods that may be useful for providing a system of methods (methodology) that can be used to extract terms from arbitrary collections of specialist text. We seek a way to treat key differences between general use of a language, for everyday purposes of interaction, pleasure, and so on, and the language used in specialist texts. A specialist text is written for a specific audience and there is an assumption of common purpose and understanding – physicists generally write for their physicist peers, for instance. But physicists use their general language to write the specialist documents. To paraphrase Sparck-Jones and Kay, specialists, such as physicists, appear to be relatively well disciplined.

For our comparisons, we consider various sets of specialist text corpora of various sizes, stored at Surrey. The purpose of this section is to determine whether a consistent systematic approach to specialist corpora is possible. If different specialisms use language in similar ways, it may be possible to systematically extract terminology from arbitrary collections of text. We consider some of the existing analytical techniques and study the results of their application to various English special language texts; the geographical origin of these texts is not significant, and orthographic variations are not treated. Furthermore, we consider how these techniques can be adapted and combined to automate the extraction of terms from arbitrary collections of specialist text.

In our treatment, we attempt to retain the distinction between words as tokens (instances occurring at various locations in texts) and words as types (the different words used), following Charles Sanders Peirce (1839-1914). We hope that ambiguities through different senses of the same "string" are not prevalent in specialist corpora.
3.3.1 Extracting individual words

Tokenization and counting of words in collections of text provides a means by which to study possible common properties of language use. The first 100 most frequent words in a corpus have been shown to comprise just under half of that corpus. This is true of most corpora including specialist and non-specialist corpora (Ahmad 1999a, Ahmad 1995). A quarter of all tokens of a reasonably-sized special language corpus (c. 0.5 million tokens) are represented by closed class words – determiners, prepositions, conjunctions – with determiners almost invariably comprising 5% or more. Specialist corpora typically have over 25 open class words (types) in this top 100, usually nouns describing objects and events in the domain. These words may indicate technical terms: “an essential part of scientific language” (Halliday and Martin 1993, p4). Work by Ahmad at the University of Surrey has shown that the first 100 most frequent types within a text collection of a given size (around half a million words) has between 30 and 40 open class words, as shown in Table 7.

<table>
<thead>
<tr>
<th></th>
<th>Linguistics</th>
<th>Nuclear Physics</th>
<th>Automotive Engineering</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Class</td>
<td>30</td>
<td>33</td>
<td>39</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34%</td>
</tr>
<tr>
<td>Closed Class</td>
<td>70</td>
<td>67</td>
<td>61</td>
<td>208</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>66%</td>
</tr>
</tbody>
</table>

Table 7: Frequency of open-class and closed class words in three Surrey Corpora

A Chi-square test for statistical significance produces a value of 1.87, showing that the distribution is not significant. Perhaps we can generalize from our samples that specialist text collections show similar characteristics? Inclusion of counts from BNC (2 open class in top 100) produces a chi-square value of 42, suggesting that the distribution then becomes significant. This treatment of open-class and closed-class words currently requires human judgement and prior domain knowledge. The ranks at which open-class words occur does not show a consistent pattern, however the percentage of the corpus represented by only 100 words, previously referred to, is of interest since it indicates consistency in language use across corpora.

Since we seek to systematize our approach, we consider a number of metrics that may be useful for determining how language behaves by providing evidence of specific characteristics of language use. The first of these is known as Zipf’s Law.

Zipf’s Law

The distribution of tokens in a text is an important property initially used in information theory. George Kingsley Zipf specified a power-law function using the notion that the rank of a token, i.e. its position in a list ordered according to its frequency, is inversely proportional to its frequency.
(Zipf 1949). This law will play an important role for us for two reasons: first, it may throw light on the possible differences between a specialist text (corpora) and a general language text (corpora) which will complement the empirical, but statistically founded, method shown Table 7; second, it may help us to deal with opportunities for using words (types) found infrequently. It has been shown that rank multiplied by relative frequency produces a constant around 0.1 (Li 1992, Yavuz 1974). Another way to consider this is that rank multiplied by frequency produces a constant that is a tenth the number of tokens in the corpus (e.g. for the Brown Corpus, Manning and Schutze 1999, pp26-27).

We begin by examination of the BNC. This corpus comprises 100,106,029 word-tokens with 669417 different word-types. The distribution of these tokens, frequency vs rank distribution that is, shows a range of extremes:

1. 10 most frequent types – the, of, and, to, a, in, that, it, is, was – comprise over 22% of all the tokens in the corpus – over 22 million tokens

2. There are 353689 word-types that are only used once each (frequency of 1), a further 86498 used only twice (two tokens each), 40312 used only 3 times, and so on.

3. Zipf’s power-law suggests that the product of frequency and rank should be a constant and that the product is about a 10th of the size of the corpus; for BNC therefore, Zipf’s constant is around 10 million. Table 8 and Figure 17 show that the law holds only for a part of the corpus.
<table>
<thead>
<tr>
<th></th>
<th>f</th>
<th>f * r</th>
<th>Error</th>
<th>Percentage Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6187927</td>
<td>6187927</td>
<td>-3822676</td>
<td>-38.19%</td>
</tr>
<tr>
<td>2</td>
<td>2941790</td>
<td>5883580</td>
<td>-417023</td>
<td>-41.23%</td>
</tr>
<tr>
<td>3</td>
<td>2682878</td>
<td>8048634</td>
<td>-196199</td>
<td>-19.60%</td>
</tr>
<tr>
<td>4</td>
<td>2560346</td>
<td>10241384</td>
<td>230781</td>
<td>+2.31%</td>
</tr>
<tr>
<td>5</td>
<td>2150885</td>
<td>10754425</td>
<td>743822</td>
<td>+7.43%</td>
</tr>
<tr>
<td>6</td>
<td>1883295</td>
<td>11299770</td>
<td>1289167</td>
<td>12.88%</td>
</tr>
<tr>
<td>7</td>
<td>1115382</td>
<td>7807674</td>
<td>-2202929</td>
<td>-22.01%</td>
</tr>
<tr>
<td>8</td>
<td>1089559</td>
<td>8716472</td>
<td>-1294131</td>
<td>-12.93%</td>
</tr>
<tr>
<td>9</td>
<td>998867</td>
<td>8989803</td>
<td>-1020800</td>
<td>-10.20%</td>
</tr>
<tr>
<td>10</td>
<td>923975</td>
<td>9239750</td>
<td>-770853</td>
<td>-7.70%</td>
</tr>
<tr>
<td>3001</td>
<td>3417</td>
<td>10254417</td>
<td>243814</td>
<td>+2.44%</td>
</tr>
<tr>
<td>3002</td>
<td>3414</td>
<td>10248828</td>
<td>238225</td>
<td>+2.38%</td>
</tr>
<tr>
<td>3003</td>
<td>3411</td>
<td>10243233</td>
<td>232630</td>
<td>+2.32%</td>
</tr>
<tr>
<td>3004</td>
<td>3409</td>
<td>10240636</td>
<td>230033</td>
<td>+2.30%</td>
</tr>
<tr>
<td>3005</td>
<td>3409</td>
<td>10244045</td>
<td>233442</td>
<td>+2.33%</td>
</tr>
<tr>
<td>3006</td>
<td>3408</td>
<td>10244448</td>
<td>233845</td>
<td>+2.34%</td>
</tr>
<tr>
<td>3007</td>
<td>3408</td>
<td>10247856</td>
<td>237253</td>
<td>+2.37%</td>
</tr>
<tr>
<td>3008</td>
<td>3405</td>
<td>10242240</td>
<td>231637</td>
<td>+2.31%</td>
</tr>
<tr>
<td>3009</td>
<td>3404</td>
<td>10242636</td>
<td>232033</td>
<td>+2.32%</td>
</tr>
<tr>
<td>3010</td>
<td>3404</td>
<td>10246040</td>
<td>235437</td>
<td>+2.35%</td>
</tr>
<tr>
<td>4991</td>
<td>1746</td>
<td>8712540</td>
<td>-1298063</td>
<td>-12.95%</td>
</tr>
<tr>
<td>4992</td>
<td>1746</td>
<td>8714286</td>
<td>-1296317</td>
<td>-12.98%</td>
</tr>
<tr>
<td>4993</td>
<td>1745</td>
<td>8711040</td>
<td>-1299563</td>
<td>-12.96%</td>
</tr>
<tr>
<td>4994</td>
<td>1745</td>
<td>8712785</td>
<td>-1297818</td>
<td>-12.95%</td>
</tr>
<tr>
<td>4995</td>
<td>1745</td>
<td>8714530</td>
<td>-1296073</td>
<td>-12.93%</td>
</tr>
<tr>
<td>4996</td>
<td>1745</td>
<td>8716275</td>
<td>-1294328</td>
<td>-12.91%</td>
</tr>
<tr>
<td>4997</td>
<td>1745</td>
<td>8718020</td>
<td>-1292583</td>
<td>-12.94%</td>
</tr>
<tr>
<td>4998</td>
<td>1744</td>
<td>8714768</td>
<td>-1295835</td>
<td>-12.93%</td>
</tr>
<tr>
<td>4999</td>
<td>1744</td>
<td>8716512</td>
<td>-1294091</td>
<td>-12.96%</td>
</tr>
<tr>
<td>5000</td>
<td>1743</td>
<td>8713257</td>
<td>-1297346</td>
<td>-12.99%</td>
</tr>
<tr>
<td>65521</td>
<td>26</td>
<td>1703546</td>
<td>-8307057</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65522</td>
<td>26</td>
<td>1703572</td>
<td>-8307031</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65523</td>
<td>26</td>
<td>1703598</td>
<td>-8307005</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65524</td>
<td>26</td>
<td>1703624</td>
<td>-8306979</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65525</td>
<td>26</td>
<td>1703650</td>
<td>-8306953</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65526</td>
<td>26</td>
<td>1703676</td>
<td>-8306927</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65527</td>
<td>26</td>
<td>1703702</td>
<td>-8306901</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65528</td>
<td>26</td>
<td>1703728</td>
<td>-8306875</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65529</td>
<td>26</td>
<td>1703754</td>
<td>-8306849</td>
<td>-82.98%</td>
</tr>
<tr>
<td>65530</td>
<td>26</td>
<td>1703780</td>
<td>-8306823</td>
<td>-82.98%</td>
</tr>
</tbody>
</table>

Table 8: Rank, frequency, their product and the difference of this product from 1/10 the size of the BNC (about 10 million).
While Zipf’s law appears to hold from about the 500th word, it cannot really be said to hold for the first few hundred words. By rank 3500, this “constant” has been breached to the low side, and continues its descent. It may be argued that BNC was sampled in a subjective way, so producing some deviation from Zipf’s law. We have looked at another commonly analysed general language corpus - the Brown corpus (47218 types, 1015945 tokens), and the results show similar behaviour in their frequency profiles, although with variations at specific ranks likely due to the number of types available. Use of Zipf’s law for analysing large corpora of newswire texts (the Wall Street Journal for 1987, 1988 and 1989 of 19 million, 16 million and 6 million tokens approximately) shows similar behaviour between these collections, with little deviation due to corpus size (Ha et al 2002).
Figure 18: Plot of frequency times rank, versus rank, for the first 3000 most frequent words in the Brown corpus. Note the proximity to 1/10 (about 100,000) for these ranks in the corpus.

From ranks 1-10 in BNC (Table 8) and Brown (Table 9), the deviation from expectation is initially significant (about 30 to 40%) but begins to settle. The pattern of frequency times rank is similar between BNC and Brown (compare Figure 17 with Figure 18), and values fall away subsequently for all of these corpora. By rank 3000, groups of deviations remain around a similar level: frequencies have settled to similar values, and while further deviations become significant from the expectation, they remain within certain ranges due to ranks (compare Table 8 with Table 9). It would be difficult to infer anything from the values at the ranks shown in the two tables, apart from noting that variation is not only limited to BNC: it is apparent in other general language text corpora also.
Table 9: Rank and frequency information for the BNC and Brown corpora showing deviation from the constant expected according to Zipf.

<table>
<thead>
<tr>
<th>r</th>
<th>f</th>
<th>f * r</th>
<th>Error</th>
<th>% error</th>
<th>f</th>
<th>f * f</th>
<th>Error</th>
<th>% error</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>1</td>
<td>6187927</td>
<td>-3822676</td>
<td>-38.19%</td>
<td>1</td>
<td>69970</td>
<td>-31624.5</td>
<td>-31.13%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2941790</td>
<td>-4127023</td>
<td>-41.23%</td>
<td>2</td>
<td>36410</td>
<td>-28774.5</td>
<td>-28.32%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2682878</td>
<td>-1961969</td>
<td>-19.60%</td>
<td>3</td>
<td>28854</td>
<td>-15932.5</td>
<td>-14.80%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2560346</td>
<td>230781</td>
<td>+2.31%</td>
<td>4</td>
<td>26154</td>
<td>3021.5</td>
<td>+2.97%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1883295</td>
<td>1289167</td>
<td>12.88%</td>
<td>5</td>
<td>23363</td>
<td>15220.5</td>
<td>14.98%</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1115382</td>
<td>-2202929</td>
<td>-22.01%</td>
<td>6</td>
<td>10594</td>
<td>-27436.5</td>
<td>-27.01%</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>1089599</td>
<td>-1294131</td>
<td>-12.93%</td>
<td>7</td>
<td>10102</td>
<td>-20778.5</td>
<td>-20.45%</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>998867</td>
<td>-1020800</td>
<td>-10.20%</td>
<td>8</td>
<td>9815</td>
<td>-13259.5</td>
<td>-13.05%</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>923975</td>
<td>-770853</td>
<td>-7.70%</td>
<td>9</td>
<td>9542</td>
<td>-6174.5</td>
<td>-6.08%</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3417</td>
<td>243814</td>
<td>+2.44%</td>
<td>10</td>
<td>36</td>
<td>6441.5</td>
<td>+6.34%</td>
</tr>
<tr>
<td></td>
<td>3001</td>
<td>3414</td>
<td>238225</td>
<td>+2.38%</td>
<td>3001</td>
<td>36</td>
<td>6477.5</td>
<td>+6.38%</td>
</tr>
<tr>
<td></td>
<td>3002</td>
<td>3411</td>
<td>226300</td>
<td>+2.32%</td>
<td>3002</td>
<td>36</td>
<td>6513.5</td>
<td>+6.41%</td>
</tr>
<tr>
<td></td>
<td>3003</td>
<td>3409</td>
<td>210033</td>
<td>+2.30%</td>
<td>3003</td>
<td>36</td>
<td>6549.5</td>
<td>+6.45%</td>
</tr>
<tr>
<td></td>
<td>3004</td>
<td>3409</td>
<td>213442</td>
<td>+2.33%</td>
<td>3004</td>
<td>36</td>
<td>6585.5</td>
<td>+6.48%</td>
</tr>
<tr>
<td></td>
<td>3005</td>
<td>3408</td>
<td>213845</td>
<td>+2.34%</td>
<td>3005</td>
<td>36</td>
<td>6621.5</td>
<td>+6.52%</td>
</tr>
<tr>
<td></td>
<td>3006</td>
<td>3408</td>
<td>217253</td>
<td>+2.37%</td>
<td>3006</td>
<td>36</td>
<td>6657.5</td>
<td>+6.55%</td>
</tr>
<tr>
<td></td>
<td>3007</td>
<td>3405</td>
<td>213637</td>
<td>+2.31%</td>
<td>3007</td>
<td>36</td>
<td>6693.5</td>
<td>+6.59%</td>
</tr>
<tr>
<td></td>
<td>3008</td>
<td>3404</td>
<td>210233</td>
<td>+2.32%</td>
<td>3008</td>
<td>36</td>
<td>6729.5</td>
<td>+6.62%</td>
</tr>
<tr>
<td></td>
<td>3009</td>
<td>3404</td>
<td>2047040</td>
<td>+2.35%</td>
<td>3009</td>
<td>36</td>
<td>6765.5</td>
<td>+6.66%</td>
</tr>
<tr>
<td></td>
<td>3010</td>
<td>3404</td>
<td>2046040</td>
<td>+2.35%</td>
<td>3010</td>
<td>36</td>
<td>6765.5</td>
<td>+6.66%</td>
</tr>
<tr>
<td></td>
<td>4991</td>
<td>1746</td>
<td>-1298063</td>
<td>-12.95%</td>
<td>4991</td>
<td>19</td>
<td>-6765.5</td>
<td>-6.66%</td>
</tr>
<tr>
<td></td>
<td>4992</td>
<td>1746</td>
<td>-1296317</td>
<td>-12.98%</td>
<td>4992</td>
<td>19</td>
<td>-6746.5</td>
<td>-6.64%</td>
</tr>
<tr>
<td></td>
<td>4993</td>
<td>1745</td>
<td>-1299563</td>
<td>-12.96%</td>
<td>4993</td>
<td>19</td>
<td>-6727.5</td>
<td>-6.62%</td>
</tr>
<tr>
<td></td>
<td>4994</td>
<td>1745</td>
<td>-1297818</td>
<td>-12.95%</td>
<td>4994</td>
<td>19</td>
<td>-6708.5</td>
<td>-6.60%</td>
</tr>
<tr>
<td></td>
<td>4995</td>
<td>1745</td>
<td>-1296073</td>
<td>-12.93%</td>
<td>4995</td>
<td>19</td>
<td>-6689.5</td>
<td>-6.58%</td>
</tr>
<tr>
<td></td>
<td>4996</td>
<td>1745</td>
<td>-1294328</td>
<td>-12.91%</td>
<td>4996</td>
<td>19</td>
<td>-6670.5</td>
<td>-6.57%</td>
</tr>
<tr>
<td></td>
<td>4997</td>
<td>1745</td>
<td>-1292583</td>
<td>-12.94%</td>
<td>4997</td>
<td>19</td>
<td>-6651.5</td>
<td>-6.55%</td>
</tr>
<tr>
<td></td>
<td>4998</td>
<td>1744</td>
<td>-1295835</td>
<td>-12.93%</td>
<td>4998</td>
<td>19</td>
<td>-6632.5</td>
<td>-6.53%</td>
</tr>
<tr>
<td></td>
<td>4999</td>
<td>1744</td>
<td>-1294091</td>
<td>-12.96%</td>
<td>4999</td>
<td>19</td>
<td>-6613.5</td>
<td>-6.51%</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>1743</td>
<td>-1297346</td>
<td>-12.99%</td>
<td>5000</td>
<td>19</td>
<td>-660000</td>
<td>-6.49%</td>
</tr>
<tr>
<td>Brown</td>
<td>65521</td>
<td>26</td>
<td>1703546</td>
<td>-8307057</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65522</td>
<td>26</td>
<td>1703572</td>
<td>-8307031</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65523</td>
<td>26</td>
<td>1703598</td>
<td>-8307005</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65524</td>
<td>26</td>
<td>1703624</td>
<td>-8306979</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65525</td>
<td>26</td>
<td>1703650</td>
<td>-8306953</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65526</td>
<td>26</td>
<td>1703676</td>
<td>-8306927</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65527</td>
<td>26</td>
<td>1703702</td>
<td>-8306901</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65528</td>
<td>26</td>
<td>1703728</td>
<td>-8306875</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65529</td>
<td>26</td>
<td>1703754</td>
<td>-8306849</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65530</td>
<td>26</td>
<td>1703780</td>
<td>-8306823</td>
<td>-82.98%</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It appears that the deviations from Zipf’s law are potentially signatures of language: although the law does not strictly hold, they do follow some kind of pattern. Zipf’s law has been modified into the Zipf-Mandelbrot law, which may better explain this initial behaviour and subsequent tail-off. The modification provided by Zipf-Mandelbrot reportedly provides a better fit to ranks below 100. This may be worth exploring subsequently.

85
Chapter 3

Now, consider a specialist corpus containing texts about Nanoscale Science and Design. This corpus comprises 1,012,096 word-tokens with 26,861 different word-types. The distribution of these tokens, frequency vs rank distribution that is, shows a range of extremes:

1. 10 most frequent types – the, of, in, and, a, to, I, is, article, first – comprise over 23% of all the tokens in the corpus – over 220,000 tokens (compared with 22% for first 10 of BNC, and 7 shared words between these two sets of 10)

2. There are 11626 word-types that are only used once each (frequency of 1), a further 3627 used only twice (two tokens each), 1882 used only 3 times, and so on.

3. Zipf’s power-law suggests that the product of frequency and rank should be a constant and that the product is about a 10th of the size of the corpus; for Nanoscale Science and Design, therefore, Zipf’s constant is around 100,000. Again, an approximation to the constant only holds only for a part of the corpus (Figure 19).

![Figure 19: Graph of frequency multiplied by rank (Zipf’s law) for all words in the 1 million word nanoscale science and design corpus.](image)

Though the top 100 types account for 50% of tokens in the corpus, and there is some pattern to the next few thousand words following Zipf, the predominance of low frequency words becomes important. About 53% of words in the British National Corpus have a frequency of 1 (appear in the hapax legomena). For these words, their rank in a frequency list is arbitrary, but they could
only provide values around the 1/10 corpus constant if they began around a rank of 10,000,000. Since BNC only contains about 669,000 words, this is not possible.

Having considered certain behaviours of the BNC, Brown and a Nanoscale Science and Design corpora, we now consider aspects of five specialist corpora (including the Nanoscale corpus) of varying sizes from 166,000 words to over 1 million words. We are looking for commonalities in behaviour of specialist corpora. If we accept a generalisation of behaviours of high-frequency words, we now look at low frequency words in these corpora.

We have compared the number of words (types) occurring at frequencies from 1 to 10 in our 5 specialist corpora with a predicted derived value not found in literature\textsuperscript{14}. This derivation shows that the proportion of words occurring with a given frequency is equal to \(1/\left(f(f+1)\right)\), where \(f\) is the frequency. The proportion of words occurring once is 50%, twice is 12%, and so on. The results of our analysis are shown in Table 10, and are shown graphically in Figure 20 and Figure 21.

<table>
<thead>
<tr>
<th>Freq(word)</th>
<th>Breast Cancer</th>
<th>Automotive</th>
<th>Nuclear</th>
<th>Finance</th>
<th>Nanoscale</th>
<th>BNC</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>166044</td>
<td>350920</td>
<td>393993</td>
<td>685037</td>
<td>1012096</td>
<td>100106029</td>
<td></td>
</tr>
<tr>
<td>Types</td>
<td>10036</td>
<td>14252</td>
<td>14937</td>
<td>28793</td>
<td>26861</td>
<td>669417</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>40.20%</td>
<td>38.44%</td>
<td>40.55%</td>
<td>36.76%</td>
<td>43.28%</td>
<td>52.84%</td>
<td>50.00%</td>
</tr>
<tr>
<td>2</td>
<td>14.92%</td>
<td>14.14%</td>
<td>15.18%</td>
<td>15.20%</td>
<td>13.50%</td>
<td>12.92%</td>
<td>16.70%</td>
</tr>
<tr>
<td>3</td>
<td>8.33%</td>
<td>7.14%</td>
<td>6.73%</td>
<td>8.58%</td>
<td>7.01%</td>
<td>6.02%</td>
<td>8.30%</td>
</tr>
<tr>
<td>4</td>
<td>5.55%</td>
<td>5.10%</td>
<td>4.85%</td>
<td>5.78%</td>
<td>4.66%</td>
<td>3.71%</td>
<td>5.00%</td>
</tr>
<tr>
<td>5</td>
<td>3.76%</td>
<td>3.64%</td>
<td>3.29%</td>
<td>3.99%</td>
<td>3.46%</td>
<td>2.54%</td>
<td>3.30%</td>
</tr>
<tr>
<td>6</td>
<td>2.81%</td>
<td>2.69%</td>
<td>3.05%</td>
<td>2.96%</td>
<td>2.64%</td>
<td>1.90%</td>
<td>2.40%</td>
</tr>
<tr>
<td>7</td>
<td>2.23%</td>
<td>2.15%</td>
<td>2.19%</td>
<td>2.30%</td>
<td>1.95%</td>
<td>1.51%</td>
<td>1.80%</td>
</tr>
<tr>
<td>8</td>
<td>2.08%</td>
<td>1.78%</td>
<td>1.83%</td>
<td>1.82%</td>
<td>1.73%</td>
<td>1.20%</td>
<td>1.40%</td>
</tr>
<tr>
<td>9</td>
<td>1.70%</td>
<td>1.80%</td>
<td>1.41%</td>
<td>1.54%</td>
<td>1.33%</td>
<td>0.98%</td>
<td>1.10%</td>
</tr>
<tr>
<td>10</td>
<td>1.29%</td>
<td>1.48%</td>
<td>1.33%</td>
<td>1.45%</td>
<td>1.19%</td>
<td>0.83%</td>
<td>0.90%</td>
</tr>
<tr>
<td>Total</td>
<td>82.86%</td>
<td>78.36%</td>
<td>80.42%</td>
<td>80.38%</td>
<td>80.76%</td>
<td>84.45%</td>
<td>90.90%</td>
</tr>
</tbody>
</table>

Table 10: Percentages of words appearing at low frequencies within 7 corpora, 2 representing general language, 2 representing specialist corpora, alongside a predicted value

\textsuperscript{14} Descriptions of this derivation can be found at: http://www.cs.cornell.edu/Courses/cs630/2003fa/lectures/ir1-basics.pdf; http://linkage.rockefeller.edu/wli/zipf/cmpsci546_spring2002_notes.pdf; and http://ciir.cs.umass.edu/cmpsci646/Slides/ir10%20text%20stats%20up.pdf (June 2004)
Figure 20: Plotted trends from the above table of percentages of words occurring in the seven corpora

It was predicted that 50% of words would have frequency 1. BNC appears to be closest to predicted values to begin with.

Figure 21: Difference in percentage between predicted quantity of words at frequencies of 1 to 10 (x-axis) made using Zipf's law, and the actual quantities in 6 corpora. Note values at frequencies around 3 and 4 tend to become positive for specialist corpora.
All the specialist corpora, however, have values around 40% for words of frequency 1. Summation of predicted values leads to the estimation that 90% of words occur with frequency of 10 or less. BNC is closest to this value at around 85%, whereas specialist corpora have around 80% of words at these frequencies. 43% of words (types) in the nanoscale science and design corpus have frequency 1. 80.8% of words here have a frequency of 10 or less. This compares with the 84.45% in BNC, and indicates the potential for convergence of such values for corpora, even when considering specialist and general language.

A first inference from these analyses is that specialist corpora follow a consistent pattern in that all these specialist corpora deviate in the same manner from Zipf’s law. High frequency words follow similar patterns. Furthermore, there is a strong difference with the BNC when low frequency words are considered: 53% of word-types in BNC only appear once, over 10% more than in any of the specialist corpora. Word-types occurring with frequencies from 2 to 10 consistently make up a lower percentage of available types in contrast to the specialist corpora. Although we have this deviation from Zipf’s law, the reduced values for specialist corpora tend to better fit with Zipf’s principle of least effort in language use, whereby the vocabulary is balanced between more words and more meanings (Zipf 1949, pp22-23): semantic shifts provide further evidence of attempts to retain such a balance. Zipf further proposes that a constant results by multiplying the number of words occurring at a given frequency with the square of the frequency itself. Zipf somewhat stretches the notion of constant by suggesting that the values 27,844 and 18,000 are approximately the same - also 2,922 and 5,996 for a different set of text (Zipf 1949 pp32-33).

Zipf originally tested his law against Joyce’s text *Ulysses*: about 30,000 types that make up 260,000 tokens (16,400 occurring once only). Zipf’s principle of least effort should have theorized that as the number of tokens increases, the number of types would increase at a reduced rate. We can show this by splitting the Nanoscale Science and Design corpus into 3 sections, 2 of about 0.25 million words, one of about 0.5 million words.

<table>
<thead>
<tr>
<th>Tokens</th>
<th>Types</th>
<th>Average frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>257,541</td>
<td>12,678</td>
<td>20.31</td>
</tr>
<tr>
<td>253,399</td>
<td>13,310</td>
<td>19.04</td>
</tr>
<tr>
<td>501,156</td>
<td>18,266</td>
<td>27.44</td>
</tr>
<tr>
<td>1,012,096</td>
<td>26,861</td>
<td>37.68</td>
</tr>
</tbody>
</table>

Table 11: Amounts of tokens and types, and their frequency ratio, in subparts of the Nanoscale Science and Design corpus

When the corpus size quadrupled, the number of types approximately doubled. We would need to test this against various sizes of randomly created subcorpora in order to determine how well this
relationship holds, but Zipf’s prediction may well be better than his law, specifically in relation to specialist corpora.

Without comparison to further general language corpora, such as the American National Corpus, the suggestion that general language corpora provide a better match to such predictions cannot be made with confidence, although the similarities in behaviours of specialist corpora is certainly worthy of further exploration. The simple similarities in values between words at low frequencies, for example, is of interest, and although Zipf’s law does not make ideal predictions, there are certain predictive powers we now have regarding both high and low frequency words which may enable us to generalize our results.

The combination of these high and low frequency analyses shows a pattern in language whereby corpora are dominated by certain words, and the frequency curves have reasonably consistent shapes. Consider the similarity for rank x frequency for the BNC (Figure 17), compared to the first 1000 words of the nanoscale science and design corpus (Figure 22). The similarities alluded to between corpora tends to indicate that comparing corpora will be useful in determining characteristics specific to any given corpus.

![Figure 22: Graph of frequency multiplied by rank (Zipf's law) for the first 1000 words in the 1 million word nanoscale science and design corpus.](image)

Z Score

A “Z score” can be used to determine the relative importance of specific items within given collections. The score makes a comparison in relation to the mean and standard deviation giving,
effectively, a strength above the standard deviation of the selected item. The score can be defined as:

\[ Z_x = \frac{X - U_x}{\sigma_x} \]

Where:

- \( X \) = frequency of word
- \( U_x \) = average frequency of all words in the corpus
- \( \sigma_x \) = standard deviation of word frequencies

We can use this score to characterise word frequencies within corpora by determining how many words are at what strength. The average frequency in the BNC is 149.54 (standard deviation about 11207). There are 25149 words in BNC with frequency of 150 or greater (over 96% of BNC words have below average frequency). A Z-score of various values provides the percentage of words shown in Table 12 below.

<table>
<thead>
<tr>
<th>Z-score</th>
<th>Breast Cancer</th>
<th>Automotive</th>
<th>Nuclear</th>
<th>Finance</th>
<th>Nano</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>166044</td>
<td>350920</td>
<td>393993</td>
<td>685037</td>
<td>1012096</td>
<td>1001000</td>
</tr>
<tr>
<td>Types</td>
<td>10036</td>
<td>14252</td>
<td>14937</td>
<td>28793</td>
<td>26861</td>
<td>669417</td>
</tr>
<tr>
<td>0</td>
<td>11.917%</td>
<td>12.097%</td>
<td>10.598%</td>
<td>10.798%</td>
<td>8.883%</td>
<td>3.757%</td>
</tr>
<tr>
<td>1</td>
<td>1.265%</td>
<td>0.996%</td>
<td>0.964%</td>
<td>0.743%</td>
<td>0.778%</td>
<td>0.139%</td>
</tr>
<tr>
<td>2</td>
<td>0.548%</td>
<td>0.449%</td>
<td>0.462%</td>
<td>0.320%</td>
<td>0.387%</td>
<td>0.064%</td>
</tr>
<tr>
<td>3</td>
<td>0.359%</td>
<td>0.274%</td>
<td>0.274%</td>
<td>0.208%</td>
<td>0.253%</td>
<td>0.040%</td>
</tr>
<tr>
<td>4</td>
<td>0.279%</td>
<td>0.196%</td>
<td>0.181%</td>
<td>0.170%</td>
<td>0.190%</td>
<td>0.030%</td>
</tr>
<tr>
<td>5</td>
<td>0.189%</td>
<td>0.147%</td>
<td>0.134%</td>
<td>0.139%</td>
<td>0.164%</td>
<td>0.024%</td>
</tr>
<tr>
<td>6</td>
<td>0.179%</td>
<td>0.119%</td>
<td>0.114%</td>
<td>0.122%</td>
<td>0.134%</td>
<td>0.021%</td>
</tr>
<tr>
<td>7</td>
<td>0.169%</td>
<td>0.112%</td>
<td>0.087%</td>
<td>0.094%</td>
<td>0.108%</td>
<td>0.018%</td>
</tr>
<tr>
<td>8</td>
<td>0.139%</td>
<td>0.084%</td>
<td>0.074%</td>
<td>0.073%</td>
<td>0.089%</td>
<td>0.016%</td>
</tr>
<tr>
<td>9</td>
<td>0.110%</td>
<td>0.070%</td>
<td>0.054%</td>
<td>0.069%</td>
<td>0.078%</td>
<td>0.014%</td>
</tr>
</tbody>
</table>

Table 12: Z-scores and the number of words remaining above that score from the British National Corpus
The data above again indicates that there is some kind of pattern to usage of words when considering those above certain strengths of the standard deviation. However, the use of percentage values hides the number of words actually being considered. While BNC has 25149 words with a Z greater than zero, the maximum number (which can be considered as rank) for the specialist corpora is 3109 for finance. At a z-score of 9, in BNC 95 types remain, for the Nanoscale and Finance corpora this is around 20, and for the remaining 3 corpora, closer to 10. Predominantly, we expect these to be the closed class words, and orders above this will provide us with little additional information.

‘Weirdness’

We have seen something of the behaviour of word frequencies for general versus specialist corpora. These behaviours are based on the entirety of the corpus, which includes the various closed-class words alluded to earlier in this section. It could be argued that in looking at high frequency words, and not rejecting these closed-class words, as noted in the previous paragraph, we are predominantly modelling the behaviour of closed class words. Special language texts tend to show a greater preponderance of specialist terms when compared to general language texts. Now, we can either remove all closed class words, or we can determine the expected frequency (probability) of closed class words and use those as rejection criteria. In the latter case, if we expect a closed class word, like “the”, to always occur a given percentage of times (generally about 6%), the fact that it occurs with such a frequency in a specialist collection is unsurprising.
When compared to a frequency distribution of general language, there are three possible behaviours:

I. Most closed-class words will occur in line with expected frequencies.

II. Open-class words – potentially terms or carrier terms – should occur comparatively more frequently, since their frequency in general-language texts will be low, or possibly even zero.

III. Some words in the general language will occur infrequently in the specialist corpus.

This imbalance in the distribution of certain words can be quantified through a ratio, referred to following Bronislaw Malinowski as *weirdness* (for example, Ahmad, Gillam and Tostevin 2000):

\[
\text{Weirdness} = \frac{R_s}{R_g}
\]

Where:

\[R_s = \frac{f_s}{n_s}; \quad f_s = \text{frequency of word in specialist language corpus}\]
\[\& \quad n_s = \text{total count of words in specialist language corpus};\]
\[R_g = \frac{f_g}{n_g}; \quad f_g = \text{frequency of word in general language corpus}\]
\[\& \quad n_g = \text{total count of words in general language corpus}\]

\(R_s\) represents the relative frequency of the occurrence of a word/term in a text sample of \(n_s\) words;
\(R_g\) represents the relative frequency of the same word/term in a general language corpus.

Results of applying the formula are:

I'. Closed-class words, such as "the" occur with weirdness around unity.

II'. Open-class words – potentially terms or carrier terms – occur with high values. If the word does not appear in the general language corpus, the result will be \(Rs/0\) – represented as "infinite" (some computer systems represent this as NaN – Not a Number).

III'. Words in the general language occurring infrequently in the specialist corpus will produce a weirdness value somewhat below unity.

Application of weirdness produces a profile of word use with values ranging from near zero to "infinite". Frequent words in the sample that appear infrequently in the general language corpus
will produce large weirdness values. This category of words is interesting for our purposes, although the other categories may be interesting for other purposes (e.g. authorship attribution).

Weirdness values can be computed using a component of System Quirk called KonText that can compute frequency distribution of tokens in specialist texts and has access to frequency distribution of 'representative' general language (Ahmad 1995, Holmes-Higgin 1995); a good example of such a collection is the British National Corpus (Aston and Burnard 1998). This method produces lists of single-token words that may be productive in making longer terms, and still others that may exist only when carried in compounds. The result of applying this mechanism suggests candidate words for further analysis.

For the five specialist corpora, at least 12% of words are occurring with “infinite” weirdness (Table 13). The Nanoscale Science and Design corpus has over 38% of words not occurring in the general language. Of these 10231 “infinites”, 6441 occur with frequency 1. If we consider there may be spelling errors within these and only retain words with frequency of 3 or greater, we still have 2322 “infinite” weirdness words remaining.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Types</th>
<th>Tokens</th>
<th>Number of types not in BNC</th>
<th>% types not in BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>10036</td>
<td>166044</td>
<td>1368</td>
<td>14%</td>
</tr>
<tr>
<td>Automotive</td>
<td>14252</td>
<td>350920</td>
<td>1794</td>
<td>13%</td>
</tr>
<tr>
<td>Nuclear</td>
<td>14937</td>
<td>393993</td>
<td>4190</td>
<td>28%</td>
</tr>
<tr>
<td>Finance</td>
<td>28793</td>
<td>685037</td>
<td>4718</td>
<td>16%</td>
</tr>
<tr>
<td>Nanoscale</td>
<td>26861</td>
<td>1012096</td>
<td>10231</td>
<td>38%</td>
</tr>
</tbody>
</table>

Table 13: Quantity of “infinite” weirdness words in the five specialist corpora: note the large numbers involved in the Nuclear and Nanoscale corpora in contrast to the Breast Cancer, Automotive, and Finance corpora.

The words occurring with “infinite” weirdness present a problem: although this ‘metric’ can be used to signal that the token may be a candidate term, their treatment requires human interpretation. We cannot compare significance numerically. As we wish to explore the notion that frequency correlates with acceptability, we would like to combine the information gained from weirdness with frequency information, and use this to suggest whether a specific word is more likely to be identifiable in a given specialism.

Since “infinite” weirdness occurs when the frequency in the general language is zero, when we have an unseen type (a type of zero probability), we need to consider how to provide a value for expectation of existence in general language: a means to treat unseen types. Various estimation and smoothing techniques have been suggested in the literature, principally for predictions of bigrams (e.g. Manning and Schutze 1999, p196 et seq; Chen and Goodman 1996; Gale 1995; Gale
and Church 1990). The “Add-one” approach cited by Gale and Church adjusts frequency using a renormalization factor described as:

\[
    r^* = (r + 1) \frac{N}{N + S}
\]

Where:

- \( r \) = initial frequency of word
- \( N \) = tokens in corpus
- \( S \) = number of types (words) in corpus

We can use this in the weirdness calculation to ensure a non-zero denominator by calculating a renormalization factor for the BNC. We have:

\[
    r^* = \frac{100106029}{100106029 + 669417} \approx r + 1
\]

The renormalization factor can be reduced to adding 1 to each frequency in the weirdness calculation, which in turn leads us to modify the calculation of weirdness, replacing \( R_g = f_g/n_g \) by \( R_g = (1+f_g)/n_g \). The resulting shift of all values upwards from zero acts like an affine transformation: all results remain colinear. The calculation of this modified value for weirdness leads to a value for our unseen types. Since the behaviour of words in specialist corpora has, earlier in this chapter, been shown to be reasonably consistent at low frequencies (to 10), and words at higher frequencies appear to show consistent behaviour in both general and specialist corpora, add-one would appear, initially, to be a more suitable means by which to treat unseen types (“infinite” weirdness words). For every word (type) in any specialist corpus, we can now produce a finite number for its weirdness. The actual effect of using this mechanism requires further study, although it appears to be a promising means. While better estimators and smoothing factors may be of use, for example in the works cited above, add-one will suffice for current purposes.

Consider, now, analysis of the Nanoscale Science and Design Corpus, for which we have used the BNC as the basis to generate weirdness values. Previously, as shown in Table 13 above, 38% of weirdness values would be “infinite”. Now, we have numeric values for weirdness of previously infinite values, for example, as shown below in Table 14. Note how this produces a different ranking to that obtained from frequency information. These values for weirdness can now be analysed further.
### Table 14: New weirdness value calculated using the add-one smoothing for selected words from the Nanoscale Science and Design corpus

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq</th>
<th>Weirdness</th>
<th>BNC</th>
<th>Smoothed Weirdness</th>
</tr>
</thead>
<tbody>
<tr>
<td>nanowires</td>
<td>619</td>
<td>INF</td>
<td>0</td>
<td>61225</td>
</tr>
<tr>
<td>nanoparticles</td>
<td>829</td>
<td>81996</td>
<td>1</td>
<td>40998</td>
</tr>
<tr>
<td>nanowire</td>
<td>360</td>
<td>INF</td>
<td>0</td>
<td>35607</td>
</tr>
<tr>
<td>nanotube</td>
<td>969</td>
<td>47921</td>
<td>2</td>
<td>31948</td>
</tr>
<tr>
<td>nanoscale</td>
<td>268</td>
<td>INF</td>
<td>0</td>
<td>26508</td>
</tr>
<tr>
<td>tunneling</td>
<td>514</td>
<td>50839</td>
<td>1</td>
<td>25420</td>
</tr>
<tr>
<td>nanoparticle</td>
<td>232</td>
<td>INF</td>
<td>0</td>
<td>22947</td>
</tr>
<tr>
<td>nanotubes</td>
<td>1379</td>
<td>27279</td>
<td>5</td>
<td>22733</td>
</tr>
<tr>
<td>exciton</td>
<td>215</td>
<td>INF</td>
<td>0</td>
<td>21266</td>
</tr>
<tr>
<td>nanostructures</td>
<td>212</td>
<td>INF</td>
<td>0</td>
<td>20969</td>
</tr>
<tr>
<td>phonon</td>
<td>179</td>
<td>INF</td>
<td>0</td>
<td>17705</td>
</tr>
<tr>
<td>mwnts</td>
<td>176</td>
<td>INF</td>
<td>0</td>
<td>17408</td>
</tr>
<tr>
<td>technol</td>
<td>348</td>
<td>34420</td>
<td>1</td>
<td>17210</td>
</tr>
<tr>
<td>nanorods</td>
<td>159</td>
<td>INF</td>
<td>0</td>
<td>15727</td>
</tr>
<tr>
<td>photoluminescence</td>
<td>270</td>
<td>26705</td>
<td>1</td>
<td>13353</td>
</tr>
<tr>
<td>nanocrystals</td>
<td>395</td>
<td>19534</td>
<td>2</td>
<td>13023</td>
</tr>
</tbody>
</table>

Plotting the log of frequency against the log of this new weirdness provides the three areas identified previously (Figure 24, with reference to I-III, p93). With log-rank on the x-axis and log-weirdness on the y-axis, we see:

I”. “Shared” closed-class words, occurring around zero (log 1) on the x-axis.

II”. Open-class words – potentially terms or carrier terms – occur with high weirdness values.

III”. Words in the general language occurring infrequently in the specialist corpus towards the left of the x-axis.
Using this modified weirdness calculation, the values for the top 1000 most weird words display the kind of properties for which Zipf's law again becomes a consideration (Figure 25).

Figure 25: 1000 most weird words in the nanoscale corpus, using the modified version of weirdness
A finer-grained analysis of the 100 most weird words shows some variation in smoothness between ranks 1-100 due to words of similar frequencies (Figure 26).

![Figure 26: 100 most weird words in the nanoscale corpus, using the modified version of weirdness](image)

Having series for both frequency and weirdness values, we can consider the means by which to produce lists of highly frequent (correlating with acceptability), highly weird (specialist) words. A combination of these two will provide a minimum criteria for word selection. Consideration purely of weirdness may lead to instances where low-frequency words are selected ahead of high-frequency words of lower weirdness.

From before, a z-score on frequency produced varying number of words to consider. Suppose we now adopt z-scores for weirdness also, and consider the combination of equal z-scores for frequency and weirdness, what kind of results can we expect? Analysis of the five specialist corpora shows the number of words produced using equal values for both frequency and weirdness (Table 15).
Table 15: Number of words produced by considering a combined value for z-score for both frequency and weirdness

As the number of words increases, the number of words at increasing z-scores varies for the different corpora. When normalized against the number of words in each corpus, the results follow similar patterns due to behaviours at high frequencies as noted before (Figure 27). The finance corpus, it appears, consistently produces a lower proportion of words above the standard deviation. Perhaps this is consistent with a low proportion of weird words according to its size (Table 13)? The breast cancer corpus provides an upper bound here. Perhaps a characteristic due to corpus size? Between these bounds, similar patterns are seen for the remaining corpora.
Figure 27: Normalized counts of highly frequent-highly weird words characterised against the z-scores that produced them

Consider, now, the frequencies and weirdness values contributing to these numbers of words for the Nanoscale Science and Design corpus (Table 16).

<table>
<thead>
<tr>
<th>z-score</th>
<th>Min frequency</th>
<th>Max frequency</th>
<th>Number words selected (frequency)</th>
<th>Min weirdness</th>
<th>Max weirdness</th>
<th>Number words selected (weirdness)</th>
<th>Combined number of words selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>4763</td>
<td>4763</td>
<td>44</td>
<td>24795</td>
<td>24795</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2553</td>
<td>4763</td>
<td>51</td>
<td>11078</td>
<td>42086</td>
<td>78</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1870</td>
<td>4763</td>
<td>68</td>
<td>11078</td>
<td>42086</td>
<td>119</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>1379</td>
<td>4763</td>
<td>104</td>
<td>3877</td>
<td>42086</td>
<td>193</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>718</td>
<td>8520</td>
<td>209</td>
<td>1856</td>
<td>85161</td>
<td>411</td>
<td>19</td>
</tr>
<tr>
<td>0.9</td>
<td>580</td>
<td>8520</td>
<td>233</td>
<td>1856</td>
<td>85161</td>
<td>449</td>
<td>21</td>
</tr>
<tr>
<td>0.8</td>
<td>514</td>
<td>..</td>
<td>269</td>
<td>1322</td>
<td>..</td>
<td>498</td>
<td>24</td>
</tr>
<tr>
<td>0.7</td>
<td>476</td>
<td>..</td>
<td>312</td>
<td>1149</td>
<td>..</td>
<td>584</td>
<td>27</td>
</tr>
<tr>
<td>0.6</td>
<td>395</td>
<td>..</td>
<td>388</td>
<td>1149</td>
<td>..</td>
<td>656</td>
<td>33</td>
</tr>
<tr>
<td>0.5</td>
<td>348</td>
<td>..</td>
<td>451</td>
<td>825</td>
<td>..</td>
<td>754</td>
<td>39</td>
</tr>
<tr>
<td>0.4</td>
<td>288</td>
<td>..</td>
<td>531</td>
<td>735</td>
<td>..</td>
<td>884</td>
<td>45</td>
</tr>
<tr>
<td>0.3</td>
<td>215</td>
<td>..</td>
<td>649</td>
<td>601</td>
<td>..</td>
<td>1241</td>
<td>62</td>
</tr>
<tr>
<td>0.2</td>
<td>159</td>
<td>..</td>
<td>826</td>
<td>459</td>
<td>..</td>
<td>1572</td>
<td>79</td>
</tr>
<tr>
<td>0.1</td>
<td>99</td>
<td>..</td>
<td>1216</td>
<td>325</td>
<td>..</td>
<td>2088</td>
<td>129</td>
</tr>
<tr>
<td>0</td>
<td>38</td>
<td>..</td>
<td>2386</td>
<td>177</td>
<td>..</td>
<td>4551</td>
<td>352</td>
</tr>
</tbody>
</table>

Table 16: Number of words selected by z-score: range of frequency and weirdness values at each z-score
At \( z=5 \), the single remaining selected word has neither highest frequency (\( =65178 \) for *the*) nor highest weirdness value (\( 85161 \)) but is the best resulting combination of these factors. The table displays a tendency of maximum values in the range to increase towards a limit and subsequently remain at this limit, while the minimum values become ever lower. The z-score for frequency and weirdness each select a number of words for consideration, however the combination limits those considered quite considerably. The majority of those words initially selected on the basis of frequency are likely to be the closed-class words, hence their exclusion by weirdness values. The initial highly weird words do not match on frequency.

This combination of frequency count, modified weirdness value, and combined z-scores produces a method for selecting words (which we may consider as candidate terms) from a specialist corpus without using prior knowledge. It may, or may not, result in the effective identification of (all) 'genuine' terms. This requires evaluation. It does, however, produce a list of high-frequency (acceptable?), high weirdness (specialist?) words, and does not rely on arbitrarily chosen (human selected) values for doing so. While evaluation will determine the effectiveness of the method, it provides a systematic approach to our analysis. Resulting words should be evaluated for terminess.

In this high frequency (acceptable), high weirdness (specialist) approach, we have not considered the effect of removing a list of closed class words from consideration. Recall that in calculating weirdness, consistently used closed-class words will occur with weirdness around 1, infrequently used closed-class words with weirdness less than 1 and some slightly more frequently used an amount above 1. Removal of these words would produce a slightly different profile to the results since words at lower frequencies will be promoted: closed class words such as determiners will generally occur at high frequencies, but with low weirdness. However, we would expect this effect to be countered by a higher average weirdness since values at the lower end of the scale are being removed, and by the fact that we are now dealing with a smaller sample size. Deciding the extent of the closed-class wordlist, and measuring the effects of its use both require treatment and evaluation.

For the Nanoscale Science and Design corpus, the top 2000 most frequent words of the BNC as a list would remove 1321 words from consideration. While frequencies for these words in the specialist corpus range between 10390 and 1, weirdness values are between 1 and 152. With average weirdness of 314, removal of these words results in an increased average weirdness value (up to 330), with frequency average down from about 38 to 18. This does not affect the highest ranked (most weird) words, however the selection criteria now produce larger results sets - 583 words for \( z=0 \) (was 352), 43 for \( z=1 \) (was 19). The effect of removing certain lists of words
requires a systematic treatment, including consideration of possibly domain-adapted words, which is beyond our current consideration, and again resulting words should be evaluated for terminess.

**TF-IDF**

Term frequency-inverse document frequency is a favourite metric of the Information Retrieval community, and measures the relevance of a document to a specific word according to a collection of documents. TF-IDF is usually used to determine document relevance against a bag of words within a query. If a word appears with high frequency within a single document in a large collection, this document will be ranked as most relevant to the query. Where words occur in large numbers of documents, their contribution is discounted by a factor of the number of documents they occur in. The result of applying this equation across a collection of documents is a ranked list of documents; the equation as such tells us little about the importance of the word compared to other words in the same collection. As it relies on an instance of a word in the first place, this method can tell us little about the terminess of a particular word, however it may be of use for other purposes.

\[
tf\text{idf}(w) = \frac{tf \cdot \log(N)}{df(w)}
\]

Where:

- \(tf\) = absolute frequency of word in a document
- \(N\) = total number of documents in the corpus
- \(df(w)\) = count of documents in which word \(w\) appears at least once

Once closed-class words have been discounted, we are interested in the highly frequent domain specific words that remain. Since the TF-IDF value for a word is discounted by the number of documents, and may be "normalized by document length", those words which are characteristic of the domain, and longer texts that contain them more frequently, are penalised. This may be suitable for providing descriptions of documents (although longer documents may provide better information) in relation to each document, but doesn’t produce a useful value for our purposes. Some have attempted to use TF-IDF to provide a measure of terminess in a given collection using the total frequency of the term. This reduces to a simple logarithmic penalty function inversely proportional to the number of documents that contain the word. The intended use of TF-IDF is to provide a series of weights for each term in every document in the collection such that the contribution to a document ranking can be determined. This author considers the use of TF-IDF for determining terminess to be contrived at minimum, and this will not be of further value. There
are other information retrieval metrics (cosine distance, Okapi, BM25 and so forth), which will also not be considered further for the reasons presented here.

3.3.2 Extracting multiword terms

While some single words may be terms in their own right, in specialist domains key concepts are likely to be lexicalised by multiword terms. Multiword terms in special languages may appear to be created in a more systematic way that conforms to the terminological system of the specialism. Zipf's principle of least effort may be undermined by words being combined into multiword terms and a profusion of such multiword terms, which somehow pack the semantics of their constituents (Halliday and Martin pp29-31).

Multiword terms in English generally exclude closed class words like prepositions and determiners from their constitution, which provides a useful heuristic in their identification. Such terms do not usually include punctuation marks or numerals, although chemical, mathematical and, indeed, nomenclatures for logic, do use a variety of hyphenation, numerals and other symbols that are important within that subject field. Highly technical texts within a specialist domain that contain a high percentage of open-class words, as shown previously, may have many of these words forming multiword expressions. These terms may form "conceptual" relations with other terms within the text and other texts.

A question to consider is this: how do we know that we can combine specific words in a certain way to produce a valid multiword expression? A number of mechanisms for identifying important words within texts have been proposed, based on the frequency of instances of the word, the distance of one word from another, the probability that two words appear together in a pattern, instances of occurrence versus a reference corpus, and so on. Collections of texts of various sizes and types have been analysed based on various theories of grammar, syntax and semantics. This section considers some predominantly statistical mechanisms.

Lexical phrases

Since multiword terms in English generally exclude closed class words like prepositions and determiners from their constitution, it is possible to consider using a list of these in semi-automatically identifying lexical phrases. Such a list, as alluded to at the beginning of the previous section, requires prior (domain) knowledge in order that domain specific words are not included in such a list. This heuristic for detecting multiword terms has been variously used and is implemented within System Quirk in an application called Ferret. It is reasonably effective for
English, but on languages such as French may have limited effectiveness where terms tend to include such closed class words.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Term candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>387</td>
<td>carbon nanotubes</td>
</tr>
<tr>
<td>346</td>
<td>room temperature</td>
</tr>
<tr>
<td>168</td>
<td>field emission</td>
</tr>
<tr>
<td>156</td>
<td>quantum dots</td>
</tr>
<tr>
<td>126</td>
<td>transmission electron microscopy</td>
</tr>
<tr>
<td>106</td>
<td>optical properties</td>
</tr>
<tr>
<td>95</td>
<td>carbon nanotube</td>
</tr>
<tr>
<td>89</td>
<td>magnetic field</td>
</tr>
<tr>
<td>88</td>
<td>wetting layer</td>
</tr>
</tbody>
</table>

Table 17: Indicative results from System Quirk's Ferret application for the nanoscale science and design corpus. Note the extraction of “carbon nanotube” (singular) and “carbon nanotubes” (plural form).

The criteria for selection of these lexical phrases, while simple, tends to ignore our word selection criteria presented above. While acknowledging the utility of such a mechanism, and the potential for possibly validating statistical extraction, we wish to explore statistical approaches to creation of multiword terms. The mutually validating approach has been presented elsewhere (Gillam and Ahmad 2002).

**Mutual Information**

Mutual Information was proposed as an *association ratio* based on the “information theoretic concept of mutual information” (Church and Hanks 1989). The ratio compares the frequency of occurrences of two words within a chosen distance of each other, to the relative frequency of the individual words within the text collection.

\[
\text{MI}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}
\]

Where:

- \(P(x)\) = relative frequency of word \(x\) in corpus
- \(P(y)\) = relative frequency of word \(y\) in corpus
- \(P(x,y)\) = count of times words \(x\) and \(y\) appear within \(w\) words of each other

Church and Hanks discuss the possible asymmetries between \(P(x,y)\) and \(P(y,x)\) – noting that this method should be order dependent. This equation depends upon the selection of two words – firstly, a word must be chosen as the subject of the information, and secondly, a word that exists
within the specified window of this word must be chosen as the second for the computation to occur. Positive MI indicates that the two words occur together more frequently than by chance.

**T-Score**

MI has been "improved" by others including Jacquemin, who criticises the use of MI as "not the most appropriate measure to establish differences among nearly synonymous words" (Jacquemin 2001 p34). Oddly, this is the first and only mention given by Jacquemin in this section of his book to synonymy, and the notion of conditional probabilities is quickly returned to with the t-score. He also makes claims about order dependence, which cannot be derived directly from MI (without consideration of $P(y|x)$). Jacquemin offers Church’s t-score, without saying whether he thinks it more appropriate than MI. Calculation of the t-score can be carried out using the following equation:

$$T(x,y) = \frac{P(x,y) - P(x)P(y)}{\sqrt{\frac{P(x,y)^2}{N}}}$$

Manning and Schütze provide examples of t-scores, for which the null hypothesis is only rejected above 2.576, and they note that MI and t-score result in identical rankings for the collocations they present. Some of the results they provide, with a corpus of 14,307,668 words as follows:

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Frequency</th>
<th>Word 2</th>
<th>Frequency</th>
<th>MI</th>
<th>t-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ayatollah</td>
<td>42</td>
<td>Ruhollah</td>
<td>20</td>
<td>18.38</td>
<td>4.4721</td>
</tr>
<tr>
<td>Bette</td>
<td>41</td>
<td>Midler</td>
<td>27</td>
<td>17.98</td>
<td>4.4721</td>
</tr>
<tr>
<td>first</td>
<td>14907</td>
<td>made</td>
<td>9017</td>
<td>1.09</td>
<td>2.3714</td>
</tr>
<tr>
<td>time</td>
<td>15019</td>
<td>last</td>
<td>15629</td>
<td>0.29</td>
<td>0.8036</td>
</tr>
</tbody>
</table>

Table 18: Example of use of Mutual Information (Manning and Schütze 1999, combined from p166 and p179)

The selection of words for treatment by both MI and t-test seems to be arbitrary. Perhaps both metrics could be systematized using our weirdness-frequency combination to make this initial selection. However, both metrics have a further difficulty: no account is taken of features of the neighbourhood of the selected word(s). It is quite possible that two words may appear mutually important using these measures, but that they appear with an even frequency across all the positions being considered and hence there is no significant collocation pattern. Furthermore, both metrics consider two words; collocations may be of any length. Any co-occurrence within the neighbourhood is deemed to be important, or alternatively MI / t-score has to be calculated at every available position. Work by Frank Smadja on collocations using the Xtract software system is interesting here.
"XTract"ing Collocations

Consideration of the importance of the individual positions within the neighbourhood of a particular word is a key characteristic of Smadja’s work on collocations (Smadja 1993). Smadja analyses a neighbourhood of five words preceding and following a nucleate. The frequency of occurrences of each word at each position around the nucleate is recorded. If the nucleate and another token consistently appear together in the same positions with respect to each other, there will be a high frequency at the position of the collocating token. This is then identified as a significant collocation pattern using a U-Score – see equation below – familiar to statisticians as a measurement of variance.

\[
U - Score = \frac{\sum_{i=1}^{n} (f_i - \bar{f})^2}{n}
\]

The significance of the results produced by this equation can be determined by applying a Z-Score, encountered earlier, for the range of frequencies of the collocating words. Smadja suggests that a U-score of greater than 10, and a Z-score value greater than unity can be used as a threshold for selecting collocating words. He goes on to identify various patterns of collocations within Associated Press (AP) news-wire texts. This method initially relies on the selection of a specific word, upon which analysis is to be carried out, and a text corpus. Criteria for selecting this word are not clear beyond “interest”. Here, our weirdness-frequency selection may well be of benefit.

In contrast to MI, collocations of (interrupted or uninterrupted) arbitrary length can be determined using this method, although Smadja’s paper only says it is possible and we have to assume that he implies use of the same thresholds at each step. For extraction of “frozen phrases”, the method is modified such that only a word that represents over 75% of the overall collocation frequencies is selected as part of the phrase. This includes collocations of closed-class words; otherwise, a phrase such as “the Dow Jones Index of....” would not be extracted fully.

Since the thresholds for collocation extractions are based on frequencies, at lower frequencies they will cease to provide multiword term candidates of greater length. As Smadja notes for an analysis of a 10 million-word stock market corpus: “Xtract has only been effective at retrieving collocations for words appearing at least several dozen times in the corpus. This means that low-frequency words were not productive in terms of collocations. Out of the 60,000 words in the corpus, only 8,000 were repeated more than 50 times”. Smadja further notes “the statistical methods we use do not seem to be effective on low frequency words (fewer than 100 occurrences). Low frequencies of occurrence are much more likely in a (specialist) corpus of
100,000 to 1,000,000 words: it may not be possible to collect more than this amount for an emerging specialism. Subsequent to statistical analysis, Smadja makes use of the tagged corpus to determine common syntactic patterns, although the benefit of doing this is not particularly clear.

We are interested in the statistical analysis primarily, and in the behaviour of patterns of collocation. Here, we refer to a process of using significant collocates as inputs to a subsequent collocation phase as re-collocation. We wish to study the effects of increasing the length of multiword patterns being generated, against the expected decrease in frequency. By selecting statistically significant collocates where the collocation occurs either one word ahead or behind the nucleate, we can analyse the instances of this bigram in the sentences of the texts in which they occur. For the bigrams, we now generate a list of statistically significant collocates which we use to form trigrams, and the process is repeated until criteria for significance no longer apply. To create collocates of increasing length, we may need to consider a relaxation of the criteria for determining this significance.

We considered a mechanism for automatic selection of the most highly frequent and highly weird words. In previous work, we have used a simple heuristic for determining important collocating words – they match the selected u-score and z-score criteria and they do not occur within the 2000 most frequent words of the BNC. This heuristic has been, and remains, effective, however a more systematic approach to this selection should be considered.

The example of collocation values below (Table 19) shows a selection of words collocating with carbon (frequency of 1506) in the nanoscale science and design corpus. These words all match the criteria for U-score and z-score selection. Based on positional information, only 11 of these 17 collocations would be selected for further analysis. Smadja’s note that this method would be ineffective for words with less than a frequency of 100 would reduce consideration of this set of results to 2 collocates – nanotubes and nanotube, and would result in missing further important collocations.
Selection of the collocation *carbon nanotubes* (with frequency of 647) for re-collocation allows the behaviour of this pattern to be evaluated also. A selection of resulting collocations is shown below (Table 20).

<table>
<thead>
<tr>
<th>Collocate</th>
<th>Frequency</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>U-score</th>
<th>Z-score (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nanotubes</td>
<td>690</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>647</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td>37130.6</td>
<td>38.44</td>
</tr>
<tr>
<td>nanotube</td>
<td>252</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>229</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>4620.16</td>
<td>13.90</td>
<td></td>
</tr>
<tr>
<td>single-walled</td>
<td>77</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>503.41</td>
<td>4.10</td>
</tr>
<tr>
<td>aligned</td>
<td>94</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>74</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>466.44</td>
<td>5.05</td>
</tr>
<tr>
<td>multiwalled</td>
<td>70</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>302.42</td>
<td>3.71</td>
</tr>
<tr>
<td>amorphous</td>
<td>58</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>46</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>182.36</td>
<td>3.04</td>
</tr>
<tr>
<td>atoms</td>
<td>51</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>42</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>152.09</td>
<td>2.64</td>
</tr>
<tr>
<td>nanotips</td>
<td>44</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>133.44</td>
<td>2.25</td>
</tr>
<tr>
<td>multiwall</td>
<td>43</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>126.61</td>
<td>2.19</td>
</tr>
<tr>
<td>properties</td>
<td>77</td>
<td>6</td>
<td>6</td>
<td>21</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>3</td>
<td>119.01</td>
<td>4.10</td>
</tr>
<tr>
<td>cnts</td>
<td>62</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>35</td>
<td>2</td>
<td>1</td>
<td>100.16</td>
<td>2.70</td>
</tr>
<tr>
<td>emission</td>
<td>89</td>
<td>13</td>
<td>19</td>
<td>32</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>95.69</td>
<td>4.77</td>
</tr>
<tr>
<td>single-wall</td>
<td>34</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>84.84</td>
<td>1.66</td>
</tr>
<tr>
<td>vertically</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>84.24</td>
<td>1.80</td>
</tr>
<tr>
<td>carbon</td>
<td>108</td>
<td>21</td>
<td>16</td>
<td>12</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>16</td>
<td>54.96</td>
<td>5.84</td>
</tr>
<tr>
<td>swnts</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>54.81</td>
<td>1.64</td>
</tr>
<tr>
<td>nanofibers</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>50.81</td>
<td>1.30</td>
</tr>
</tbody>
</table>

Table 19: Collocations of *carbon* in the nanoscale 1 million word nanoscale science and design corpus

Interestingly, these seem to indicate various types of *carbon nanotubes*, i.e. *single-walled carbon nanotubes*, *aligned carbon nanotubes*, *multiwalled carbon nanotubes*, *multwall carbon nanotubes* and *single-wall carbon nanotubes*. The question we have here is: are these types of carbon nanotubes distinguished by values for the same property? If we accept as variants *single-wall* and *single-walled*, *multiwall* and *multiwalled*, does *aligned* have anything to do with *walledness*? If not, we have two forms of classification of carbon nanotubes. Determining this distinction requires expert input.

At low frequencies, not statistically validated, we appear to have *aligned single-walled carbon nanotubes* (2) and *large-diameter single-walled carbon nanotubes* (2). On investigation of the 2 independent texts containing *aligned single-walled carbon nanotubes*, both refer to a publication.
of “Electric-field-directed growth of aligned single-walled carbon nanotubes”. If it is a term in the accepted title of a publication, surely it must be a valid term? Similarly, *large-diameter single-walled carbon nanotubes* (2) are referred to in two separate text references to a further publication regarding “Electrical properties and devices of *large-diameter single-walled carbon nanotubes*”.

Our *aligned carbon nanotubes* appear to be *vertically* aligned (15, with satisfactory U and Z) and *laterally* aligned (2, with unsatisfactory U and Z). Without recourse to experts, we could suggest that these are being discriminated upon the same property. Purely using statistical evidence would disregard these terms, however it does provide indications of the potential results if, somehow, relaxed statistical evidence could be used for terminology discovery.

**Collocation span sampling**

A related method for analysis of collocation patterns has been presented by Barnbrook, who discusses determining the importance of collocating words with nucleates using variations on the existing means of calculating z-score, t-score and mutual information (MI). Barnbrook selects a nucleating word (or node as he refers to it) and its set of neighbourhoods 5 words either side. This he treats as a subcorpus, and frequencies within this subcorpus are compared to expected frequencies, based on relative frequencies of the corpus as a whole. Specific values are selected for thresholds of interest — Barnbrook does not provide clear criteria for selecting these values.

For example, Barnbrook considers *Frankenstein*: this text comprises 75,214 tokens of which the word *the* occurs 4,194 times. Taking a second word, *place* with 64 occurrences, the neighbourhood produces a subcorpus of 640 words. Frequencies of words in this subcorpus are then analysed and compared to the expected value.

**Example:** for the word *the*,

\[
\text{relative frequency (probability of occurrence)} \frac{4149}{75214} = 0.05576
\]

subcorpus expectation is \(0.05576 \times 640 = 35.69\)

frequency of *the* in subcorpus = 37

standard deviation \(\sqrt{640 \times 0.05576 \times (1 - 0.05576)} = 5.80\)

\[
z\text{-score} = \frac{37 - 35.69}{5.80} = 0.2259
\]
$$t\text{-score} = \frac{37 - 35.69}{\sqrt{37}} = 0.2154$$

$$MI = \log_2 \frac{37}{35.69} = 0.052$$

For the 3 values, Barnbrook suggests:

a. "A useful (though not very precise) cut-off measure" for z-score is around 3 (p96),

b. "Absolute statistical significance is harder to assess with the t-score, but the words with a score of 2 or over are likely to be the most interesting" (p98), and

c. a cut-off value for MI of 1.58 (p99).

Barnbrook’s values do not allow for consideration of the significance of positional information, as Smadja does, so the mechanisms present collocating words although may not actually indicate significant collocations. For the nanoscale science and design corpus, consider the collocates of nanotube shown in Table 21 (having taken a neighbourhood of 5 words and analysing the subcorpus):

<table>
<thead>
<tr>
<th>Collocate</th>
<th>Frequency</th>
<th>Rel freq (corpus)</th>
<th>Expectation</th>
<th>Stdev</th>
<th>Z-score</th>
<th>t-score</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>891</td>
<td>0.06</td>
<td>458.91</td>
<td>20.72</td>
<td>11.20</td>
<td>8.83</td>
<td>0.59</td>
</tr>
<tr>
<td>carbon</td>
<td>265</td>
<td>0.00</td>
<td>10.97</td>
<td>3.31</td>
<td>76.76</td>
<td>15.60</td>
<td>4.59</td>
</tr>
<tr>
<td>a</td>
<td>231</td>
<td>0.02</td>
<td>165.77</td>
<td>12.72</td>
<td>5.13</td>
<td>4.29</td>
<td>0.48</td>
</tr>
<tr>
<td>metal</td>
<td>49</td>
<td>0.00</td>
<td>5.83</td>
<td>2.41</td>
<td>17.89</td>
<td>6.17</td>
<td>3.07</td>
</tr>
</tbody>
</table>

Table 21: Barnbrook’s values for various collocates of nanotube in the nanoscale science and design corpus

Criteria for values for Barnbrook’s Z- and t-scores do not provide a means by which to dismiss closed-class words, however MI appears to suggest carbon and metal as interesting collocations. Smadja’s U- and Z-scores concur with this result for carbon, but the U-Score value for metal does not provide such an agreement. As Table 22 shows, this is due to metal not occuring in a dominant position across the neighbourhood.

<table>
<thead>
<tr>
<th>Collocate</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
<th>+4</th>
<th>+5</th>
<th>U-Score</th>
<th>Z-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>metal</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>5.49</td>
<td>5.804959</td>
</tr>
</tbody>
</table>

Table 22: Smadja’s values for the collocation of metal with nanotube in the nanoscale science and design corpus

While all 3 measures pass Barnbrook’s criteria (perhaps the values for these criteria are not sufficient for all sorts of corpus analysis, or for larger sizes of corpora), Smadja’s U-Score criteria
is not met since there is little distinction in frequency amongst the various positions. These mechanisms may provide some supporting validation for collocations, though they do not appear to be so useful for actually extracting them.

3.3.3 Word variants

Having explored both single words and multiword extraction, indicative lists of words and phrases can be considered to form a set of candidate terms. These lists may contain evidence of variants. In the case of variants due to word ordering, as conveniently ignored in the bag-of-words approach of many search engines, these variants would be expected to show some similarities (e.g. Bowker 1998). The existence of slight variants of terms, including pluralisation where only a few characters generally vary and orthographic variations, e.g. characterize/characterise and tunnelling/tunneling, suggests the possibility to group term candidates to provide further evidence of, possibly, concepts, or of some sort of hyponymy or meronymy ordering. Specialists observe many varieties, species, objects and classes. One variety/specie/object/class does not call for a specialism. Plurals may more strongly indicate the existence of a concept, and pluralised multiword expressions more so. Specialist writing is not only suffused with open-class words, especially nouns, but also with plurals of these nouns. For special languages, the ratio between singular and plural forms tends to vary, depending on the acceptance of ideas, understanding of the audience and, of course, nature of the text. Scientific texts introducing an idea to an audience, or describing a specific object or action will tend to make heavy use of the singular form - 'X has the following properties...'; 'X has been discovered in', where X could be replaced by, for example, 'colon cancer' or 'optical transition'. Where objects are more 'accepted' by the audience - that is they may have been explored in greater depth and so become classes of object - plurals predominate.

For consideration of pluralisation in English, we can derive some simple rules that may apply with reasonable generality, for example, a word or phrase occurring both with and without the "s" ending. Certain exceptions will hold for specific pairs of words (a simple example being "these" and "theses"), which stop lists act to reduce. Other variants ("ies" to "y") indicate substitution rather than contractions ("s" to "" ) to generalise this. Lexically, there may be no plural, although ontologically this is not problematic ("planets" but not "Jupiters"). Pluralisation, therefore, provides additional evidence for the existence of a concept, but does not exclude non-pluralised, or indeed non-singular, forms from providing such evidence.

As we have done to provide Weirdness values, we could consider comparison with ratios of singulars to plurals for specialist corpora against general language. This poses two problems:
firstly, the high-weirdness criteria in our approach immediately rules out all "infinite" weird words from the comparison (recall that for the nanoscale science and design corpus this is about 38% of the words). Secondly, those not immediately ruled out are occurring at low frequencies within the general language anyway. The remaining meaningful comparison, therefore, is the ratio of singulars to plurals. What does this actually tell us? We may infer that significantly higher rates of plurals convey ideas better accepted, that higher rates of singulars imply that acceptance is lower, and perhaps that we can converge the frequency values for these to provide a higher confidence in concept formation. The implications of such convergence, especially where subtypes, discovered through collocations, may not evidence both the singular and plural forms, requires further investigation.

Rather than deriving lists of rules for pluralisation, we consider an approach that looks for similarities between term candidates. It should be possible to create groups of terms with related meanings – “term clusters” (Lewis and Croft 1990). This can be achieved using the following steps:

- For each word, expand it to its set of (unique) n-grams (in our case, patterns of three letters).
- Compute a word-word similarity matrix, using, for example, Dice’s coefficient (Salton 1983) to measure the number of matching n-grams.

Patterns made up from each candidate term are compared using a selection strategy where the match is 80% or more (Srinivasan and Ruiz 1998).

\[
Trigrams = \frac{2x(n_{match})}{n_{term1} + n_{term2}}
\]

Where: \(n_{match} = \) number of matching trigrams

\[n_{term1} = \text{number of trigrams in term 1}\]

\[n_{term2} = \text{number of trigrams in term 2}\]

If we were to take patterns for, say, *tunneling diode* and *tunnelling diode*, we get the following match:

<table>
<thead>
<tr>
<th>Tun un nce nel eli lin ing dio iod ode</th>
<th>Tun un nce nel ell li lin ing dio iod ode</th>
<th>2*9/10+11</th>
<th>85.7%</th>
</tr>
</thead>
</table>

112
In the database community, this type of processing is referred to as q-grams (Gravano et al 2001). It is used for determining the edit distance between two strings, as could be determined using Levenshtein distance (named after Vladimir Levenshtein), Hamming distance (named after Richard Hamming, but only defined for strings of equal lengths) and other metrics. Two strings with a small edit distance between them should share a number of q-grams. An example for positional q-grams of length q=3 is:

<table>
<thead>
<tr>
<th>String</th>
<th>q-grams</th>
<th>Edit distance (Levenshtein)</th>
<th>Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>john_smith</td>
<td>##j, #jo, joh, ohn, hn, n_s, _sm, smi, mit, ith, th%, h%%</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>john_a_smith</td>
<td>##j, #jo, joh, ohn, hn, n_a, _a, a_s, _sm, smi, mit, ith, th%, h%%</td>
<td>(insertion of _, a)</td>
<td>11</td>
</tr>
</tbody>
</table>

If we ignore the position information, there are 11 common q-grams, but positionally, only 5, with the remaining 6 offset by 2 positions. For the tunnel(l)ing diode, the first 4 are positionally identical, with the remaining 5 offset by 1 position. This combination of grams, edit distance and positional information provides a number of means by which word, and indeed term, variance could be calculated.

### 3.4 Modelling knowledge resources

#### 3.4.1 Constructing terminologies

We have discussed key identifiers used in terminology resources and how to extract terms from text. Much is said about how people within and across enterprises should share each other's knowledge, or should at least be aware of what is known. The raw knowledge in a terminology database is no exception. The use of an existing terminology database can expedite translation, documentation and teaching and learning; indeed, sharing this raw knowledge may expedite the application of knowledge itself.

We now consider how terminologies can be built and how can they be used, or re-used, for other purposes. Every terminology data model attempts, to some extent, to provide a means to store as much information as deemed necessary in relation to the term; programs based on these data models may also have some form of import or export routine. The variety of export routines between programs indicates the need either for a mapping between the results of these mappings, or for the adoption of a common export form. While the latter is, perhaps, a longer term objective likely to be hindered by technology providers keen on maintaining a competitive advantage, the former presents a more realisable goal. Descriptive analysis of a terminology data model can
show how closely it conforms to a specific set of (metadata) descriptors, such as ISO 12620, and also to some common structure.

The development of a terminology data model, and the interchange between terminology formats and data models would therefore seem to be contingent on a very similar process of identifying the structural nature of the collection (existing or to be created) and determining the Data Categories contained or to be contained. The remainder of this section looks at terminology interchange formats that have been developed, and the principles involved as a means to understand the process of producing terminology resources – the terminology production line.

To create a terminology data model, as any other data model, it is first necessary to consider what is to be described. For this purpose, ISO 12620 acts as a useful reference source and we can make decisions about whether we need to deviate from it. The second step is to determine the structure of the data. This includes the determination of arity between the various selected data categories chosen: for example, should a concept have a single definition, a definition per language group, a single or infinite number of contexts, does a term need a grammatical gender? Each of these decisions influences the resulting terminology collection that is developed once the model is implemented and populated. The choice of modelling tool for making this determination will also affect, to some extent, the model – an entity-relationship diagram is more suited to the implementation of a relational database than an object-oriented model.

Making changes to a constructed data model has a number of effects. For example, the addition of a single data category within a relational database requires at minimum an additional field in a data table, and subsequent alterations to the user interface. At worst, the structure of the entire database may be compromised and require a new model for its accommodation – in which case, the existing entries all have to be analysed with respect to this new model; a new data category will contain no data for existing terms. It may be necessary to freeze the modification of the terminology collection for a period in which an updated version can be released – where millions of entries are contained, modification of the data model can require significant work to make existing data conform to this model. The migration between database models can be assisted by the terminology interchange formats, increasingly based on XML, which can be restructured in batch fashion using XML applications such as XSLT. Tools for manipulating XML structures can provide a more general framework for the update of a data model since there is no need to support data model update in every possible database format.

It can be shown that degrees of commonality exist between terminology collections by considering the values acceptable for these data categories in different collections. Here, we can use International Standards as reference systems for some of these DCs, as exemplified in Section
3.2, to ensure greater degrees of interoperability and interchange. Commonality can be determined as presented in Figure 28 below for two Terminology Databases (TDBj and TDBk) that share a set of Data Categories (DCi) from a common database (DCiDB) of values, for example values for language from the (currently expanding) series of ISO639 standards.

![Figure 28: Compatibility of terminology collections and the specific example of the Language field](image)

Having considered commonality in types of content, commonality of form becomes important. Although data values may be shared for data categories, the way they are structured may differ. It is important to consider the possible data modelling variance between terminology collections. For this purpose, an international standard that describes a metamodel for terminology was created, ISO 16642, which presents a "Terminological markup framework" (TMF). ISO 16642 introduces an abstract structure that can be used to represent a variety of terminological collections. It provides for the production of Terminology Markup Languages (TMLs) and can be used to represent terminology collections that can be interchanged through formats that have some pedigree, such as MARTIF (ISO 12200). TMF is based on a so-called meta-model (see Figure 29) consisting of a variety of containers within which data categories can be placed. These containers are referred to as structural nodes to which data categories are said to be anchored.
It is possible to create numerous data models for terminology, depending on the particular modelling requirements of cardinality alone, and a variety of initiatives, including EU projects TWB, TRANSTERM, INTERVAL and SALT, have produced slightly differing interchange formats for terminology collections so as to abstract away from the concrete database format. The labels for the values they represent are important, for example, synonyms may be identified by syn or synonym, or may exist as a term at the same level in a concept entry. Such data modelling variance needs to be understood. A number of data models for terminology have been developed in which the identifiers and modelling variance may, at first, produce significant variation within the data. At an abstract level, however, these models contain terms (that may relate to concepts), definitions, identifiers for particular languages and so on. The abstraction enables the potential for an interchange language that can make various resources interoperable. Here, we are not concerned with the details or limitations of specific database models. The following sections discuss some of the principal terminology interchange formats.

**Machine-readable Terminology Interchange Format (MARTIF)**

Existing as ISO 12200, MARTIF is reputedly the most widely used interchange format, albeit a highly structured and detailed format. MARTIF developed from the Text Encoding Initiative (TEI) terminology working group, from which the SGML-based TIF was created. The purpose of TIF was to allow users of terminology databases, regardless of the software, hardware and implementation methodology, to exchange terminology data. TIF was developed due to the increasing use of (diverse) terminology management software, changing attitudes towards reuse of terminology collections, and a need for third-party use of terminology (by subcontractors for example). TIF contained semantics that were not expressed in the SGML: for example, in the SGML nesting, any information that appeared in a tig (term information group) before the term
Chapter 3

element applied to the whole tig. MARTIF introduced the notion of the ntig (nested tig) to reduce the need for this interpretation and improve the interchange process (see, for example, Aaron et al 1998).

The interpretation of the resultant interchange document, combined with the complexity of developing MARTIF-conformant software has, in some cases, been a hindrance to its use. Part of this complexity can in part be blamed on its implementation using SGML. As the SALT project (in which the author was heavily involved) showed, MARTIF could be made fully conformant to TMF with minor modifications. The result of this conformity, and the need for computer-oriented terminology collections led to the one of the principal results of the SALT project, the extensible Language for Terminology (XLT), which is better known in the industrial community as the TermBase Exchange format (TBX). This conformity was carried out by other partners of the project, and is described in SALT project deliverables.

Interval Interchange Format (IIF)
The learning-curve necessary for the implementation of MARTIF led (prior to the SALT project) to the development of IIF within the EU project INTERVAL project (Gillam and Salway 1997). INTERVAL was concerned primarily with the consolidation (merging) and validation of terminology and lexicographical collections. As such, a simple format was required in which a maximum quantity of information could be exchanged by several project partners with a minimum quantity of effort. At the time of this project, the Web was just coming into its own, and web collections of terminology data were being made available, generally consisting of a number of terms each with a few other characteristics. The INTERVAL project was concerned with terminology validation, focussing on quantitative and qualitative analysis. The goal was to create a minimal DTD that would allow for tasks such as determining the correctness and currency of term use to be undertaken in a single format, with validation information being passed back to the linguists/terminologists for action in their default terminology management systems. 
IIF was used to capture Web-based collections of terminology, as well as collections that could be exported from partners, and to validate them against collections of Web-based texts, and also to reuse these collections in other applications.

Developing software applications around IIF removed the need for consideration of the underlying data models of terminology collections. IIF was a one-page SGML specification that enabled this interchange and reuse. IIF included "wildcard" tags that could be used to store application-specific data beyond that needed for IIF (which within TMF could be achieved using the "annot" mechanism); IIF was very much developed to facilitate activities in a project as rapidly as possible, although in the knowledge that it was, to some extent, compatible with
MARTIF. Within the INTERVAL project, IIF was used as the basis for terminology interchange between 4 partners, and as the input format for terminology validating and consolidating tools. Within the SALT project, the author showed that the IIF, with slight modification, could be made TMF conformant, and data from the INTERVAL project was quickly provided for use in further such studies. Each of the identifiers in the above table was contrasted with the Data Categories of ISO 12620 and the structure of the metamodel from ISO 16642. The analytical investigation of the IIF, and its relation to the structural model as defined in ISO16642 and the data categories in ISO12620 allowed IIF to become a TMF-conformant terminology markup language (TML). As such, it was possible to show that IIF and MARTIF were, finally, compatible (SALT project deliverable D4.15, Gillam 2001). Applying this approach to other terminology collections shows the work needed to make these collections reusable within this framework. TMF acts as a meta-interchange language to assist in the reusability of these collections.

3.4.2 Constructing “ontologies” (for knowledge bases)

Various text-based formalisms have been described in the literature for the interchange of systems of concepts. Recent focus has been on XML-based systems. These include Simple HTML Ontology Exchange (SHOE), Ontology Inference Layer (OIL), Web Ontology Language (OWL) and the DARPA Agent Modelling Language + OIL (DAML+OIL, McGuinness et al 2002). Some of these languages make use of the Resource Description Framework (RDF) and its schema (RDFS). Differences and similarities between such languages have been discussed (e.g. Gomez-Perez and Corcho 2002): the 6 ontology languages discussed all supported subclassing for producing taxonomies and had varying degrees of support for other relations, functions, axioms and so on. OIL and DAML+OIL appeared, from this study, to support the majority of expressiveness required for an ontology interchange language. The authors further define lightweight and heavyweight ontologies: lightweight ontologies contain taxonomically organised concepts, relations, functions and possibly instances, and heavyweight ontologies contain this and also axioms.

An ontology interchange language therefore provides the basis for description of an ontology. Application of and reasoning over an ontology, and the development of knowledge-based systems requires computer programs that can provide this reasoning: checking for integrity of the ontology, ensuring consistency in description, identifying issues of inheritance and so on. Software applications that provide the ability to import/export the formats mentioned previously, and to allow editing of the ontology, are maturing and widely available. These include OntoPrise’s OntoEdit system, Manchester University’s OilEd, and Stanford University’s Protégé, which is arguably the most developed of these three. Protégé has interfaces to the Java Expert
System Shell (JESS) to enable the direct development of Expert Systems using a lightweight or heavyweight ontology.

Figure 30: User interface of the Protege system showing an example ontology for newspapers

Work on constructing ontologies, especially that of Alexander Maedche, suggests the potential for automatically deriving ontologies from text, referred to by some as “ontology learning” (Maedche 2002, Aussenac-Gilles, Biebow and Szulman 2000, Mikheev and Finch 1995, Gómez-Pérez and Manzano-Macho 2003); perhaps on this basis we can posit terminology learning? These systems typically use existing language resources for this task: these include the use of part-of-speech taggers (ASiUM, Corporum, LTG Text Processing Workbench, Mo’K Workbench, SVETLAN’), phrase patterns (Prométhée, Caméléon, SOAT), Wordnet (Welkin), and user-built ontologies (DOE). Some systems use measures such TF/IDF (KEA, Text-To-Onto) for suggesting terms - though information retrieval specialists use term and word interchangeably - to the user. A further survey of “learning” ontologies from free text includes references to work in clustering, inductive logic programming, association rules, frequency-based, pattern-matching and classification, although in the majority of cases syntactic parsing is common (Maedche and Staab 2001a). The burden of constructing the ontology from the results of these operations rests with the user, although it appears that these ontology learners are, in fact, constructing terminologies!

Maedche has been involved with the production of the Text-to-Onto system (Maedche and Volz 2001), now part of the KAON workbench (Maedche and Staab 2003). This workbench takes a
corpus-based approach to ontology construction. A text corpus is analysed for word frequency, TF/IDF (although in testing KAON the implementation appears slightly suspect), Entropy, and a "C-Value", although documentation for the latter has not been found. There is also a means to filter results based on part-of-speech. As constraints to the system, the user is meant to set a frequency threshold and a maximum length for an extracted term. When results are presented, there is some clustering of words based on basic stemming, and the frequencies of these stemmed items are combined. Extracted terms can be added manually to an ontology model (called an OL-model), which necessitates the (human) construction of the tree-structure for the terms. Once added to this model, some relationships between these terms can be extracted. Certain relations are coded into the system including "such-as" and "as-well-as", although the criteria for acceptance are not apparent.

![Figure 31: User interface of the KAON workbench showing various items of information extracted about "terms" that have been coded into a hierarchy, and have some relations extracted.](image)

The Acquisition of semantic Knowledge using Machine Learning Methods (ASIUM) system reportedly learns semantic knowledge (Faure and Poibeau 2000, Faure and Nédellec 1999). POS tagging is carried out using an application called SYLEX. A POS tagger is also reported in the context of Edinburgh's Knowledge Acquisition Workbench (KAW) (Mikheev and Finch, 1995). POS tagging seeds a collocation process, which looks for noun phrases and verb phrases. For determining significant collocations, a threshold is "calculated using Zipf-Mandelbrot law". The authors note the significance of what they call "term inclusion": "Many terms include other terms as their components. This surface lexical structure corresponds to semantic relations between
concepts represented by these terms”. This certainly provides an interesting parallel to the work presented here, and to make a further parallel, the approach reportedly uses a term bank to identify term inclusion, though details of the term bank are not apparent.

The OntoLearn application claims to perform terminology extraction in a specific domain, and then applies Wordnet to the discovery of semantic relations between terms (Navigli et al 2003). OntoLearn uses some form of comparative analysis across different domains and some relevance of a term to a given domain is computed. OntoLearn uses a “domain consensus” measure to ensure use over a number of documents, which looks uncannily like TF-IDF. Wordnet is then used against the senses of each component of the term to find the appropriate sense. Having carried out a process involving the learning of specified relationships, the resulting ‘ontology’ is used to seed the translation of multiword terms (to the Italian version of EuroWordnet). OntoLearn’s champions suggest that the approach works well for component words that translate appropriately, although they do not treat the case where no correspondences exist, which they say is most frequent.

These so-called ontologies may share common items; descriptions of concepts, hierarchical arrangement, axioms and so on. This would require comparisons between the ontologies at the structural and conceptual levels. By separating the concepts from their relations – especially hierarchical relations – it is possible to consider, firstly whether two systems share the same concepts, and secondly the degree to which the hierarchical relations are similar. It may be that the missing concepts force alterations to the hierarchical relations, or that the concepts themselves, although appearing to have surface similarity do indeed refer to different things. This would seem to parallel the harmonization of terminology collections as detailed in ISO 860. Having relegated some aspects of ontology population to terminology population, we can argue here that such a harmonization should first occur at the terminological level, such that it can be wrapped into an ontology at a later stage. We could use methods of Formal Concept Analysis as proposed by some authors (Sowa 2000, Stumme and Maedche 2001), or we could apply mechanisms for ontology alignment such as that proposed by Maedche and Staab who present a means by which to determine the potential similarity between two separately constructed ontologies (Maedche and Staab 2002, 2001b). They define a string matching method for determining the overlap between two lexicons, which immediately commits the operation to overlaps between lexicons in the same language, or perhaps to translations of those lexicons to a common language. They further consider hierarchical structures to which these strings are matched, and the relationships between the concepts. Their method begins by considering string matching using Levenshtein’s edit distance and averaging over the similarity rather than the distance. This average provides a measure, which as they acknowledge may be deceptive due to
similarities of semantically detached words such as *power* and *tower*, for similarity of lexicons—they are assuming that the same string, or a highly similar one, will have been used to represent the same concept. The similarities are used to propose the lexical intersection between the two ontologies being considered. The union of these lexicons produces the mapping to each concept within the ontologies (although it is possible to consider that there may be concepts within the ontologies which do not have lexical entries, since there is no constraint on this). At the intersection of the lexicons, we have the means by which to consider conceptual similarity. For this, Maedche and Staab propose a measure of taxonomic overlap (TO), a kind of *coefficient of interoperability* between ontologies.

### 3.5 Conclusion

In this chapter, we have considered how to populate terminologies such that they may be used within ontology systems and ontological inferences made over them. We reviewed terminology collections, their constituents, standards that they should conform to by which they can be more widely utilised, and the means by which it is possible to get towards such collections using primary resources (text collections). To reduce the (man-power) resources required for creating such collections, we considered various techniques for extracting specific words and phrases from collections of text, focused around statistical, rather than linguistic, mechanisms. We considered Zipfian distributions of words in texts and how, although not following Zipf's expectations, distributions differ in a similar manner from this. Specifically, within collections of specialist texts, there appears to be a *controlled* use of language, which may be measurable in various ways. As attempts to measure this specialist language, we explored Quirk's notion of frequency correlating with acceptability and looked at frequency profiles for a number of specialist text collections. We also acknowledged the Malinowskian notion of the *coefficient of weirdness*, and investigated how an initial list of candidate terms (or carrier terms) might be produced from a combination of these two notions, which may provide a systematic treatment for LSP. Subsequently, we considered how this list might be used to seed examinations of collocations. Having generated a tree-structure of candidate terms, perhaps displaying "semantic inclusion", we noted how this could be used as the basis for a terminology collection, and subsequently for an ontological representation of the domain. By considering how to automatically extract terms from collections of text, with reference to a standards-conformant way of representing them, we bridge a divide between prescriptive and descriptive approaches to terminology. Prescriptively, we decide what we should have with respect to structure; descriptively, we determine what actually exists in text, and how this can be used to populate the system.
While text analysis is a central issue in developing resources for a variety of purposes including terminological and ontological ones, available literature only presents limited comparisons between a few existing techniques applied to one or two collections of specialist text. Other studies apply a few mechanisms to highly general text, and ageing, collections such as the Wall Street Journal, AP newswire, and the British and American National Corpora (BNC, ANC). Lexicographical studies tend to consider one word, or a few words, at relatively low frequencies within such collections. As a result of these limited explorations, it is difficult for researchers to intelligently choose appropriate mechanisms for the analysis of specialist texts, or to appreciate the extent to which specific mechanisms are, or remain, an effective proposition. While a full statistical appreciation of the mechanisms is beyond the scope of the current work, future work can be considered for producing such an analysis. In treating multiple corpora from different specialisms, as opposed to different varieties of general language, we would hope to further generalise the approach such that it would work mechanically with an arbitrary collection of (currently English) specialist text from a specific subject field.

A specific problem has been that statistical methods produce limited evidence for valid term formation: as terms increase in length (measured by the number of contained words), the frequency (likelihood) of occurrence tends to decrease. Here we have some difficulty since, in following the work of Quirk et al whereby frequency correlates with acceptability, longer terms exist with reducing frequencies. The likelihood of terms of increased length should be greater than chance, and we require some means by which to discern term patterns against simple collocations (e.g. bread and butter). The combination of term inclusion, frequency of occurrence, a function of rarity of words in contrast to so-called general language (weirdness), exclusion of closed-class words, and other suitable metrics suggests a method for synthesising these approaches, assessing the value provided by the metrics, and using them to generate terms of arbitrary length from arbitrary collections of text. It may be worth considering, in future work, whether collocation span sampling can help to systematize the heuristic for selection of interesting collocates. Although the systematic treatment, including the statistical justification, of each of these aspects is beyond the scope of the present thesis, empirical results and their evaluation should suggest some validity to the approach and, particularly, to its mechanisation by which terms can be extracted with some degree of confidence, placed into a representation, and subsequently used elsewhere.

A question we will not be able to answer is whether a given word is domain-specific. For example, where an idea is explained using an analogy, does the analogous item now constitute a domain-specific term? How is the belongingness of a word to a given domain measured? Is it a definite yes/no answer, or is it some fuzzy set? There are specific words which constitute closed-
class words for English, and which would not be expected to be used within a term. The same
cannot, however, be stated for French – consider “pen name” versus “nom de plume”: is the “de”
terminological here, but not elsewhere? A systematic approach to the domain-specificity of
words would be beneficial, however human judgement is currently needed for such a decision
since, for each word, this requires some background domain knowledge. The “Catch 22” of this
statement, and as noted for using a dictionary in Chapter 2, is that in order to know the language
(of the domain or dictionary), you have to know the language (of the domain/dictionary)! This is
a difficult specification for a computer program. At the same time, the use of linguistic patterns
also requires knowledge about language structure in general, and specific structures as may occur
within specific domains (or sublanguages). Again, such prior knowledge is not easily
programmed. We could consider using the most frequent words of a general language corpus to
provide a useful exclusion list, since we are not really interested in the general language. Here the
question of a cut-off point for the list becomes important: BNC contains the following words
(ranks according to frequency in brackets): chemical (2376), carbon (3866), chemicals (4236),
carbonate (12291), biochemical (13637), carboniferous (13721), bicarbonate (19853), chemically
(19849). For certain domains, such words may be important indexicals. For others they may
participate in key multiword terms. A list of the 2000 most frequent words of BNC would seem
to be a useful heuristic / compromise while attempts can be made elsewhere to automate the
provision of such a list.

Other mechanisms for collocation analysis suggested in the literature that we did not consider are
the dice coefficient and Pearson’s Chi-square and Log-likelihood ratios. There is, however, only
so much that can be done to analyse frequency values for 2 co-occurring strings within a given
distance of each other. The reported accuracy of Smadja’s method for terminology extraction
leads us to suggest using his method for knowledge extraction, based as it is on positional
information which the majority of other mechanisms ignore, will be most appropriate.

In what some may think a serious omission, we have not considered part-of-speech taggers and
syntactic patterns that may be extracted using these. POS tagging requires a training corpus and
some form of correction. Its performance can therefore only be as effective as the representativity
of the training corpus. Anecdotal evidence suggests that POS taggers that are trained on general
language corpora become significantly less accurate when used on specialist corpora, and here
background domain knowledge would be valuable. The various mechanisms investigated in this
thesis operate at the lexical level using various statistics to suggest words worth considering
further: this may form a useful precursor to POS tagger training, and indeed POS information may
be of use for validating the list of candidate terms. Should we subsequently be able to seed
linguistic analysis from this, we will have provided an interesting and valuable contribution to the treatment of special language, at least as far as English is concerned.

In the next chapter, we present a demonstrator for populating terminologies from arbitrary collections of (specialist) text, and using the terms within ontology systems such that ontological inferences made over them. The combination of terminology identifiers (section 3.2), term extraction mechanisms (3.3) and knowledge resource models (3.4) all support this effort.
4 System Description

The purpose of this chapter is to outline a method that could be used for populating a term base by describing various aspects of a terminology extraction system. The resulting term base would be ready for use by translators/terminologists and documentation professionals. A novel use of such a term base is as a basis of a Knowledge Base (KB). Section 4.5 concludes this chapter with comments on the work reported in the previous three chapters.

Firstly, we re-iterate a key message from Chapter 3: terms, or rather candidate terms, can be extracted from texts, their possible interrelationships identified, and where possible data for a term’s attributes (definition, context and so on) may be obtained. The operative word in candidate term is candidate. The candidate may or may not be approved by (one or more) terminologists, or equally importantly by the members - experts - of a specialist domain community. A systematic study of terminology can assist in providing an outline terminology collection through the production of a collection of these candidate terms and their composition into a standards-conformant interchange format. The organisation of this collection, with respect to conceptual relationships, could provide for a candidate ontology, or at least a candidate concept-oriented terminology.

Developments in International Standards for terminology and particularly the emergence of the terminological metamodel (Chapter 3), suggest that it should be easier to produce data for a term data base using a terminology markup language (TML) in conformity with the standard. The emergence of ontology-oriented markup languages as identified in section 2.2.2, supported by other applications of XML, such as ChemML, TimeML and so on, implies the potential for marking up data to be imported into a system that understands the (semantics of the) markup language: such a markup language can be referred to as an interlingua. Processing the interlingua is a less fraught effort than developing direct interfaces to individual data-based systems. If systems can import and export this interlingua, the information (or knowledge) can be shared and re-used.

Populating a terminological resource is, then, a process of identifying terms and exporting the information collected about them to the interlingua. Generating a “complete” terminology collection from a document collection becomes a more realistic, though ambitious, proposition. It may, in fact, be easier to consider the bootstrapping of such resources direct from text than to iteratively refine an existing resource: refining an existing resource would require treatment of the
existing relationships, which may require additional efforts. Advances in the expressivity of markup languages, through the description of data structures able to express relationships between "objects", such as subclassing and the object-attribute-value triples of RDF, indicate that a computer system may be able to make inferences over the term base, as opposed to simply displaying or exporting the data. The initial work required to produce terms for such a resource would be a key contribution.

Use of markup languages has a long tradition in terminology dissemination. Development of languages that can be used to delineate data relating to various attributes of a term leads us to believe that use of a terminology markup language that conforms to ISO standards can help in building a knowledge base (KB). A KB relies explicitly on the inferential power of algorithms to reason over its contents; the knowledge base reasons over the vocabulary of a domain that has been appropriately structured, for example in frames or semantic networks. Such an appropriately structured vocabulary could, perhaps, be provided by this marked-up terminology.

In this chapter, we describe techniques to assist the process of populating such a terminology collection, and subsequent use of such a collection within an ontology system for intelligent systems development (as a KB). The techniques and methods, developed with various colleagues, have been implemented by the author and evaluated by yet other colleagues. These techniques and methods have been applied to collections of language of various domains, validated through publication in peer-reviewed journal and conference publications, and used as the basis for other work, including at least seven other PhD studies.

The initial focus of our method is the automatic extraction from arbitrary collections of text, of domain terminology. For a terminology collection, the necessary output is a collection of terms, perhaps with associated definitions, contexts, source information and so on. We assume the existence of software capable of interpreting output from such a method that is delimited using a markup format conformant to TMF (described in Chapter 3).

The motivation we have is: how do we formalize a text-based empirical method for automatically extracting terms that represent concepts, relationships between the terms and also relationships between the concepts such that they can form the potential basis for both a terminology and an ontology? The formalization of the method, applied to the examination of arbitrarily selected domain texts will help in discovering the terminology/ontology of the domain. We wish, eventually, to be able to explore the notion of "ontology learning" from free text by consideration of terms, concepts and relations, international standards for terminology and the population of terminology systems.
Consider one of the classic methods of acquisition of knowledge due to Buchanan et al (in Waterman et al 1983, pp127-167) and specifically the mechanism for populating a Knowledge Base (KB) from text using a text understanding program. We seek to outline a method that could be used to populate such a KB, by taking a raw collection of texts in an arbitrary (specialist) domain and presenting the resulting ‘understanding’ of the text in such a way that aspects of the domain knowledge can be (automatically) modelled from this understanding. The recent focus on ontology would result in an abstraction away from a specific KB: the KB now becomes an instance of use of the ontology. We can consider this as a refinement of Buchanan’s approach (see Figure 32 below: the original process is shown with a dotted arrow while the new refinement is shown by the two heavier arrows).

![Figure 32: Refinement of Buchanan's approach to converting from text to knowledge base via text understanding program, adapted from Waterman et al 1983, p132](image)

It will be demonstrated that the automatic seeding of such a KB can be carried out based on human or machine selection of a few parameters, although the final output remains subject to, and contingent on, human (expert) judgment and acceptance. Since the method is based on the extraction of information from a collection of texts in a specialist domain, and requires a baseline comparison to general language, we assume the existence of a specialist corpus and a general language corpus. For the latter, we use the British National Corpus (BNC). The former can be captured from a variety of sources, for example existing text collections, collections harvested from suitable web searches, collections from domain-specific journals, and so on.

The advent of the Web, and increasing use of terminology markup, has made a number of terminological and lexicographical resources available for public use. It may be possible to verify extracted terms against such resources. The data of the candidate terms, encoded with reference to ISO standards, may then be presented to a human expert/terminologist for approval/validation.
4.1 Method Description

A collection of texts comprises instances (tokens) of words (types), which can be indexed by positional information, as well as information such as grammatical marks. For our purposes, we initially only consider the words (types) of the corpus, so for a particularly defined corpus \( W = \{ w_1, w_2, \ldots, w_n \} \) is a set, the length of which is the number of types in the corpus. For general use, we define cardinality and rank as follows:

**Definition 1 (Cardinality):** For a (finite) set \( S \), \( \#S \) denotes the cardinality (number of elements) of that set —

\[ \# : S \rightarrow \mathbb{N} \]

The cardinality of a set of length \( n \) is \( n \).

**Definition 2 (Rank):** A rank on a set of words is an injective function concerned with the ordinality of \( w \) —

\[ \text{rank} : W \rightarrow [1..\#W] \]

The rank of the last item in a set of length \( n \), howsoever that set has been ranked, is \( n \).

4.1.1 Extracting individual words

Following tokenization, counting the number of tokens of a type in the corpus provides the frequency of the type in that corpus, and the relative frequency is given by dividing by the total number of tokens:

**Definition 3 (Frequency):** Frequency values of words, \( W \), in a text corpus are given by

\[ f : W \rightarrow \mathbb{N}. \]

The size of a corpus, in tokens, is

\[ \sum_{w \in W} f(w) = C \]

**Corollary (Frequency Ranking):** If considering rank based on frequency, rank respects frequency in that \( f(w_1) > f(w_2) \Rightarrow \text{rank}(w_1) < \text{rank}(w_2) \). If \( f(w_1) = f(w_2) \), rank of \( w_1 \) and \( w_2 \) is arbitrary, since if \( \text{rank} : W \rightarrow [1..\#W] \) respects the above definition, then if
Chapter 4

f(w₁) = f(w₂) and w₃ is such that rank(w₁) < rank(w₃) < rank(w₂), then f(w₃) = f(w₁). We can use this to rank w ∈ W_{GL}

Equation 1 (Relative Frequency): Relative frequencies are given by

\[ r(w) = \frac{f(w)}{C} \]

from which it follows that

\[ \sum_{w \in W} \frac{f(w)}{C} = 1 \]

This has been referred to elsewhere as word probability, and more recently as concept probability (Yuhua et al 2003)

For a specialist corpus SL, definitions 3 and equation 1 produce \( f_{SL}(w), r_{SL}(w) \) for \( w \in W_{SL} \) and a value for \( C_{SL} \).

For a general language corpus such as the British National Corpus, we can establish similar values \( f_{GL}(w), r_{GL}(w) \) for \( w \in W_{GL} \) and a value for \( C_{GL} \). We use these values to produce the values for smoothed weirdness (equation 2), and determine z-scores (equation 3) for frequency and weirdness.

Equation 2 (smoothed weirdness): A value of weirdness that includes add-one smoothing (see section 3.3.1) is given by

\[ \tau(w) = \frac{N_{GL} f_{SL}}{(1 + f_{GL}) N_{SL}} \]

By consideration of the relationships between \( W_{SL} \) and \( W_{GL} \) and their values of weirdness, it should be possible to determine whether the corpora are identical or the extent to which they overlap.

Equation 3 (z-score): The z-score of a value \( v_i \) in a set of values \( V = \{v_1, ..., v_N\} \) is given by

130
By equation 3, we compute z-score values on $r_{SL}(w)$ (equation 1) and $\tau(w)$ (equation 2) for $w \in W_{SL}$ as suggested in section 3.3.1. With appropriate z-score values, we provide a set of words for further analysis.

**Example**

A general language corpus, the British National Corpus (BNC) contains 100,106,029 tokens consisting of 669417 types. Our special language corpus about Nanoscale Science and Design comprises 1,012,096 tokens which consists of 26861 types. If we consider two words within the specialist texts (SL), *nanotube* and *carbon*:

- *carbon* appears both in SL and GL corpora; frequency in SL is 1558 (definition 3); frequency in GL is 2463 (add-one); weirdness of *carbon* is $\tau(\text{"carbon"}) = 62.54$ (equation 2).
- *nanotube* appears ONLY in the specialist corpus; frequency is 969 (definition 3); frequency in GL is 0 (add-one); $\tau(\text{"nanotube"}) = 95843$ (equation 2).
- With $z=1$, 19 words (of the 26861) are selected ($#W' = 19$) (by equation 3).

### 4.1.2 Extracting multiword terms

We can use the words selected in the previous section as input set to Smadja's collocation method, at his step 1.2. We extend from Smadja's "a given word w" to a set of words automatically generated on the basis of a combination of high frequency and high weirdness (to provide larger or smaller input sets, we could change the z-score value). We adopt his use of U-score and z-score, but do not tag the corpus. Furthermore, we move from consideration of individual words to multiword strings. We denote a set of multiword strings as $M = \{m_1, m_2, \ldots, m_n\}$, where $m$ is composed of individual (sequentially occurring) words.

**Definition 4 (collocation distance):** Let $D$ be values between negative and positive integer value $d$, excluding zero: $D = [-d, +d] \setminus \{0\}$.

The collocation frequency $c$ is given by $c : W \times W \times D \rightarrow N$. $c(w_1, w_2, d)$ is the frequency with which $w_1, w_2$ occur precisely $|d|$ apart in the order $w_2, w_1$ if $d$ is positive, and $w_2, w_1$ if $d$ is negative.
Equation 4 (collocation strength): For $x \in D$ (definition 4), $c(w_i, w_j)(x)$ determines the list of collocations for given $w_i, w_j$. The collocation strength ($U$-score) of $w_i, w_j$ is:

$$U(w_i, w_j, d) = \frac{\sum_{x \in D} (f_c(w_i, w_j, x) - \bar{f}_c(w_i, w_j))^2}{\#D}$$

Definition 5 (relative collocation strength): Let the relative strength of the collocating word $w_j$, as according to Smadja, be given by a z-score (equation 3) on values of $f_j(w_j)$ given $w_i$.

Smadja cites $U(w_i, w_j) > 10$ and $z(f_j(w_j, w_i)) > 1$ as useful inequalities for selection of significant collocations.

A further $k$-score is used to determine significant collocation positions. For creation of compounds, or multiword expressions, in English, consideration of $d=1$ should suffice (we still consider $d=5$ for determining the value of the collocation in this instance). For languages such as French, and for determination of local grammars, this constraint can be relaxed. We do not make use of this value at present.

Definition 6 (collocating phrase): Two significantly collocating words are concatenated according to the distance between them. We define $\otimes : W \times W \rightarrow P(W^*)$

For example, possible collocating phrases for nanotube and carbon that may be generated (including the different phrase lengths) are:

$$w_j \otimes w_i = \text{“carbon” } \otimes \text{“nanotube”} = \{\text{carbon nanotube}, \text{carbon based nanotube}, \text{carbon based aligned nanotube}\}$$

$$w_i \otimes w_j = \text{“nanotube” } \otimes \text{“carbon”} = \{\text{nanotube carbon}, \text{nanotube of carbon}, \text{nanotube made of carbon}\}$$

Example

For nanotube and the neighbourhood $d=5$ (definition 4) we have collocations with carbon of:

"nanotube", "carbon" = [8, 5, 1, 2, 229, 0, 0, 2, 2, 3]
Similarly, for *multi-walled* we have:

> "nanotube", "multi-walled" = [.., 0, 11, 4, ...]

- U-score for *carbon* and *nanotube* is $U(\text{"carbon", "nanotube"}) = 4620.16$ ($> 10$) (equation 4)

- Given "nanotube", "carbon" produces $z(f(\text{"carbon"})) = 31.281$ ($> 1$) (definition 5)

For the pair *carbon* and *nanotube* with distance +1, we form the pattern *carbon nanotube* (definition 6).

If the distance had been −1, we would be considering the pattern *nanotube carbon*. If we considered *multi-walled* and *nanotube*, since the significant distance is −2, we would have to consider the pattern *multi-walled * nanotube*, where * is a wildcard. Since we constrain to $d=1$, we ignore such patterns.

**Recollocation**

While hinting at recollocation, Smadja does not describe it sufficiently for implementation purposes, indeed Smadja wishes to discover "rigid noun phrases or phrasal templates", for which he considers a probability of $> 0.75$ for each subsequent word, given a word pair.

We have extended this notion of collocation by looking at collocations to produce multiword expressions using the definitions and equations given above. This process is recursively applied so that starting from the collocation of two single words, we produce bigrams and consider collocations with these bigrams to produce trigrams, and so on until the statistical significance of the last $n$-word compound is below the required values. We have not systematically examined the explored the effects of reduction of the thresholds given by Smadja. We have, however, considered the effects of recollocating the first 5 results in each case, with interesting results.

For each collocating phrase (definition 6), we analyse the contexts within which these phrases occur, hence each set of contexts is likely to be a different input corpus. Retaining context information from the single word analysis means that each subsequent step analyses less text.

Again, unlike Smadja, we consider only $d = 1$ for deriving collocations, but are still interested in how the values for $U(w_i, w_j)(x)$ and $z(f_i)$ can be applied to determine more than can be found from his simple probability. We repeat this recollocation until $U$ and $z$ constraints can no longer be satisfied.
From the initial collocation, and subsequent recollocations, we form candidate trees that evidence "term inclusion" through left- and right-extension of the collocating phrase. The effect of this method may be to produce unbalanced trees of collocation patterns of increasing lengths.

Example

For carbon nanotube from the previous example, and the neighbourhood \( d = 5 \) (definition 4) we have collocations with arrays, of:

"carbon nanotube", "arrays" = [0, 0, 0, 0, 0, 14, 1, 0, 0, 0]

We would repeat this analysis for carbon nanotube arrays to give, for example,

"carbon nanotube arrays", "ordered" = [..., 5, .....]

which produces ordered carbon nanotube arrays, and so on, for conditions on \( U \) and \( z \).

4.1.3 Aggregation – Word variants

We consider here the computation of a word-word similarity matrix, using the Dice coefficient (Salton and McGill, 1983) to measure the number of matching n-grams. Since this method on a word-word basis would only calculate the amount of matching words, we expand each word to its constituent n-grams (in our case, we consider trigrams as patterns of three letters).

Equation 5 (dice coefficient): For each word or phrase we seek the set of all substrings of length \( n \) characters in each component word. We define gram: \( W \times N \rightarrow P(\text{char}^*) \)

For each \( w_i, w_j \) and length \( n \), the dice coefficient is:

\[
dice(w_i, w_j, n) = \frac{2 \#(\text{gram}(w_i, n) \cap \text{gram}(w_j, n))}{\#(\text{gram}(w_i, n)) + \#(\text{gram}(w_j, n))}
\]

The denominator could be determined by union of these substrings, although a slightly different result would be obtained. Comparisons between terms located through seeded collocation statistics, such as those presented, and those located by identification of lexical phrases can be compared using this technique to determine relationships between terms (Gillam and Ahmad 2002). Tri-gram patterns made up from each candidate term are compared using a selection strategy where the match is 80% or more (Srinivasan and Ruiz 1998). This value may be increased or decreased depending on the strength of match required.
Example

For phrases normal-incidence inas gaas quantum-dot detector and normal-incidence inas gaas quantum-dot detectors and length 3, we produce sets such as

\{nor, orm, rma, mal, al-, l-i, i-n, -in, inc, nci, cid, ide, den, enc, nce, ina, nas, gaa, aas, qua, ant, ntu, tum, um-, m-d, -do, dot, det, ete, tec, ect, cto, tor, or\}

with an additional element for the second phrase.

The intersection is equivalent to the first set, and the union contains many duplicates. The match (equation 5) is given by \((2 \times 33) / (33 + 34)\)

The principal concern about such a mechanism is the possibility of producing relations between unrelated words or expressions. An example of this would be the similarity between "powering" and "towering". For specialist languages, it is hoped that these will be avoided due to consideration of common collocations. It will only be possible to determine this experimentally. A subsequent concern is that increasing phrase lengths will result in collocations of different lengths being considered as synonymous, e.g.

**carbon nanotube arrays : ordered carbon nanotube arrays = 0.85**

Two modifications suggest themselves: increase the acceptance threshold; modify the calculation to account for the number of words contained. The latter may affect the treatment of e.g.

**single wall carbon nanotubes : singlewall carbon nanotubes**

Finally, we have to consider the effect of convergence based on synonymy, since this may also require the convergence of any subtrees as may exist within the structure. There is an ontological question here related to concept formation (expression using singular versus plural forms), although we will not be able to answer such a question in the scope of this thesis. Further evaluation of this technique is required.

4.2 Combined algorithm

The various components outlined above can be combined into a single algorithm for the production of a candidate conceptual structure from an arbitrary collection of texts:
### Chapter 4

#### 4.2.1 Modelling knowledge resources

According to Maedche, and as noted previously by other authors, Wordnet can be “wrapped” for use as an ontology (Maedche 2002, described in 2.3.2). It should be possible, then, to create a similar wrapper for a (concept oriented) terminology collection, and hence bridge the gap between terminology and ontology, using existing terminology collections to seed domain ontologies. We consider again the terminology standards ISO 12620 and ISO 16642 and, based on these, provide an ontology wrapper. An ontology wrapper for terminology collections defined in this way assists...
in alleviating the knowledge acquisition bottleneck by providing a direct route to domain ontologies from terminologies.

TMF does not define a hierarchy for terms, but the structure of a flat-file representation of (conceptually oriented) terms: the conceptual orientation is orthogonal to the tree-structured metamodel and defined using specific data categories from ISO 12620. TMF defines containers of terms based on language independent concepts (Term Entries), language dependent groups of terms (Language Sections containing Term Sections) and related groupings. TMF identifies the TermEntry (TE) level as being in correspondence with concepts, \( C \). We consider that terms (denoted by ISO 12620) within this concept, which within such collections are assumed to have a degree of synonymity, represent the lexicon mapped to a specific concept, \( L^C \). We have analysed the combination of ISO 12620 and ISO 16642, and provide the following table to show how an ontology wrapper, following Maedche, could be provided from this combination.

<table>
<thead>
<tr>
<th>Ontology ( O )</th>
<th>ISO 16642 / ISO 12620</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C )</td>
<td>TermEntry (ISO16642-TE)</td>
<td>A TermEntry corresponds to a concept ( C ). Words contained in the TermEntry are stored (in the lexicon ( L )) as terms and mapped to the specific concept ( C ).</td>
</tr>
<tr>
<td>( L^C )</td>
<td>term (ISO12620A-01)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( H^c )</td>
<td>broader concept generic ('is a') (ISO12620A-070201)</td>
<td>Hierarchically-motivated conceptual relationships within ISO 12620 directly map to ( H^c ).</td>
</tr>
<tr>
<td></td>
<td>superordinate concept (ISO12620A-07020201)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>superordinate concept generic (ISO12620A-07020202)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>subordinate concept (ISO12620A-070203)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>subordinate concept generic (ISO12620A-07020301)</td>
<td></td>
</tr>
<tr>
<td>( R )</td>
<td>broader concept partitive ('has a') (ISO12620A-070202)</td>
<td>&quot;Other&quot; concept relations.</td>
</tr>
<tr>
<td></td>
<td>superordinate concept partitive (ISO12620A-07020203)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>subordinate concept partitive (ISO12620A-07020302)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>concept relation (ISO12620A-06)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>generic relation (ISO12620A-0601)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>partitive relation (ISO12620A-0602)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>associative relation (ISO12620A-0604)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>coordinate concept (ISO12620A-070204)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>coordinate concept generic (ISO12620A-07020401)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>coordinate concept partitive (ISO12620A-07020402)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>related concept (ISO12620A-070205)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>related concept broader (ISO12620A-07020501)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>related concept narrower (ISO12620A-07020502)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sequentially related concept (ISO12620A-070206)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>temporally related concept (ISO12620A-070207)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spatially related concept (ISO12620A-070208)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>causally related concept (ISO12620A-070209)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>associated concept (ISO12620A-070210)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>antonym-concept (ISO12620A-10180602)</td>
<td></td>
</tr>
</tbody>
</table>

Table 23: Mapping between ISO 16642/12620 and the definition of an ontology structure for providing an ontology wrapper

137
Chapter 4

Consider the concept entries within a sample of 527 concepts (Terminological Entries) taken from the DHydro\textsuperscript{15} Hydrographic Dictionary, subsequently used in the SALT Project. Each concept is identified by an alphanumeric string, e.g. HR-3880 and contains a number of terms in English, French and Spanish. Relations such as subordinate concept generic (1 instance), related concept (121 instances) and antonym (31 instances) exist within this sample. An example “instantiated ontology structure” for a few entries within this collection can be seen in Figure 33 below.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{ontology_structure.png}
\caption{(Following Maedche). An instantiated ontology structure for a set of concepts from the DHydro Hydrographic Dictionary}
\end{figure}

According to Maedche’s notation, the ontology structure is given by: $\mathcal{C} = \{c_1, c_2, c_3\}$, $\mathcal{R} = \{r_4\}$, $\mathcal{H} = \{(c_2, c_1)\}$, $\mathcal{R}_{rel}(r_4) = (c_2, c_3)$ and $\mathcal{A} = \emptyset$

And the lexicon is: $\mathcal{L}_{en} = \text{\{“magnet”, “heeling magnet”\}}$, $\mathcal{L}_{fr} = \text{\{“aimant”, “aimant de correction de bande”\}}$, $\mathcal{L}_{es} = \text{\{“corrector magnético”, “imán de escora”\}}$, $\mathcal{L} = \mathcal{L}_{en} \cup \mathcal{L}_{fr} \cup \mathcal{L}_{es}$, and $\mathcal{R} = \text{\{“related concept”\}}$

For standardised terminology, we expect one preferred term per language, per concept, along with perhaps some accepted terms and some deprecated terms. For a specialist domain, we can accept one concept or relation being referred to by one or more lexical items, but not vice-versa. One lexical entry may, however, refer to several relations, although those relations would be of the

---

\textsuperscript{15} DHYDRO (http://www.loria.fr/projets/MLIS/DHYDRO/), a former project within the European MLIS (Multilingual Information Society) programme.
same kind. This potential many-to-many mapping requires further investigation elsewhere to
determine the most appropriate definition for this behaviour.

For the function $F$, we have the following mappings: $F(\text{"magnet"}) = c_2$, $F(\text{"heeling magnet"}) = c_2$, $F(\text{"aimant"}) = c_2$, $F(\text{"aimant de correction de bande"}) = c_3$, and the function $G$ mapping the single relationship: $G(\text{"related concept"}) = r_3$.

Maedche initially posits $F$ and $G$ as many-valued functions. For a specialist domain, we would prefer $F$ to be a function in the normal meaning of the word (i.e. not many valued), since we prefer one lexical entry to refer only to one concept. That is, given $F(t_1) = c_1$, we cannot have $F(t_1) = c_2$. However, the inverse relationship of one concept being referred to by many lexical entries is still desirable, that is, $F^{-1}(c_1) = t_1$, $F^{-1}(c_2) = t_2$, and so on. We still retain $G$ as many valued and look at “causes” as an example (“causes” represents the 12620 Data Category causally related concept). We want $G(\text{"causes"})$ to be both $r_1$ & $r_2$ where $rel(r_1) = (c_1, c_2)$, $rel(r_2) = (c_2, c_3)$, and, e.g. $F(\text{"smoking"}) = c_1$, $F(\text{"cancer"}) = c_2$, $F(\text{"death"}) = c_3$.

Within ISO 12620, conceptual relations relate concepts to other concepts at some level. There are further relationships, some of which are quite obscure such as “coordinate concept partitive”, defined as: “A subordinate concept having the same nearest superordinate concept and same criterion of subdivision as some other concept in a given partitive concept system”, for which the necessity is likely to be rare. Furthermore, there are relationships that are non-taxonomic, including “associative relation” (between two concepts).

Having provided such a structure for an ontology wrapper, it is possible to consider axioms in $\mathcal{A}_0$ described in first order logic that operate on the concepts and relations, for example, consider spatially-related-concept. We can define spatially-related-concept as being both symmetric and transitive:

$$\forall x, y . \text{spatially-related-concept}(x,y) = \Rightarrow \text{spatially-related-concept}(y,x)$$

$$\forall x, y, z . \text{spatially-related-concept}(x,y) \land \text{spatially-related-concept}(y,z) = \Rightarrow \text{spatially-related-concept}(x,z)$$

and so may be able to make use of various combinations of conceptual relations defined this way with other such relations.

The mapping provided above is sufficient for the definition of the ontology and lexicon. We can, however, consider the extension of this model to retain information with respect to relationships
between terms, which may be of importance in establishing relationships between parts of concepts, but which can be extended to lexicographical collections and term-oriented terminology collections. Consider, for example, the concept in DHydro identified by HR233, which contains the following information:

<table>
<thead>
<tr>
<th>Language</th>
<th>Full form</th>
<th>Abbrev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr</td>
<td>Association Internationale d'Hydrologie Scientifique</td>
<td>AIHS</td>
</tr>
<tr>
<td>en</td>
<td>International Association of Scientific Hydrology</td>
<td>IASH</td>
</tr>
<tr>
<td>es</td>
<td>Asociacion Internacional Hidrologia Cientifica</td>
<td></td>
</tr>
</tbody>
</table>

Table 24: Example concept entry from the DHydro collection showing the relationships between full forms and abbreviations in the three languages of the collection

Relationships between terms, such as degrees of synonymy, and in the above case abbreviations, can be defined as a set of lexical relations $R^L$. We define a lexical relation (interpretation) function:

$$lr: R^L \rightarrow 2^{L^C \times L^C}$$

that associates to each element $p$ of $R^L$ a relation $lr(p) \in L^C \times L^C$. For $(t_1, t_2) \in lr(p)$. We may then write $p(t_1, t_2)$ [or $t_1 p t_2$]. So, for example, we can have: “International Association of Scientific Hydrology”, “IASH” $\in L^C$, “abbreviated form” $\in R^L$, (“International Association of Scientific Hydrology”, “IASH”) $\in lr$ (“abbreviated form”). We then write:

“abbreviated form” (“International Association of Scientific Hydrology”, “IASH”).

Examples of lexical relationships within ISO 12620 include, but are not limited to those shown in Table 25 below.

<table>
<thead>
<tr>
<th>Dictionary ID</th>
<th>DC ID</th>
<th>DC Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.2.1</td>
<td>ISO12620A-02010801</td>
<td>abbreviation</td>
</tr>
<tr>
<td>A.2.1</td>
<td>ISO12620A-02010802</td>
<td>short form of term</td>
</tr>
<tr>
<td>A.2.1</td>
<td>ISO12620A-02010803</td>
<td>initialism</td>
</tr>
<tr>
<td>A.2.1</td>
<td>ISO12620A-02010804</td>
<td>acronym</td>
</tr>
<tr>
<td>A.2.1</td>
<td>ISO12620A-02010805</td>
<td>clipped term</td>
</tr>
<tr>
<td>A.3.2</td>
<td>ISO12620A-0302</td>
<td>false friend</td>
</tr>
<tr>
<td>A.10.18.5</td>
<td>ISO12620A-101805</td>
<td>homograph</td>
</tr>
<tr>
<td>A.10.18.5.1</td>
<td>ISO12620A-10180501</td>
<td>homonym</td>
</tr>
<tr>
<td>A.10.18.5.2</td>
<td>ISO12620A-10180502</td>
<td>homophone</td>
</tr>
<tr>
<td>A.10.18.6.1</td>
<td>ISO12620A-10180601</td>
<td>antonym-term</td>
</tr>
</tbody>
</table>

Table 25: Examples of relationships between terms within ISO 12620
ISO 12620 contains other relationships between terms and concepts which we could further define with functions between terms and concepts (for example, \( f: \mathbb{R}^X \to \mathbb{Z}^{C \times T} \)). The full set of formal descriptions for ISO 12620 may be useful to the terminology community in assigning further items to this standard, however for the purposes of this section, we have demonstrated that the mapping of terminology to ontology, informed by the standards of ISO TC37 is a realistic exercise. Subsequent proof of this mapping may be of interest, however this work is at such an initial stage that the possibility it opens up requires further exploration and evaluation elsewhere before its effectiveness can be concluded.

4.3 System Description

We have described a method for the automatic extraction from arbitrary collections of text, of domain terminology. For a terminology collection, the necessary output is a collection of terms, perhaps with associated definitions, contexts, source information and so on. We have considered how to formalize the text-based empirical method for automatically extracting terms that represent concepts, and made consideration of the relationships between the terms and also relationships between the concepts such that they can form the potential basis for a terminology and, subsequently, an ontology.

The input, an arbitrary collection of text in a specialist domain, produces two outputs: a terminology markup format (TMF as shown above), from which we can produce terms in a terminology markup language (TBX, say) and in an ontology interchange language (RDFS, say). The first phase of analysis uses a single value, a z-score, to select words based on a combination of high-frequency and high-weirdness. This produces a set of words (single tokens). These words are propagated into the collocation phase, where bigrams are produced, and iterated over (recollocated) to produce a tree-structured list of multiword terms of various lengths. This stage requires values for the u-score and collocation z-score, along with constraining the accepted collocation position to one word either side. Finally, indications of word variants are provided which may be of use in aggregating certain results to reduce the tree structure by combining synonyms.

The production of our results can be adapted by changing the values of various parameters used in their calculations (mostly z-scores). This description produces the simple system architecture shown in Figure 34 below.
Chapter 4

Figure 34: System Architecture Diagram

An initial prototype was made by combining System Quirk’s KonText (v6.0), augmented and improved by the author, with programs developed *ab initio* by the author including ColloQator version 1.0 and the MatchesN program. Results were analysed using Microsoft Excel. This provided a first test of the recollocation capability. The ability of ColloQator to provide recollocation information for only the top ranked pattern proved to be beneficial to work being carried out by a large number of colleagues at Surrey, and results produced by this application have been published by these colleagues also, providing for independent third-party evaluation.
The extension of the method to include seeding the collocations through high-frequency high-weirdness combinations was considered in the INTERVAL project (Gillam and Ahmad 2002), though not implemented or evaluated. The current version of the system implements most of the functionality discussed above. The system has been developed using the object-oriented Java programming language, using Java libraries for XML, XSLT and the aptly titled Horrible Spreadsheet Format (HSSF) functionalities.

Figure 35: The ColloQator research tool

The Horrible Spreadsheet Format (HSSF) libraries provide a means by which these results can be exported to Excel – a more familiar environment for some. In this transformation, the tree-structure is lost, with each item in the tree becoming an Excel Worksheet, which contain the collocation values produced (Figure 36).
This tree-based representation of extracted collocation patterns can also be exported as a candidate ontology to Protégé (Figure 37).
4.4 Testing and evaluation

4.4.1 Case study: Carbon nanotubes

A corpus of 1,012,096 words comprising 404 learned articles from the Applied Physics Letters section on Nanoscale Science and Design (average of about 2500 words per article) was analysed. The 404 articles produce a list of 26861 words. Based on smoothed weirdness, and a z-score of 1, a subset of 19 words was considered for collocations. These 19 words are used as the seeds in our collocation process. For each word, we look at the words collocating within a distance of -5 to +5. We use $U > 10$ and $z > 1$ to provide collocations, and subsequent recollocations. Some of the resulting collocations, with frequencies, are:

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>aligned carbon nanotubes</td>
<td>48</td>
</tr>
<tr>
<td>vertically aligned carbon nanotubes</td>
<td>15</td>
</tr>
<tr>
<td>aligned carbon nanotubes kai</td>
<td>4</td>
</tr>
<tr>
<td>multiwalled carbon nanotubes</td>
<td>46</td>
</tr>
<tr>
<td>multiwalled carbon nanotubes mwnts</td>
<td>13</td>
</tr>
<tr>
<td>single-wall carbon nanotubes</td>
<td>24</td>
</tr>
<tr>
<td>single-wall carbon nanotubes swnts</td>
<td>4</td>
</tr>
</tbody>
</table>

Interestingly, single-walled carbon nanotubes is also extended by swnts ($f = 19$), however when we consider multiwall carbon nanotubes, the mwnts extension does not satisfy the conditions. There is, perhaps, some tension between wall and walled within this collection. Analysis by hand (non-expert) would suggest that vertically aligned carbon nanotubes is valid, while single-wall carbon nanotube and multiwalled carbon nanotubes appear to only be extended by abbreviations (over-zealous removal of punctuation). Extending this analysis to lower frequencies (by relaxing $U$ and $z$ constraints), we can find the following terms of increased length:

- conventional horizontal-type metalorganic chemical vapor deposition reactor
- ridge-type ingaas quantum-wire field-effect transistors
- trench-type narrow ingaas quantum-wire field effect transistor

From the resulting list of term candidates, we can consider the matching mechanism for determining potential term clusters (possible synonyms)

If we consider multiwalled carbon nanotube we find two other terms of a similar nature: multiwalled carbon nanotubes and multiwall carbon nanotube. These variants match with a value greater than 0.9, so we can present them as candidate synonyms. Using this approach, validity of
the candidate terms can be manually evaluated at the same time as the resulting 'tree' (conceptual structure) generated from it. A resulting concept tree, containing candidate terms and their frequencies, is presented below in Figure 38, where the candidates in boxes with solid outlines are considered valid by the author.

Figure 38: Tree representation of the candidate compound terms being extracted using this method

From this tree-based representation, we produce the terminology and ontology markup. We use the subtype/supertype relationships in ISO 12620, along with other identifiers in 12620, and structural elements of ISO 16642 to form the basis for a hierarchical conceptual structure of a terminology collection. The example below shows the resulting markup based on TMF’s XML-based generic mapping tool (GMT) for *aligned carbon nanotubes* and *vertically aligned carbon nanotubes* (we assume here that *vertically aligned carbon nanotubes* IS-A *aligned carbon nanotubes*).
For candidate synonyms, we create a further term entry (TE) containing a language section (LS) for English, which has a term section (TS) for each synonym: this would represent one concept. The combination of ISO 12620 and 16642 facilitates the derivation of complex markup formats such as TBX. Although the information provided above is in quite a verbose format, it ensures consistency of data category use, and the application of XML-based filtering (using XSLT) enables automatic conversion to a TBX compliant format, from which systems capable of importing TBX will be able to populate their database models. Other formats that conform to this combination will be able to do the same, providing they have the necessary conversion filters.
We use the information above in the concept-base for an ontology, which is possible since the basis for these ontology interchange languages is the supertype/subtype relationship. The TMF version is converted, using XSLT, to the Resource Description Framework Schema (RDFS). In this conversion, there is a degree of information loss, since RDFS does not cater for much of the information available in a terminology format, however the translation to an ontology language shows the ability to directly populate an ontology.

```xml
<rdfs:Class rdf:ID="aligned_carbon_nanotubes">
  <rdfs:label>aligned carbon nanotubes</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:ID="vertically_aligned_carbon_nanotubes">
  <rdfs:subClassOf rdf:resource="#aligned_carbon_nanotubes">
    <rdfs:label>vertically aligned carbon nanotubes</rdfs:label>
  </rdfs:subClassOf>
</rdfs:Class>
```

Ontology editing applications that understand RDFS, including Protégé, can use such output to seed their ontologies for further development.

### 4.4.2 Peer review

Parts of the method discussed above have been presented and peer-reviewed in various papers based on prior work. These include:
<table>
<thead>
<tr>
<th>Focus</th>
<th>Description</th>
<th>Authors/year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Terminology</strong></td>
<td>Analysis of a semi-automatically collected corpus of text containing contextual references to AIDS, presenting an initial method was presented for validating terminology collections, expanded upon in this thesis.</td>
<td>Gillam and Ahmad 2002</td>
</tr>
<tr>
<td><strong>Extraction</strong></td>
<td>Analysis of a History of Science text, Newton’s Opticks, with the goal of making valuable, fragile, historic materials in the archives of the Royal Society more widely available to the public.</td>
<td>Gillam et al 2002</td>
</tr>
<tr>
<td></td>
<td>Analysis of a corpus of 435,000 tokens about the semiconductor devices known as Tunnel Diodes, with results partially validated by an expert in the subject, Professor Michael Kelly, then at the University of Surrey.</td>
<td>Ahmad and Gillam 2001</td>
</tr>
<tr>
<td><strong>Modelling</strong></td>
<td>Developments relating to ISO 639 for Language Codes: description and use of the Linguasphere System of language tags, and conformity with ISO 12620 (forthcoming revision).</td>
<td>Dalby, Gillam, Cox and Garside 2004</td>
</tr>
<tr>
<td><strong>Knowledge</strong></td>
<td>International standards for marking up terminology, details of their use and conformance.</td>
<td>Gillam, Ahmad, Dalby and Cox 2002</td>
</tr>
<tr>
<td><strong>Resources</strong></td>
<td>Bootstrapping ontological representations from text based on a convergence of statistical and linguistic methods via terminological representations</td>
<td>Gillam and Tariq 2004</td>
</tr>
<tr>
<td></td>
<td>Hand-crafting of a knowledge-base from text that modelled features of neurons of the visual system in which the difficulties of populating a knowledge base with information about neural cells were discussed and prototype frame and conceptual graph-based systems were prototyped. These systems were built before the existence of the Unified Medical Language System (UMLS).</td>
<td>Gillam and Ahmad 1996</td>
</tr>
</tbody>
</table>

Aspects of the work have been validated in EU co-funded projects. The first of these was INTERVAL, where the two independent reviewers of the project regarded the initial version of ColloQator and the framework as ‘excellent’, and these items were further evaluated at an EU demonstration review attended by representatives of over 30 other projects. The second was SALT, where consideration was made of the mechanisms for terminology interchange, and which resulted in the publication of an international standard. Subsequently, the theoretical basis of the expansion of the work from terminology extraction to ontology/concept acquisition has been carried out in the development of this thesis. Further projects to which contributions have been made were also peer-reviewed, including the ACE and GIDA projects.

Evaluation of the collocation system has been carried out by five previous Surrey PhD students and is part of continuing work by a further seven PhD students. Their works concentrate on aspects of (a) the efficacy of terminology/ontology management systems in information extraction and (b) the use of terminology, its evolution, and the concomitant conceptual evolution, in emergent domains. A number of papers have been published by various colleagues since 1997 in
which results from this system have been used to test various hypotheses in fields such as safety, dance, finance, crime, and so on.

4.4.3 Evaluation of results

The most significant evaluation for any system is comparison to human performance. In certain subject fields, a "gold standard" is produced by which other systems are benchmarked, e.g. in the information retrieval community. There is currently no such gold standard for either terminology extraction, or ontology acquisition; the author made a plea for the creation of a gold standard terminology which could be used for scientific comparison of terminology extraction methods at the 2002 TKE conference since current comparisons are highly ad hoc. It is possible, of course, to consider the documents produced by the European Union as forming a useful corpus for such a task to be carried out with, however there would be significant debate about treating, for example, terminology contained in Eurodicautom as a gold standard. Such a gold standard would therefore seem to be some way off but could form an interesting basis for future work.

It is recognised that expert time is valuable, and our work aims to reduce the "knowledge acquisition bottleneck" by reducing initial efforts required in domain understanding. A slightly different approach to evaluation is therefore taken by considering other published literature, in this case texts again from the University of Surrey Library. The benefit of this approach is that the same resources can be referred to by other interested parties, and the "explanations" within will not change, whereas experts may vary their responses depending on a number of factors.

Three books were selected, primarily based on their recency and availability, that dealt with aspects of nanoscience:


None of these texts contained indexed references to nanowires, nanostructures, nanorods or nanocrystals, perhaps due to the recency of these terms. We consider, therefore, what we can determine about carbon nanotubes from these texts.
From the first of these texts, we consider the index. Listed under *Carbon(s) or carbonaceous materials* are *nanotubes*. These nanotubes appear to be either *multi-walled* or *single-walled*. The single-walled variants appear to relate to a *transistor* (Figure 39).

Interestingly, the page to which this refers contains both a *Carbon Nanotube Transistor*, and a *single carbon nanotube field-effect transistor*. There is no direct mention of a "single-walled carbon nanotube transistor", and here an expert would be required to explain this association.

The second text refers to *carbon nanotube* in its index, but subelements of this tend to be other characteristics such as *curvature, diameter, growth rate, oxidation* and so on. Chapter 3 of this text is titled *Structure of a Single-Wall Carbon Nanotube* (page 35), a footnote to which claims that "Many carbon nanotubes that are observed experimentally are multi-wall structures". Chapter 5 of this text (page 73) refers to both *single-wall carbon nanotubes* and *multi-wall carbon nanotubes* – slight variations in word use from the first text (Figure 40).
This chapter describes synthesis methods for carbon nanotubes, with primary emphasis on single-wall nanotubes. Two relatively efficient methods to synthesize single-wall carbon nanotubes have been identified: laser vaporization and carbon arc synthesis, and both methods depend on the use of catalysts. Other techniques such as vapor growth are also reviewed. Also discussed in this chapter are the synthesis of multi-wall carbon nanotubes, the purification of carbon nanotubes, the insertion of metals into the hollow core of carbon nanotubes, and the doping of carbon nanotubes with alkali metals.

Figure 40: Fragment of text from Saito, Dresselhaus and Dresselhaus (Chapter 5, page 73) discussing synthesis of different types of carbon nanotubes

The third text’s index has carbon nanotubes (page 41) and nanotubes (268, 269). In this text, the variants of nanotubes are referred to on page 268 (Figure 41)

Advanced topic: catalytic growth of nanotubes

Growth can involve a catalytic agent. This is particularly important in the formation of nanotubes [27]. Fullerenes such as \( \text{C}_{60} \) are formed in any number of reactive systems in which carbon is volatilized and allowed to condense. Single-walled and multi-walled nanotubes (SWT and MWT, respectively) of carbon are formed when certain metals are added.

Figure 41: Use of single-walled and multiwalled in relation to nanotubes in the Kolasinski text

The importance of these “walled” nanotubes, and the variation in labels being used, is apparent from these three texts. We have identified single-walled, single-wall, multi-walled, multi-wall and multiwalled. With reference to Figure 38 previously, evaluation according to “in-text” use, and indexed use, appears to part-validate the tree of candidate compound terms extracted from the collection of texts, as shown below in Figure 42. While this is not a comprehensive evaluation of the approach, it is certainly encouraging, and suggests the possibility for further opportunities in such mutually validating approaches.
4.5 Conclusion

In this chapter, we have presented a system for the production of a candidate conceptual structure from an arbitrary collection of texts in a specific field of interest. In Chapter 3, we saw how patterns of language use had various statistical similarities across different specialisms. Based on these similarities, we can suggest that the method presented is suitable for extracting conceptual structures from such arbitrary text collections. The structure extracted relates back to Chapter 2, since it provides evidence of how knowledge is being made at least partially explicit within such texts. The link between knowledge and language use is not a simple one, though this is perhaps due to Zipf’s principle of least effort, of which ellipsis is a prime example: a vertically aligned carbon nanotube may be referred to throughout the remainder of the text as the nanotube. This provides a challenge for text analysis, since we assume that nanotube, when it occurs in text, is the same nanotube. In one text, this may be a single-walled nanotube, in another, a vertically aligned nanotube. Here we have a problem of both reference and counting. Such a problem we cannot treat here.

The system presented is intended for use in ongoing research and development to help our understanding of term formation, collocation, local grammars, and conceptual structures. The ontology wrapper for terminology, described using Data Categories from ISO 12620, has been demonstrated, so we have partially reduced the seeding of an ontology to the previously unsolved problem seeding a terminology collection. The mapping from terminology, built semi-automatically using a variety of extraction techniques, to ontology for use in developing...
Chapter 4

intelligent systems is a contribution worth evaluating. By acquiring the ontology from different collections of text, and perhaps over time (diachronically), we can consider how different versions of the ontology change with the domain. It may be possible to determine how such a specific ontology can be made more general, and indeed for adaptability and reuse this may be necessary. This would form a more considerable, although longer term, aspect of such an evaluation. Elements of the method have been evaluated in peer-reviewed publications, with various results presented in these publications and in conferences. These provide confidence in the efficacy of the method.

In focussing on an emerging discipline, Nanoscale Science and Design, we found no existing dictionary or ontology that could be used to validate our results. Instead, we looked at how small elements of what had been extracted were elaborated and indexed in textbooks of the discipline. What we found is encouraging since it tends to validate the method for this extraction, beyond that already published. It may be interesting to consider how the experts produce indexes, and why they produce variants of a particular term, since this takes us back to the link between knowledge and language. Longer term, we would like an expert to evaluate the structures that are automatically extracted from these collections of text. Here, we would consider multiple experts in each discipline, and multiple disciplines. We could devise an experiment whereby an expert and a number of students of each specialism hand-craft ontologies, and the outputs are compared to the system, but this would require us to explain the principles of terminology/ontology. We have to consider, also, that any iteration of such an experiment is slightly tainted by such understanding, perhaps skewing the result. An alternative would be to evaluate whether domain experts can craft an improved ontology given a seed ontology extracted directly from text, although the domain expert has no understanding of ontology *per se*. This, perhaps, would produce a thesis in itself— or maybe several theses.

The method presented, it appears, can produce a structure of the concepts for the expert, without the expert needing to understand the principles of terminology science or knowledge representation. This could be considered as a contribution towards overcoming the “knowledge acquisition bottleneck” since the results can be used for the creation of knowledge bases. There are benefits to the use of such a method. The population of terminological resources has various benefits, for example:

- terms with their contexts can assist terminologists carrying out tasks involving translation; a text collection can be indexed by its key terms;
- texts can be summarised with reference to specific terms (focussed);
• documents can be managed by their terms, updated when terms are deprecated, and workflow (specifically, versioning) of documents assisted in this fashion;

• aspects of terminology standardisation can be assisted

Likewise, population of an ontology —so-called “ontology learning” — can have similar benefits, for example:

• it can assist in the development of intelligent information systems where instances of the concepts in the ontology can be used in the development of rule- or case-based expert systems;

• it can be used in systems for information retrieval (query expansion) or information extraction;

• it has implications for knowledge management systems;

In comparison to other leading systems, the system discussed has specific advantages. In Text-to-Onto, maximum term length and frequency threshold are determined by the user. The user is also expected to add any extracted terms to the tree and creating the ordering manually. Too low a frequency threshold produces a long list of terms for post-editing. Too high will remove the interesting multiword expressions (since these are considered together). Text-to-Onto does not provide contextual information to assist in this process. Text-to-Onto does, however, provide information about relationships between extracted terms, and although the exact nature of these relationships is not clear, certain linguistic patterns are suggested. Our method automates the selection of interesting single words by frequency, and the program decides term length by collocation. Furthermore, by semantic inclusion, we produce an initial tree for user adaptation. This approach would seem to reduce the burden of interpretation on the user. While over-generation of results is a concern in our system, using an appropriate ontology editor inappropriate subtrees can be quickly pruned, and trees that are more appropriate at higher levels can be moved there. Our system provides contextual information for the collocations of words at all levels such that a (human) judgement about the correctness of the terms could be used. Subsequently, linguistic patterns could be used to augment the tree, and this has been considered elsewhere (Gillam and Tariq 2004). Systems such as KAW, ASIUM and others depend in large part on the accuracy of the part-of-speech tagger employed. Our method uses frequency information only, so is not dependent on the effectiveness of training a tagger, a process that requires prior domain knowledge. It may be interesting, post hoc, to test various parsers on contexts of the extracted terms, especially those of significant length (> 5), in evolving specialist domains to determine the effectiveness of a combined approach, and this may relate to work in
local grammars. We could remove the *a priori* need for POS tagging, and may be able to consider further means for evaluating headedness and correctness of the extracted term. Our lack of reliance on POS tagging initially is a significant difference to other work, and work in local grammars may well show that reliance on POS tagging in specialist domains is not necessarily an appropriate initial step.
5 Discussion and Outlook

I have looked at the link between knowledge and language, specifically explicit knowledge as presented in the text of specialisms. Based on the notion that a collection of texts of a specialism expresses some of the knowledge of that specialism, I have looked at how it might be possible to automatically extract specialist terminologies from these texts, and how the things (and processes) in specialist domains are expressed using terms that may be composed of one or more words. The possible mapping between terminological concepts as stored within a terminology collection to an ontology was demonstrated, making use of standardised identifiers from ISO 12620 on Data Categories and ISO 16642 on a Terminological Markup Framework. Indeed, the work presented has made reference to ISO standards throughout. Populating terminologies and ontologies are both broad goals, with significant subsequent work required for associating the information to other information both in the ontology and in the terminology collection. Here are some of the themes touched upon on this thesis.

Semantics

Chapter 2 investigated the link between knowledge and language with reference to Ogden and Richards' meaning triangle, from a terminological perspective, and onto the relationship to Knowledge Representation and hence to the burgeoning field of ontology. Chapter 2 briefly touched upon how we think about language, and how we think in language to express "things" in the world. On difficulty of describing language is that we need to use language to do so, and this is possibly more difficult in the terminology of specialisms due to different devices in language.

The essence of terms

Chapter 3 demonstrated how international standards for the collection, storage and sharing of scientific terminology could assist in the sharing and understanding of these subjects for both human and machine. The very essence of terms and terminology collections was considered with reference to the definition of a term as stated in ISO 1087-1 and other relevant standards. Sowa has described an ontology as a catalog of the types of things that are assumed to exist in a domain of interest (D) from the perspective of a person who uses a language (L) to talk about D. The parallel between this definition of a term according to 1087-1, and Sowa's description of an ontology provides us with further justification for bridging the terminology-ontology divide.
Term distributions and collocations

Various metrics for terminology extraction were considered with reference to Zipf’s law and his principle of least effort, frequency distribution, weirdness (after Malinowski), and Smadja’s statistics for patterns of collocation, and how to use these in combination for identifying terms and organising them hierarchically by inclusion. While these metrics are well-known, their combination is perhaps innovative.

Terminology and semantics

Grounded in the work of Wüster, recent work in terminology standards through ISO, specifically in ISO TC 37 concerned with “Terminology and other language resources”, lead us to consider that these standards could support the development of ontologies by considering the conceptual relations between terms. The hierarchical organisation, along with other relationships between terms can be used for populating so-called ontologies that many see as being vital to the Semantic Web initiative and also important in the Semantic Grid, and in developing the more elusive Knowledge Grids. Once a terminology collection or ontology has been populated, other considerations occur, such as updating the collection, or mapping to collections developed elsewhere: an unchanging collection is difficult to comprehend.

Terms, concepts, and corpora

Chapter 4 proposed and demonstrated a method based on the terminological analysis of text collections that uses an arbitrary collection of text in a specialist domain as its input and automatically extracts a candidate terminology / candidate ontology. The system that prototypes this approach has output suitable for human interpretation at various stages. The candidate ontology can be modified using freeware ontology editors such as Protégé, and subsequently this provides a means by which to instantiate knowledge bases. While the international standards for computer collection of terminology are available, we used an RDFS encoded export format for use with Protégé.

Various aspects of this work have been carried out since 1995 for the purposes of extracting key information from a domain, and modelling and harnessing that information such that intelligent systems could be built. The starting point of my personal research was a study of aspects of neurons in the brain, specifically the visual cortex, to provide for a system of classification of neurons themselves. Over time, this research has evolved to encompass terminology management and, lately, ontology. Various elements of the work were evaluated in research projects referred to in Chapter 1, and in studies that have led to the publication of research papers. Evaluation of the approach has in part been carried out through peer-review of these publications, with
subsequent expert evaluation expected to follow. The archetype system for extracting terminology from text and subsequently populating an ontology, referred to elsewhere as ontology learning or the production of a seed ontology, is being developed further to provide the means by which to carry out such an evaluation. Here we have demonstrated the efficacy of the approach.

I believe that the work presented contributes in some ways towards solving the well-documented and outstanding AI problem of producing an initial model of an arbitrary specialist domain from background resources without significant hand-crafting effort and involvement of a domain expert: the so-called "Knowledge Acquisition Bottleneck". This bottleneck is usually overcome through extensive interactions with domain experts, involving a number of expert interviews. The research presented has explored issues of terminology extraction from domain texts, the need for and use of knowledge representation, and the means by which terminology extraction tasks and knowledge representation can be integrated seamlessly in times to come.

By producing better representations and algorithms, we may be able to adapt systems to various tasks in more effective ways, but we should not expect that such improvements will lead to more autonomous, intelligent, or evolving computer systems. To produce intelligent systems, there is a need for some form of knowledge of the terms, concepts, facts, and rules of thumb – the basic building blocks of knowledge. Such systems need this “common” sense, and, ideally, an ability to communicate in a natural language, such as English, to autonomously enlarge the knowledge base, and to perform experiments to make further headway in understanding. Humans easily delimit knowledge search and can converse with degrees of common understanding – though ambiguities and social differences can be interesting. Machines need to have this information encoded. Humans do have knowledge which can be difficult to encode – knowledge of how to adapt in a situation, e.g. crossing a busy road, playing a sport, driving a car – and an ability to take action based on prediction, such as steering a boat correctly according to the expectation of events that could be caused by a large approaching wave. As humans cannot easily understand these actions, reducing the actions down to a machine-processable level is perhaps a longer-term challenge.

The knowledge acquisition framework, and the associated system, is the product of 15+ years of work at Surrey. The author of this thesis has worked on extending the scope of text-based terminology acquisition work at Surrey to a text-based concept acquisition work. Further work has already been taken up based on this method by about 7 other PhD students at the University of Surrey using collocation patterns, local grammars and studying ontology generation. The archetype system, its predecessor and its successors will further evolve this work.
5.1 Possible impacts

If we have information, extracted from texts that can be represented in an ontology interchange format, we then have a means for populating an arbitrary KB. This ontology populating activity is referred to elsewhere as ontology learning. The result can be used to support the development of intelligent systems (for example Expert Systems), assist domain modelling activities, and perhaps contribute towards the development of the Semantic Web / Semantic Grid / Knowledge Grid. Methods for automatic extraction of complex terms and term structures were presented with reference to a number of standards that either exist or are in development to foster the exchange and reuse of these results.

The ability to understand the relationships between words and other items within, for example, web texts, is increasingly of importance with the ever increasing amounts of information being created. It has been estimated that 1-2 exabytes of information are produced every year (Maybury 2002), and this keeps increasing. The original purpose of the Web was to share information about scientific experiments. The language used to describe scientific subjects can be difficult to understand for non-experts, or indeed those beginning multi-disciplinary research activities, and so specialist languages remain of significant interest for this kind of work.

British-born World Wide Web inventor Sir Tim Berners-Lee, FRS, coined the term “Semantic Web” to describe an intelligent successor to the current Web that can interact with both humans and machines alike. The challenge in creating the Semantic Web is of developing systems able to link, understand and reason over information from disparate sources. This requires common, machine-processable, descriptions of "things", ranging from information about how texts relate to one another, to how words and pictures within the texts can be interpreted. Through combinations of metadata identifiers, domain ontologies and intelligent agents, communications will be mediated to provide access to a range of services. The challenge lies in the move from the unstructured nature of the current Web, to either highly intelligent adaptive agents capable of the disambiguation and linking of information, or the use of more consistent forms of data through well-developed highly granular systems of markup. For the development of the Semantic Web, some authors (Alani et al 2003) have identified the need for domain-specific semantic knowledge, and others have approached this subject through ontology-learning as a means to provide this (Navigli et al 2003).

In section 1.1, part of the hope for this research was that it “may help to look afresh on issues raised in knowledge management”. Applehart et al argue about maintaining a "common vocabulary" for Knowledge systems. We contend that this is possible by keeping a check on terminological usage, to ensure that the vocabulary is kept up-to-date. We have shown methods
for extracting common vocabulary from existing texts. Once identified, new patterns could potentially be extracted from newer texts in similar ways, although the method presented here may require modifications to enable this in specific domains. Identification of such patterns could help identify changes in knowledge: a functional aspect of a Knowledge Management system that is potentially beneficial. The common vocabulary also makes a number of other monitoring tasks, such as document authoring, easier. This leads us to promote the method presented as useful for creating this "common vocabulary". In essence, what we have described is the application of statistical analysis with reference to techniques of linguistic analysis, the combination of which provides part of the structural data necessary for such KM tasks. It should be possible to develop and maintain a common language by incremental analysis of terminology use through the method outlined. Extraction techniques have been presented that identify new terms in relation to a reference corpus such as the BNC. The extension would be to make the existing corpus into a reference corpus (or monitor corpus) to identify and extract new terms from new texts. This could lead to discovering evidence of innovations or process changes. The ability to identify interesting new ideas, through monitoring the output of authors in specific domains, could be beneficial to industry at large; clearly a desirable function of a Knowledge Management system.

5.2 Contributions

The key contribution is the automatic extraction of terminology from an arbitrary collection of text in a specialist domain. The result of this extraction is a candidate conceptual structure. The utility of this contribution can be considered as the use of automatic terminology extraction in populating a standards-based terminology format and, subsequently, an ontology format.

In comparison to other leading systems, the system discussed shows the power of automation: in comparison to Text-to-Onto, maximum term length and frequency threshold are no longer need to be determined by the user. A key difference to most systems used for extracting such conceptual structures is the non-reliance on part-of-speech tagging. The statistical approach to terminology extraction as presented may be more easily extended for use in other languages, or provide an initial seeding for such tasks subsequently.

Initial results are interesting, and suggest further exploration is required to establish the potential of this method, including further peer-reviewed publications.

The practical contributions made in this respect are:

- testing the applicability of Zipf’s law to specialist text
• automation of selection of candidate terms (single words);
• implementation and extension of Smadja’s collocation algorithm
• consideration of clustering of terms;
• mapping between a standards-based terminology and a burgeoning ontology format

The last of these points enables precisely the intended use of the results within both a terminology system and an ontology engineering system.

This thesis, hopefully, contributes to:

• research agendas of language and knowledge, specifically terminology and ontology
• the scientific community’s ability to exchange information
• improving the tools used in language research
• fusing methods in terminology extraction with knowledge representation

5.2.1 Evaluation of research questions

We consider, here, the extent to which work done in respect of the thesis provides answers to the research questions (presented in 1.2.1). Each question is reproduced (in italics) with a short summary provided in relation to how the work presented may have helped to answer it.

1. How different are languages of specialisms from general language?

Although patterns of language use have similarities, deviations from expectations as according to Zipf’s law are certainly of interest, particularly in consideration of the more restricted use of low frequency words as evidenced by there being only 40% of words used once in the specialist collections, in contrast to the expected 50% and the above 50% value for BNC. This question is usually answered through introspection or through questions of use. I have attempted to determine whether the difference could be studied more objectively, and further work may be able to provide an objective consideration of this difference.

2. What is the link between language and knowledge? Does language use accurately reflect what is known?

The link between language and knowledge has been studied since time immemorial, and I have considered it here also. The way in which people encode, or pack, information into a simple “tag”
such as a word or phrase that stands for it, and can use this for reasonably effective communication, still requires further study. Language is used productively in knowledge dissemination: a small number of lexical units are used to describe an equally small number of concepts, and it is through the production of expressions comprising multiple such lexical units that concepts exist in profusion. These combined units "pack" additional information such that the combined expression becomes more than the sum of its parts. Furthermore, people can adapt to different granularities of description, and change what they understand by such tags (semantic shift) reasonably easily. Such adaptivity is a challenge for computer systems. This link remains an open question, though I have suggested how these combined units can be treated computationally.

3.  Is it possible for a machine (computer) to develop an understanding of language?

Despite significant research efforts, the immediate answer to this question would have to be: not yet. However, as I have presented, there are repeated patterns in the use of language of specialisms that may provide an eventual basis for such an understanding. If we were unable to present repeated patterns, we could not compute them and hence mechanisation would certainly not be possible. Whatever made humans predisposed to language, or led somehow to the combined development of brain and language, and why it has developed in different ways in different geographical or social contexts, is not yet understood. Language is a somewhat fluid mechanism for communication that can be exploited for purposes of subversion, and such intentions are not easy to identify mechanically. Controlled vocabularies improve the possibility, and considerations of limited subsets of language can be effective to some degree, but this remains an open question.

4.  Can the population of "domain ontologies" for use in the Semantic Web benefit from work carried out in the population of terminology collections on the extraction of information from text?

From the work presented, there is some benefit to be had. Whether these terminology collections, augmented with conceptual information, can be utilised in the Semantic Web needs to be evaluated: the attitude in the Semantic Web community is rather strange in that terminology and ontology are either used interchangeably in the literature or an ontology is described as a collection of terms or words, e.g. Wordnet. Understanding how these words/terms are conceptually related when the terms in the collection number in the millions, and in multiple languages, becomes a task for which human consideration becomes difficult at best, and the challenge becomes one of understanding the resource before its use can be evaluated. My work would appear to suggest that there is benefit to be had, but further evaluation is necessary.

163
5. Can studying the link between language and knowledge contribute to the scientific community's ability to exchange data?

If we can identify key domain concepts through their representation by terms, and the associations between these concepts/terms, this may help the scientific community to exchange data. Scientific journals produce human-oriented materials, from which talented researchers are, perhaps, able to recreate the experiments presented. The need for each reader of such material to extract the key differences between one experiment and the next may currently hinder efficient scientific experimentation. This in some ways resembles the notion presented by Tim Berners-Lee of “screen-scraping” on the web: in this instance, researchers carry out experiments with various parameters. They encapsulate the parameters and results of use of these parameters in the terminology of their discipline. Other researchers then use computers to decipher the terminology such that they can get to the parameters, and carry out similar experiments, so continuing the cycle. Perhaps an “answer” to this is: don’t use language to exchange scientific knowledge? In lieu of acceptance of this answer, my work as presented would seem to make such a contribution.

6. Can studying the link between language and knowledge lead to improvements in the tools used in language research?

This question has not been answered by the research presented. Certainly by studying different aspects of language, and differences within and between languages, we may eventually improve the tools used in language research. If the nature of the link between language and knowledge can be found, and is shown to be computable, it would be expected to lead to improvements. This question remains open.

7. Is it possible to combine techniques in terminology extraction with those in knowledge representation?

The work I have presented suggests a positive answer to this question, although only currently in providing an outline of an ontology, or rather the basis for an essence of the ontological description. Further work is needed for evaluating this provision, and subsequently considering how to create a more extensively populated ontology.

8. Can we develop a computer system to assist in the semi-automatic construction of concept systems or ontologies: hence, can we bootstrap a domain-specific ontology from a text collection of consensus within a domain of discourse?

The work I have presented suggests a positive answer to this question also. Again, though, this is a limited provision. It may be undermined by “common” knowledge that is not contained within
the texts of the domain, or from the perspective of the work presented that which is not presented within the arbitrary collection under consideration. Each text, even in a large collection, is only meaningful if read in the context of a larger number of texts: in scientific contexts, the terminology, and ontology, will be consensual. An arbitrary collection of such texts should provide some degree of consensual agreement of “knowledge”, hence it should be possible to bootstrap some form of domain-specific ontology from such a collection of text, but the limitations of this need to be understood.

5.3 Questions for Future Research

There is a wealth of future work that could be considered. The various research questions presented in Chapter 1, and evaluated earlier, would certainly benefit from further exploration. It is possible also to consider extension of the method presented, to combine it with other work (at Surrey) on the extraction of subtypes using lexicogrammatical cues. This would extend the population of the ontology, and may help eventually to provide attributes and values for the various concepts being identified by consideration of how characteristics are described in text. How the concepts associate might then provide the final element, axioms, of the domain model and future intelligent systems could be developed through the processing of texts alone. This is perhaps the holy grail of intelligent systems, and is still some way off. With consideration of the research questions previously, the following new research questions could be considered:

- By contrasting different ontologies in the same domain, and here we consider the association to subject classification systems, how could they be merged? What about multilingualism where we may require a different understanding of a word?\(^{16}\)

- What can be gleaned from pluralisation of compounds versus singulants, and what can be concluded if either exists, but its counterpart does not (e.g. what if we found “carbon nanotube”, but not “carbon nanotubes”?)

- Can a systematic evaluation of statistical measures provide conditions for acceptability of low frequency terms, and can we produce a fully statistical selection method?

\(^{16}\) Application of Zipf’s law to various lengths of phrases has been considered for English and Mandarin (Ha et al 2002) in analysis of frequent bigrams, trigrams and so forth as they occur; they do not consider whether the patterns are collocational or terminological.
Chapter 5

- Can the automatic selection of domain terms contribute to feature selection for neural networks (e.g. in providing feature-vectors)?

- Can we use the results presented to support activities such as studying change of language over time (ontology-based)?

- Can we provide a consistent basis (reference collection or test suite) for systematic analysis of the differences between general and special languages: can we compare extraction methods systematically in a similar fashion?

The rapidly burgeoning field of ontology within computing means that publications in this area are ever on the increase, and as such, what is not possible today may well be tomorrow. At some stage, this field will stabilise and begin to develop peer-reviewed standards to parallel those within the field of terminology science. Since after 70 years of development in terminology science, and 50 years in the ISO community, the standards are still not settled, this endeavour could be some time in coming.
References


Dalby, D. (2000) "Linguasphere Register of the World's Languages and Speech Communities". Hebron (Wales). Two volumes: ISBN 0 9532919 1 X and 0 9532919 2 8


References


References


References


Appendix A – Publications


