Terminology and the construction of ontology

Lee Gillam, Mariam Tariq and Khurshid Ahmad

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

{l.gillam, k.ahmad, m.tariq}@surrey.ac.uk

Abstract. This paper discusses a method for extracting conceptual hierarchies from arbitrary domain-specific collections of text. These hierarchies can form a basis for a concept-oriented terminology collection, and hence may be used as the basis for developing knowledge-based systems via ontology editors. This reference to ontology is explored in the context of collections of terms. The method presented uses both statistical and linguistic techniques. The result of such an extraction may be useful in information retrieval, knowledge management, or in the discipline of terminology science itself.

Keywords. terminology extraction, conceptual hierarchies, knowledge-based systems, ontology

1. Introduction

Organization of information is important for all scientific activities. For science to be explored, its phenomena should be both observable and repeatable. Publication of landmark scientific texts, including the first scientific journal, Philosophical Transactions (by the Royal Society in 1665), and works such as Newton's Opticks, Darwin's Origin of Species and so forth, have provided "in-text" organization of scientific information. Increasing numbers of such publications, and the need to provide better organization of information are doubtless among reasons that we now have grandly named systems which we can use to classify information, such as the Universal Decimal Classification (UDC) and the Lenoch Universal Classification (LUC). Pioneers of these classifications used terms such as *universal* to refer to all that exists. Both these systems provide a hierarchical structure for organizing information, although their structural bases differ. These large-scale general classification systems have a particular identifiable problem: they do not keep pace with developments in specific subject fields since they are not easily modified.

Though never named as such, these systems of classification represent earlier manifestations of *ontology*. The Encyclopaedia Britannica (EB) defines ontology as "the theory or study of being as such; i.e., of the basic characteristics of all reality". Use of the word *universal* suggests that these systems are providing some theory or study of being, so promoting their ontological status. More recently, "ontology" has been used to describe the (computer-readable) representation of information about the world in a form in which it can be reasoned over. This still provides for a theory or study of being, however the focus now is utilitarian: the term *ontology* is used extensively in literature on information extraction, Knowledge Representation, and with reference to the Semantic Web, as a thing to be used, rather than a study of things. Use of ontology as representations of the world have led to its consideration as a tool for developing solutions to problems of translation (Navigli, Velardi and Gangemi 2003), information retrieval (Oard 1997, Guarino, Masolo and Vetere 1999), knowledge management (Maedche et al 2003) and other issues related to knowledge-based activities (Alani et al 2003). The creation of any conceptual system, including a subject classification system or the modern-day ontology, still requires significant human effort. Subject experts, information retrieval professionals, and artificial intelligence researchers specify and design such systems largely by hand. These experts bring to bear their experience, documentation and knowledge in building the systems. The knowledge of experts is already documented to a discernible extent. These systems, classifications, terminologies or ontologies, may be subsequently standardized, for example the British Standard (BS 1000) for UDC, and the current and emerging parts of the ISO 639 standard for classifying resources according to language.

The question here is whether a candidate conceptual system of any kind could be obtained by a systematic examination of documentation in the specialist domain. A computer implementation of a systematic examination could be used to generate candidate conceptual systems that could be adapted, rejected, or possibly standardized. Such a text-based approach has a potential advantage that domain texts may signal changes in concepts that require modifications to the conceptual system, which may provoke subject experts to make such changes.

This paper seeks to explore an ambitious notion of deriving candidate conceptual systems from arbitrary collections of text in specialist domains. The basis for this exploration is terminology: terms, or rather *candidate* terms, can be extracted from texts, and that inter-relationships can be identified. Potentially, other data for a term's attributes (definition, context and so on) could be extractable from such texts, however this is beyond the scope of this paper. The specific organization of these terms, and how they relate to one another should provide such a candidate conceptual system (ontology).

We describe an approach to the automatic creation of ontologies from arbitrary collections of text in specialist domains. This activity requires an understanding of how concepts are articulated in natural language. The 'inevitable' link between knowledge and language suggests that it will not be easy to move away from a *world as words* view for producing computational representations. The paper discusses how to build a candidate ontology from free text using a combination of statistical and linguistic methods. Production of candidate ontologies may help to overcome a principal obstacle in developing intelligent systems: acquiring knowledge accurately and quickly – the so-called "knowledge acquisition bottleneck" (Buchanan and Shortliffe 1984, Luger and Stubblefield 1993). Knowledge engineers to this day manually craft the knowledge of the 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 domain.

2. A terminological perspective on ontology

The term *ontology* has emerged from philosophy to be used extensively in literature on artificial intelligence. For some authors, an ontology is produced by hand-crafting a representation of a specific domain, or by renaming an existing language resource: here, Wordnet and its EuroWordnet variants have been variously renamed (Sowa 2000, Oard 1997). To us the key role of an ontology is to identify areas of knowledge, associate these areas of knowledge with other areas, and demarcate these areas with key terms. The practical import of an ontology is to create systems for storing and retrieving fragments of knowledge; for example, an ontological understanding of a subject allows one to understand inter-dependence of certain key concepts, and related keywords, and the independence of other concepts. The dependencies, or otherwise, may help in query expansion through the detailed specification of 'objects' within a knowledge base.

In what some may consider *the* definition of terminology, ISO 1087-1, a *term* is defined as a "verbal designation of a general concept in a specific subject field". If a system such as UDC can be considered as an ontology, a term, by this definition, already has an ontological status. Supposing, now, that multiple concepts can belong to the same subject field, and that these concepts may also exhibit conceptual organization, we have an association between two conceptual structures both of which may be referred to as ontologies. The effect of this combination is rarely considered since the principal challenge for terminology is to construct the system of terms for some purpose, primarily translation. Use of systems of terms for purposes beyond translation is perhaps a motivation here: terminology science may be able to contribute significantly to the current debate on ontology, perhaps. The automatic extraction of terms could assist in the definition of the latter ontology – comprising terms - although it does not help to understand the former ontology – the classification system.

These endeavours are to some extent verified by work of certain knowledge representation exponents. There is a direct association that we can make between the ISO definition of term, and the description Sowa provides for an ontology: “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 for the purposes of talking about D” (Sowa 2000: 492). A ”catalog” of terms provides evidence of things that exist in a specific subject field, which is variously synonymous with a domain, and perhaps we can draw a direct association between the ISO “verbal designation” and Sowa’s language L. To further state this link between terminology and ontology, Sowa 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**” (emphasis added) (2000: 497). Sowa’s view of ontology seems to provide a link between the EB definition and Gruber’s commonly cited “explicit specification of a conceptualization” (Gruber 1997).

For some, “ontology” as a *thing* itself covers a variety of resources. Lassila and McGuinness have presented an ontology spectrum that presents various levels of formalization (2001). Along this spectrum are:

- **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

Based on Lassila and McGuinness, 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.

Where a thesaurus is considered an ontology for these authors, Oard regards the thesaurus as use-oriented: an ontology which is “specialized to information retrieval” (1997). He suggests that ontological considerations help in the “knowledge based retrieval” of free text. Thesauri have been used for query expansion to overcome limitations in simple keyword-based retrieval due to variability of human indexers and user keywords (Efthimiadis 1996). There is some confusion in the literature as to whether or not general purpose lexical networks, like WordNet and EuroWordnet are ontologies. Oard sides with those who do whilst Sowa regards WordNet as a “terminological ontology” (2000). Maedche is fairly agnostic: he provides a formal means by which a resource such as Wordnet could be used as an ontology (2002). Some authors are unsure about the role of WordNet in the analysis of technical texts: WordNet is “overly general” and unlikely to contain domain specific terminology (Faure and Nedellec 1999). We tend to agree, and both Sowa’s definition incorporating “domain”, and the standards-based definition of terms being in a “specific subject field” provide additional strength to this position.

Maedche’s work on ontology “learning” suggests the potential for mapping between terminology and ontology via an ontology structure (Maedche 2002). If such a mapping between terminology and ontology can be made, large-scale validated terminology collections may be of value to ontology developers. Existing standardized collections of terminology developed in accordance with international standard ISO 704 would provide a ready-to-use peer-agreed resource for such activities.

This notion of ontology learning is interesting. Typically, approaches to ontology “learning”, and here consideration is constrained to extraction from free text, are based on syntactic parsing (Maedche and Volz 2001, Maedche and Staab 2003, Faure and Nédellec 1998, Mikheev and Finch 1995). Some authors augment such an approach using TFIDF[1], word clustering using simple stemming, and coded linguistic relationships (Maedche and Staab 2003). Some authors have noted the significance of “term

1.1.

[1] Referred to as *domain consensus* by yet other authors (Navigli, Velardi and Gangemi 2003)

inclusion" and reportedly use term banks to identify term inclusion (Mikheev and Finch 1995). Other authors use WordNet to discover semantic relations between terms, although they are unconvincing with regard to term identification (Navigli, Velardi and Gangemi 2003). Papers on automatic extraction of terms and, more ambitiously of ontology (ontology learning), and on information extraction, open by enumerating that there are three different types of techniques that can be used for the respective enterprises: (i) statistical (Salton 1971, Jing and Croft 1994), (ii) linguistic (Grefenstette 1994) and (iii) hybrid techniques. This division is somewhat artificial in that the first two techniques are interdependent: What is being collected and arranged are linguistic units – words, phrases (statistics) and study of these units (linguistics) is based largely on the frequency counts of these units. The hybrid techniques include sophisticated classification systems (neural networks and other learning algorithms) but these techniques still rely of the frequency counts of linguistic units. We will avoid drawing these artificial boundaries as statistical and linguistic information are not only interdependent but represent different facets of how information/knowledge is communicated in language.

Encouraged by Sowa and Maedche, a method is proposed for automatic derivation of ontologies from collections of specialist text, informed by work in terminology science and using mechanisms for extracting and organizing terms from text corpora. If we describe an approach to terminology extraction, informed by recent developments in international standards for terminology that can be used to seed such terminology collections, and map between terminology and ontology via an ontology structure, there is a potential reduction in the burden of terminology and ontology acquisition.

3. A method for extracting concept structures

A three-part method for understanding the ontology of a domain is proposed based on a study of collocations and lexical semantic relationships between terms in a specialist text corpus. The three parts comprise analysis of a) word use patterns; b) collocation patterns; and c) linguistic patterns.

The first part is inspired by studies of texts of specialist domains, sometimes pejoratively called technical texts, where parsimony of thought and expression is used to reduce ambiguity in communication and to give authority to the writers of the specialist texts and to the domain community. In particular, specialists make good use of word formation devices available to them in their natural languages; the productive manner in which a small set of words is used to construct vocabulary, through inflection, derivation, blending and compounding for example, to describe a specialist activity is for us a productive use of the morphology of a language system. There are key differences in word use in specialist texts when compared with texts of everyday use and these differences can be used to reduce subjectivity in term, concept and ontology extraction.

The second and third parts relate to the practical problems of extracting patterns that provide evidence of the formation of compound phrases, both collocational and linguistic. The first of these is that of Frank Smadja for retrieving collocations from texts (1993). The second method is that of Marti Hearst for seeking to automatically acquire hyponyms from large text corpora (1992). Both methods (Smadja and Hearst) rely on manual selection of an important term for further consideration. We have embellished both methods by reducing the reliance on the human being: given a corpus of texts, our system bootstraps itself by automatically selecting words/terms that have the potential to form lexical hierarchies and networks. The candidate compound terms, typically compound nouns, extracted using Smadja's method have at least one of the more frequent terms, usually acting as a head of the compound. Hearst's method can use such profusely occurring terms and identify lexical semantic relations that these terms have with one another: the elegance of Hearst's work, perhaps based on the original observations of Cruse (1986), is that both the hyponyms and hypernyms can either be single word or multiword term.

From this analysis, we produce "trees" showing collocational extensions and lexical semantic relationships. The trees obtained from Smadja's method and from Hearst's method can be *joined* and *simplified* (see Sowa 1984 for example) to obtain a much richer picture of the inter-relationships between terms, and, by implication, between concepts. The unified tree can then be exported to one of the many knowledge representation and reasoning systems for use in the development of knowledge bases. Within these knowledge representation systems, the trees can be further refined.

The unified tree is our account of the ontological commitment of the specialists whose texts we have examined: we make a small leap to suggesting that this tree is a candidate ontology. We have used this method for discerning the ontological commitment in domains as diverse as breast cancer (Tariq et al 2003), financial trading (Gillam 2002)), health policy communication (Gillam and Ahmad 2002), forensic science (Ahmad et al 2003) and, in this paper, a specialization of the multi-disciplinary subject, nanotechnology.

In the remainder of this section, we discuss the three parts of the method, and this section concludes with an algorithm that combines the parts.

3.1. Word use patterns

Schröder has argued that any language system includes an open-ended sequence of sublanguages. Most sublanguages are special languages, which belong to a definite subject field. Any special language "represents the totality of linguistic means used in a limited sphere of communication on a restricted subject in order to enable cognitive work to be done and mutual information to be conveyed by those acting in the said domain" [and is distinguished] from sociolects, [ ] defined as sublanguages of social and/or professional groups […..in that] special languages are always *functional languages* and belong to a subject field' rather than to a 'certain group' (Schröder 1991: 4-5).

Special languages deal with a range of named or designated *entities*: objects, events, acts, processes, abstractions and generalizations, to name but a few. These entities may have different qualities and quantities, may behave differently, and the behaviour may be further sub-classified. Special languages vocabulary comprises nouns, adjectives, (full-) verbs, and adverbs; these are sometimes referred to as words of the open classes; classes whose stock is constantly changing. The closed class stock of a language system includes conjunctions, prepositions, determiners, and verbs-to-be.

The comparison of relative frequencies of words between a general language corpus, for example the 100 million word British National Corpus (Aston and Burnard 1998) and a special language corpus indicates the *variance* in the use of the words in general language and their term-equivalents in special languages. 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).

Zipf's law, while arguably limited in its predictive capability, provides a common independent reference to which we can make a systematic comparison of the differences between analysis of specialist texts (corpora) and general language. It has been shown that rank multiplied by relative frequency produces a constant around 0.1 (Li 1992, Yavuz 1974), or to put it another way, 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 Schütze 1999, pp26-27). Zipf's law, and consistent deviation from the expected values it provides both a means to quantify differences between special language and general language, where deviations from this law may be a signature of text. Furthermore, this tends to justify making comparisons between the two.

For this analysis, we are concerned with distribution of frequency, and the comparison of a specialist corpus with a general language corpus. We use BNC as the general language corpus, and consider frequency and weirdness values as discussed elsewhere (e.g. Gillam and Ahmad 2002). High frequency provides the simplest measure of use; weirdness values provide a means to determine how "interesting" a given word is in contrast with the general language. Since function words tend to be consistently used across corpora, high frequencies are tempered by low values for weirdness.

The difficulty with existing descriptions of weirdness is that it provides a singularity when words do not occur in the general language corpus: this results in as "infinite" value. To overcome this, we have modified the description of weirdness to incorporate a simple smoothing technique - "Add-one" (Gale and Church 1990) - to adjust frequency according to a renormalization factor. Other smoothing techniques have been suggested in the literature, principally for predictions of bigrams (e.g. Manning and Schütze 1999: 196 et seq; Chen and Goodman 1996; Gale 1995; Gale and Church 1990). The renormalization factor for the BNC can be reducing to adding one to the value of the frequency of occurrence in BNC (due to the contrast between the number of tokens and the number of types). 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. 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 promising. This modified value for weirdness is calculated as follows:

$$\tau(w) = \frac{\mathrm{N}_{\mathrm{GL}} f_{SL}}{(1 + f_{GL})\mathrm{N}_{\mathrm{SL}}}$$

Where: f_{SL} = frequency of word in specialist language corpus
f_{GL} = frequency of word in general language corpus
N_{SL} = total count of words in specialist language corpus
N_{GL} = total count of words in general language corpus

We can now produce numeric distributions for both frequency and weirdness values. We can use these distributions to suggest a list of words of further interest based on a combination of high-frequency (correlating with acceptability following Quirk) and high-weirdness. To further remove subjectivity, we use a common statistical measure of significance, z-score, to appraise and combine these such that words are automatically selected.

3.2. Collocation patterns

Collocations are 'recurrent combinations of words that co-occur more often than chance and that correspond to arbitrary word usages.' (Smadja 1993: 143). 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. His collocation method analyses a neighbourhood of five words preceding and following a *nucleate*. The frequency of occurrence 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 identified as a significant collocation pattern using a u-score familiar to statisticians as a measurement of variance. Two further values are suggested, both z-scores that may be of use in determining significance of these collocations. He also suggests that the same mechanism could be applied to producing larger collocations, but provides a caveat that the results tend to fail when frequencies of lower than 50 are considered. Subsequently, syntactic information is added to the constituents of the collocates using a statistical tagger: we do not consider this here. Smadja suggests that a u-score of greater than 10, and a z-score value greater than 1 can be used as thresholds for selecting collocating words. He identifies various patterns of collocations within Associated Press (AP) news-wire texts.

Smadja's method initially relies on the (manual) selection of a specific word, upon which analysis is to be carried out. Criteria for selecting this word are not clear beyond lexicographical "interest". The automatic weirdness-frequency selection mechanism, suggested above, may provide useful input to this method. In further modifications to this method, while we use the given calculations for producing interesting values, at each iteration we only use a collocation if the value one word from the nucleate satisfies the thresholds. While this approach is valid for processing English texts, it may need to be reconsidered for collections in other languages.

Since the thresholds for collocation extractions are based on frequencies, at lower frequencies they may cease to provide multiword term candidates of greater length. Smadja notes that 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). Our analysis on specialist corpora tends to make use of corpora of around 100,000 to 1,000,000 tokens, so low frequencies of occurrence are more likely in these corpora: it may not be possible to collect more than this amount for an emerging specialism.

Our investigations of this algorithm lead us to believe that collocations with frequencies much less than 50 can be considered to be interesting: the likelihood that seven non-function words occur in an unbroken sequence within any text is quite small, so any supporting frequency information is important.

From the list of frequent-weird words, we produce a tree comprising a root node with the leaves representing each word. For each word, we produce the collocations that satisfy the thresholds, and where our conditions for position and Smadja's thresholds are met, the leaf word becomes a node to which valid collocations are attached. In this analysis, we remove words from consideration that we think do not form interesting collocations. For this, the 2000 most frequent words of the British National Corpus is a useful set. The contexts within which the collocations occur become the input 'corpus', and further collocation patterns are analyzed within these contexts. Each further collocation iteration extends this tree, until no further valid collocations can be produced.

3.3. Linguistic patterns

The notions of conceptual schemes, thesaural frameworks, and ontology are articulated not only through the deliberate and frequent use of words related to key concepts, or *terms* denoting concepts in the Platonist sense, but through a set of lexical semantic relationships involving the terms. For example, terms in a domain are often related to each other through a range of semantic relations such as *hyponymy* and *meronomy*. These semantic relations are often exemplified in a language through the arrangement of certain terms in recurrent grammatical patterns that can be subsequently analyzed. In this context, Cruse has discussed the notion of *diagnostic frames*: a triplet of phrases - *X* REL *Y* where X and Y are noun phrases (NPs) and REL is a phrase generally expressed as *IS A, IS A TYPE OF/KIND OF* and *PART OF* for illustrating hyponymic and meronymic relationships respectively.

To understand the ontology of a domain we suggest that hyponymic and other semantic relationships like meronymy should be examined. Hearst has outlined a method for "automatic acquisition of hyponyms from large text corpora [e.g. two different text types, the 8.6 million word *Grolier's American Encyclopaedic* and 20million word *New York Times* corpus]" some 10 years ago. Hearst extracted words (noun-phrases) that had hyponymic relationship with each other through linguistic patterns, and used her method to "critique" the structure of a "large hand-built thesaurus". Hearst has suggested that hyponymic relationships are often marked by phrases like *such as*, *including*, that relate a hyponym and its hyperonym (*injury* including *broken bone*, *the bow lute*, such as *the Bambara ndang*). She has suggested the following patterns where such a marked-up relationship can be found:

NO.	POTENTIAL 'HYPONYMIC' PATTERNS
1	NP_0 **SUCH AS** { NP_1, NP_2, ,…………….(**AND|OR**) NP_n}
2	**SUCH** NP_0 **AS** { NP_1, NP_2, ,…………….(**AND|OR**) NP_n}
3	{ NP_1, NP_2, ,……………., NP_n} (**AND|OR**) **OTHER** NP_0
4	NP_0 (**INCLUDING|ESPECIALLY**) { NP_1, NP_2, ,.(**AND|OR**) NP_n}

Figure 2. Pattern indicating that NP_i ($i \neq 0$)is a hyponym of NP_0. These lexico-syntactic patterns indicate hyponymy relation and satisfy the *desiderata* that the patterns: occur frequently and in many text genres; (almost) always indicate the relation of interest; can be recognized with little or no pre-coded knowledge

3.4. Combined algorithm

The three parts to the method can be combined into the following algorithm:

STEP	**TASK**
Setup	Select a corpus of specialist texts in an arbitrary domain and a general language frequency list (BNC).
1	**Patterns of word use**
Input	Select a value for z-score (e.g. 1)
1.1	Tokenize the corpus and collect frequency information for each word
1.2	Compute “smoothed” weirdness values using general language frequency list
1.3	Reject words where z-score of frequency less than chosen z-score AND z-score of smoothed weirdness less than chosen z-score
Output	List of “interesting” (domain specific) words
2	**Patterns of collocation**
Input	Tree-structured version of “interesting” words resulting from above. Values for u-score and z-score
2.1	Augment tree by taking collocations at positions –5 to +5, ignoring collocates in the list of the 2000 most frequent words in BNC.
2.2	Remove leaves not satisfying Smadja’s thresholds AND not at positions –1 / +1. Contextual information is associated to these collocations
2.3	Using collocation pattern and its contexts, repeat from 2.1 until tree can no longer be extended.
Output	Tree of collocating (candidate) terms
3	**Linguistic Patterns**
Input	A set of linguistic patterns *REL*;
3.1	Find terms from the collocation tree resulting from the previous step that satisfy each *REL*; produce a new tree containing hypernymic relationships where both terms come from this tree.
3.2	Extend the tree from 3.1 where only 1 known term is involved in the relation
3.3	Extend the tree from 3.2 for “unknown” terms
Output	Tree of lexically related terms with varying degrees of confidence associated to the terms identified
4	**Unification**
Input	Resulting trees from steps 2 and 3
4.1	Merge trees resulting from steps 2 and 3 for expert refinement and pruning
Output	A candidate conceptual structure (tree)

The algorithm above provides an overview of the proposed method. It is interesting to note, although we have yet to explore this, that the algorithm can bootstrap itself if we converge steps 2 and 3 more closely. For step 3.2, if we identify a new term (candidate) involved in the relation not deemed important in the collocation phase, we can perhaps identify its collocational structure through further analysis (determining the most highly-frequent, highly-weird component word) and develop further collocations through this. Subsequently, we can look for additional linguistic patterns involving this new sub-tree. The algorithm concludes once we have exhausted the set of patterns. The approach is interesting since it combines frequent “interesting” words with what are likely to be infrequently occurring linguistic patterns as a means of mutual validation. This combination is the subject of further work.

4. Case Study: Nanotechnology

A note on nanotechnology and a 1 Million word corpus on the topic: Nanotechnology is an emerging and highly controversial 21st century area of research. Nanoscale devices take advantage of the characteristics of materials at the atomic level. Proponents of nanoscale devices look to a future of ultrasmall supercomputers, and cancer-tagging molecules. One particular aspect of nanotechnology is the nanotube. A nanotube is a tubular molecule composed of carbon or of atoms other than carbon. Discovered in 1991, carbon nanotubes are just a few billionths of a metre across. The nanotube is a cylindrical molecule, usually of Carbon, each formed of sheets of atoms arranged helically. The arrangement and the properties of the constituent atoms have been observed in laboratory conditions to have revolutionary physical properties – the measured strength of the artefacts made of nanotubes exceeds that of any other currently known, despite their size. Nanotubes are superconductive and can switch from one state to another at speeds currently not known.

The fact that this new ‘wonder’ can be made from an element as abundant as Carbon has led to a significant interest and intellectual and financial investment in nanotubes research. This interest has led a large number of journal papers, popular science articles, national government position and policy papers, and newspaper reportage on the subject and potential of (carbon) nanotubes. One journal, *Applied Physics Letters*, has carried over a million words in some 400 articles in just a few years. Prestigious journals including *Nature*, *Science*, *Philosophical Transactions of the Royal Society* carry learned papers and commentary in this subject area.

This interdisciplinary field is rapidly evolving and having incorporated the ontological commitments of the key contributors, physicists and chemists, nanotechnology researchers are developing their own ontological commitment. New lexical hierarchies and part-whole networks of terms will evolve so as to articulate the synthesized ontological commitment. Extraction of these hierarchies and networks will help in understanding and visualizing the implicit ontology.

A corpus of 1,012,096 words was analyzed. This corpus contains 404 learned articles from the *Applied Physics Letters* section on *Nanoscale Science and Design* (average of about 2500 words per article). The corpus comprises 26861 words (types). The representative general corpus used was the British National Corpus.

4.1. Word use patterns: Nanotechnology

A total of 10231 words were found to have infinite weirdness – 38% of these words do not occur in the BNC; these include typographical errors as well as neologisms and scientific and technical words related to nano-science and technology. Table 1 below shows a selection of words in our corpus, some of which are not in the BNC or occur in BNC with a low frequency, and have a high frequency of distribution in this specialist corpus. For example the lemma *nanotubes* occurs 2348 times (1379 times for the plural *nanotubes* and 969 for the singular *nanotube*) with a relative frequency of 0.0023%: in our corpus we will find the lemma *nanotube* about 23 times for every 10000 words, including the closed class words. Both *nanoparticles* and *nanowires*, highly specialized forms of the nanotube, have infinite weirdness but lower frequency distribution (829 and 619 tokens respectively) than the 'parent' nanotube.

Table 1. A selection of nano-based terms in our corpus with weirdness values

Term	Frequency	Weirdness	BNC frequency	Smoothed weirdness
nanowires	619	INF	0	61225
nanoparticles	829	81996.07	1	40998
nanowire	360	INF	0	35607
nanotube	969	47921.71	2	31948
nanoscale	268	INF	0	26508
nanoparticle	232	INF	0	22947
nanotubes	1379	27279.27	5	22733
nanostructures	212	INF	0	20969
nanorods	159	INF	0	15727
nanocrystals	395	19534.65	2	13023

A z-score value of 1 for both frequency and weirdness produces a list of 19 words which contains nanotubes, nanoparticles and nanotube as candidates for subsequent analysis. Changing this value for z-score alters the number of words selected by this combination. This relatively straightforward mechanism allows us to systematically vary the amount of words that we consider. Table 2 below shows the number of words selected by different values for z-score.

Table 2. Number of words produced by combined z-score values for both frequency and weirdness

	Nanoscale
Tokens	*1012096*
Types	*26861*
z-score	
5.0	1
4.0	5
3.0	6
2.0	8
1.0	19
0.9	21
0.8	24
0.7	27
0.6	33
0.5	39
0.4	45
0.3	62
0.2	79
0.1	129
0	352

4.2. Collocation patterns: Nanotechnology

Given that the term *nanotubes* has high frequency and high weirdness, with a z-score of 1 for frequency and weirdness, our system selects this as one of the words for consideration in collocation patterns. We then compute frequencies of collocates of this term with words in a sentence neighbourhood of 5. Recall that the term *nanotubes* has a frequency of 1378, *nanotubes* collocates with 1811 different words within this neighbourhood. One of these collocations, which occurs 690 times across the neighbourhood, is with *carbon*, 647 of these are at position –1, which would produce the phrase *carbon nanotubes*. Applying Smadja's thresholds, the number of statistically significant collocates reduces from 1811 to 22 (98.8% of collocates ignored). Our constraint with regard to positions +1 and –1 reduces this list further to consideration of *carbon nanotubes*, *z nanotubes*, *nanotubes cnts* and *nanotubes grown*, a list of only 4 (see Table 3 for details).

Table 3. Example collocations of nanotubes

Word	**Collocate**	**-5**	**-4**	**-3**	**-2**	**-1**	**1**	**2**	**3**	**4**	**5**	**u-score**
nanotubes	carbon	3	7	0	6	***647***	***0***	2	9	8	8	37131
nanotubes	single-walled	0	1	2	72	***7***	***0***	0	1	0	0	455
nanotubes	aligned	1	1	6	51	***25***	***1***	3	2	5	2	237
nanotubes	multiwalled	0	1	0	46	***9***	***0***	1	5	0	0	184
nanotubes	cnts	2	0	0	0	***0***	***35***	0	0	0	0	109
nanotubes	multiwall	1	0	1	30	***6***	***0***	2	0	0	0	78
nanotubes	emission	11	24	13	4	***0***	***0***	0	9	8	0	55
nanotubes	z	0	2	0	0	***24***	***2***	0	0	0	0	51
nanotubes	single-wall	1	2	0	24	***5***	***0***	0	0	1	0	50
nanotubes	mwnts	0	0	0	0	***0***	***18***	1	0	0	1	29
nanotubes	vertically	0	0	15	3	***0***	***0***	1	2	1	0	19

Having established a statistically significant collocate such as *carbon nanotubes*, we take the contexts that it appears in as an input 'corpus', and seek collocations with this phrase to find three-word expressions.

Smadja's algorithm to extract three word collocates that included *carbon nanotubes* (with a frequency of 646). By continuing to apply this mechanism, we can derive further extensions for the compounding as shown in Table 4.

Table 4. Extended collocations of nanotubes

Phrase					**Frequency**
	aligned	*carbon*	*nanotubes*		48
vertically	aligned	*carbon*	*nanotubes*		15
	aligned	*carbon*	*nanotubes*	kai	4
	multiwalled	*carbon*	*nanotubes*		46
	multiwalled	*carbon*	*nanotubes*	mwnts	13
	single-wall	*carbon*	*nanotubes*		24
	single-wall	*carbon*	*nanotubes*	swnts	4

Such collocations we can quickly appraise. The *mwnts* and *swnts*, while potentially important collocations, shown the association of a candidate term to its abbreviated form. We have not considered such devices in this work as yet. On investigation *kai* appears to be the name of an author. There is an interesting variation between *single-wall* and *multiwalled* carbon nanotubes: perhaps, some tension between *wall* and *walled* within this collection.

Assuming that *nanotubes* is the linguistic head of the 2-, 3- and 4- word collocations, our program generates a hierarchical tree (Figure 1). This tree helps not only in visualizing the relationship between the various terms but can be exported to a knowledge representation system, for instance Stanford University's Protégé system, to be stored as a semantic network. This network can be used in a reasoning system.

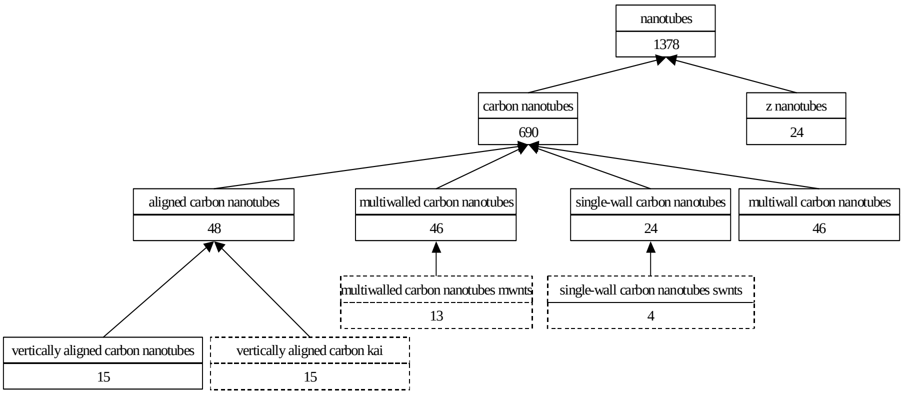
nanotubes
1378
carbon nanotubes
690
z nanotubes
24
aligned carbon nanotubes
48
multiwalled carbon nanotubes
46
single-wall carbon nanotubes
24
multiwall carbon nanotubes
46
multiwalled carbon nanotubes mwnts
13
single-wall carbon nanotubes swnts
4
vertically aligned carbon nanotubes
15
vertically aligned carbon kai
15

Figure 1: Tree representation of the candidate compound terms extracted from the 1 million word corpus using this method

4.3. Linguistic patterns: Nanotechnology

Sentences containing linguistic patterns comprising relationships between key words *X* and *Y* in a semantic relationship are extracted automatically from a corpus of free texts with the help of the set of cues. Firstly, we can use the resulting collocations from the step above as key words *X* and *Y* to discover patterns between these collocations.

Subsequently, we can look for phrases that comprise (a) an adjective plus preposition (for example, *SUCH AS*), or (b) an adjectival pronominal (for instance of *AND OTHER*, and *OR OTHER*). To establish whether the relational cue belongs to the grammatical categories in (a) and (b), the sentences containing the frame (*X* REL *Y*) can be tagged using a statistical tagger, e.g. MXPOST (Ratnaparkhi 1996). Regular expressions are then used to detect the tagged sentences that follow the required pattern. The correct sentences are parsed to extract the hypernym-hyponym pairs.

722 sentences were extracted using a set of 8 cues, out of which 55% embodied a domain-related hyponymic relationship. Out of all the cues, *such as* was the most productive, being used in 66% of the valid sentences. Below we list some example sentences illustrating the use of the cues: *such as, and other, including* and *like*.

1. This method has been successfully applied in recent years in the synthesis of various metal nanostructures *such as* nanowires, nanorods, and nanoparticles.
2. Occasional multiwall carbon nanotubes *and other* carbon nanostructures were also found following annealing at higher (> °C) temperatures.
3. The present method will be extended to find and fix nanoparticles *including* polymers, colloids, micelles, and hopefully biological molecules/tissues in solution.
4. This technique is promising because many different types of nanowires, *like* nanotubes or semiconductor nanowires, are now synthetically available

From the sentences above, various terms can be linked together based on the hyponymic relationship, for example (the arrow indicates subtype → supertype):

1. [nanotube], [semiconductor nanowire] → [nanowire] → [metal nanostructure] (sentences 1, 4)
2. [micelle], [polymer], [colloid] → [nanoparticle] → [metal nanostructure] (sentences 1, 3)

Sentences such as 2 and 4 above may confirm a synonymy relationship between *multiwall carbon nanotubes* and *multiwalled carbon nanotubes.*

Again, we can construct a tree from this information such as that presented in Figure 2 below.

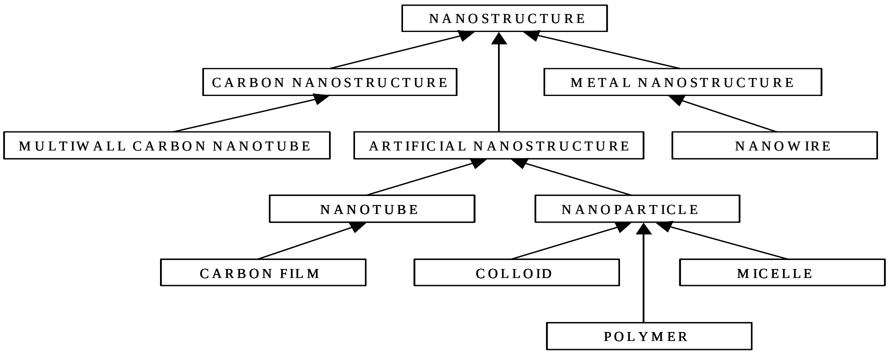

Figure 2: Tree representation of a candidate partial hierarchy – sub-tree of nanostructure

The above tree, derived from the candidate hyponyms was shown to an expert in nanotechnology who provided an initial validation of the hyponymic relationships shown above.

4.4. Unification

The result of combining the statistical and linguistic methods produces a “tree” of terms and relations, organised hierarchically (principally term inclusion).

The partial graphs above, produced by consideration first of collocations and secondly of linguistic patterns, can be merged. For example, collocates of *nanowire* can be linked to the *nanowire* node of the sub-graph

[fe nanowire array], [thicker nanowire array], [thin nanowire array] **is_a** [nanowire array] **is_a** [nanowire] **is_a** [metal nanostructure]

[amorphous boron nanowire] **is_a** [boron nanowire] **is_a** [nanowire] **is_a** [metal nanostructure]

Such graphs need to be appraised for characertistics at the same level, for example the *boron nanowire* as a type of *nanowire* versus *nanowire array* as a structure of *nanowire*.

These graphs could similarly be expanded for other extracted relations and collocations, for example those of *nanoparticle* and other subtypes of *metal nanostructure*.

4.5. A nanotechnology terminology / ontology

The above operation of joining and simplifying, which can be performed by knowledge representation and reasoning systems routinely, allows one to construct larger hierarchies and splice hierarchies. The two operations performed on networks that relate the (values of) attributes of, say, a superordinate object will allow the values of the attributes to be inherited (automatically) by the subordinates and instances; indeed a subordinate may have more than one parent and one can see cases where multiple inheritance is operative.

The use of recently developed and developing terminology standards ISO 12620 for “Data Categories” and ISO 16642 for a “Terminological Markup Framework” enables us to produce a terminology markup language (TML) that represents these results.

ISO 16642 presents a metamodel for terminology comprising a number of “containers” by which terminology resources are generally composed. The terminology metamodel describes a terminology data collection (TDC) that contains one or more term entries (TE), containing one or more language sections (LS), which has one or more term sections (TS). Global information (GI), complementary information (CI) and term component sections (TCS) may also occur.

Each of the containers from ISO 16642 can have (terminological) data categories (DC) included, which are defined in ISO 12620, including the “term” DC itself and a variety of term-related, administrative and relational (e.g. *superordinate concept* and *subordinate concept*) DCs. The combination of the metamodel and a data category selection (DCS) provides the basis by which interoperability between terminology formats can be determined, and can be represented using feature-structure representations (forthcoming ISO 24610). These are combined with the notions of *style* and *vocabulary* from ISO 16642 to produce an XML-conformant encoding such as MARTIF (ISO 12200), or the TermBase eXchange (TBX) format developed by the Localisation Industry Standards Association (LISA). The combination of these standards with the extraction method can provide the basis for a terminology collection. Once terminology systems conform to these standards for import/export, production of candidate terminology collections can be produced by this method and directly used in such systems.

Instead of a terminology system however, we consider how to use such trees within a knowledge representation system that can read files encoded in the supertype/subtype relationship-based Resource Description Framework Schema (RDFS). Here, we map each term to the content of an *rdfs:label*, and suitable concept identifiers (*rdf:ID*) are used to present the classing and subclassing (*rdfs:Class*, *rdfs:subClass*). In this conversion, there is a degree of information loss, since RDFS does not cater for much of the information needed for a terminology format, and it is not expressive enough to cater for natural languages, however the mapping to an ontology language shows the ability to directly populate such an ontology system. Ontology editing applications that understand RDFS, including Protégé and OilEd, can use such output to seed their ontologies for further development.

The mapping of our automatically extracted results to RDFS enables us to produce a candidate ontology that can be edited (pruned, adapted and so on) and visualised within Protégé (Figure 3). Issues of multiple inheritance, use of synonyms and abbreviations, can then be handled within the ontology editor by a domain expert and used to develop knowledge-based systems. Results of this method are still quite preliminary, and require (human) evaluation and determination of the appropriate parameters for term extraction, but show early promise.

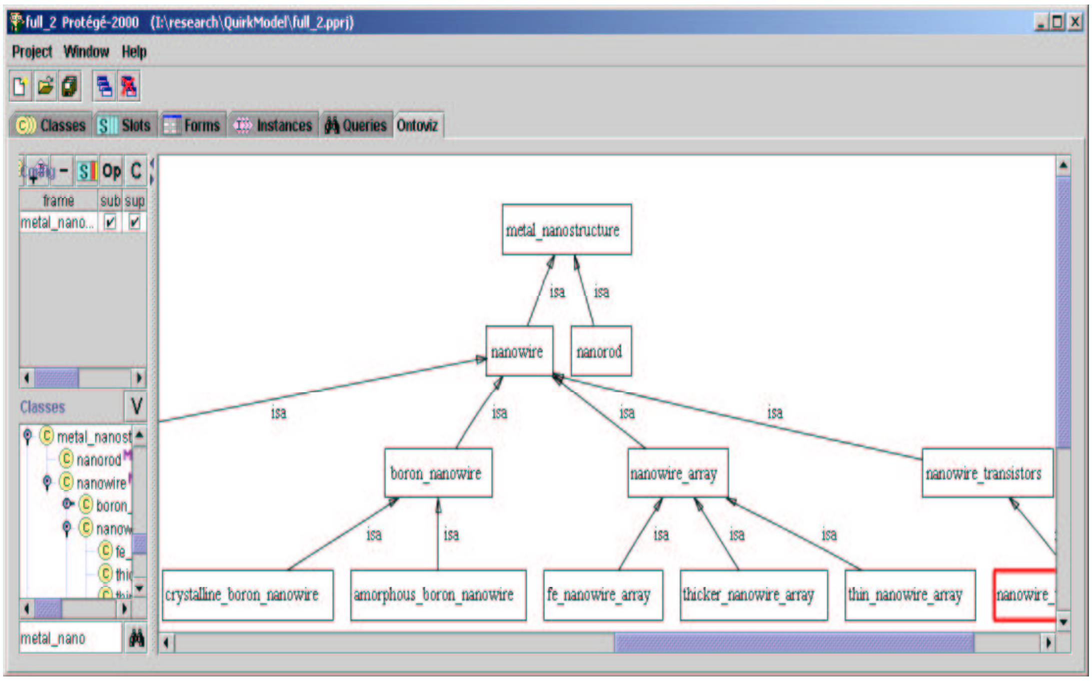

Figure 3: Screen shot of the Protégé Ontology Editor displaying a section of the automatically constructed Carbon Nanotube candidate ontology in RDFS format.

5. Afterword

We have described methods and techniques for constructing networks of terms, extracted from text corpora, which show knowledge in a subject domain is organised. Initial discussions with domain experts have validated the first results being produced with some degree of confidence. We have also shown how candidate terms can be extracted automatically, including collocations, and how other semantic relationships between terms can be extracted. Once collocational and semantic relational networks are produced, one use these networks for tasks such as query expansion: traversing from superordinate to subordinate – focusing a query- and vice versa – broadening a query.

This paper has discussed the automatic generation of a domain ontology by consideration of approaches to terminology extraction and management including current and forthcoming international standards. Many other approaches to generating a domain ontology (including KAON, KAW, ASIUM, Corporum, LTG Text Processing Workbench, Mo'K Workbench, SVETLAN', Prométhée, Caméléon, SOAT) appear to be dependent *ab initio* upon the effectiveness of either the part-of-speech tagger, and therefore confidence in the results will depend in large part on how well-trained the tagger is, and/or on existing resources such as phrase patterns or lexical databases (e.g. Wordnet). Arguably, these approaches can only be as effective as the prior linguistic knowledge they embody: the correctness of the tagger, the coverage of the patterns and the quality of the database will all affect the results obtained. An initial statistical approach will be biased towards the make-up of the text corpus, but will obtain consistent results across different collections: at worst, it will be consistently wrong. Augmentation with phrase patterns acts as a means to expand, or potentially contract, such a result, and may also confirm the result from the statistical phase. Critics may argue that the KAON workbench provides such functionality already (Text-to-Onto), however KAON (v1.2) relies on *a priori* knowledge or expectation (e.g. minimum term frequency to retain) and creation of the concept tree is left to the user. The statistical component of our method is parameterised, so absolute frequency values are not necessary: an order of magnitude above standard deviation provides for value independence. The result of the statistical phase suggests an initial concept tree, which is augmented by a linguistic phase, and we are considering how to make this adaptive, i.e. self-bootstrapping. The resulting skeleton terminology/ontology can subsequently be modified within an appropriate environment. Our approach attempts to avoid use of POS information as far as possible, in the hope that domain adaptation will be easier to manage, and the approach may more readily adapt to other languages. Our system also allows the user to confirm the extracted ontology by referring back to the indexed text. A criticism from a terminological perspective of KAON is that it does not allow this: subsequent effort is necessary.

The joining and simplification operations help to extend the collocational (and semantic relational) networks, involving common terms in each of the network, and the joining of the two networks, especially if drawn from two different papers in a corpus, will help in visualizing the domain knowledge in a way which is only the preserve of the experts.

In addition to the use of text corpora one might explore the use of extant concept-oriented, standardised terminology data bases in conjunction with a text corpus. The collocational patterns and conceptual relations within the extant term bases can be used to verify and validate the results of a corpus-based bootstrapping of an ontology of a domain against an existing verified and validated term data base. This work is currently under progress at Surrey.

We concede that, as work-in-progress, the approach presented requires significant evaluation against other approaches, as well as a comparative evaluation across corpora, and also to determine the extent to which the statistical operations are reliable for lower frequency terms of greater length. Though Smadja is skeptical about collocations at lower frequency, we have found terms of upto length 7 at a frequency of 1 through relaxing Smadja's thresholds. Such terms include:

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

and we can argue that the probability that this combination occurred in peer-reviewed text by accident is somewhat low.

The combination of statistical and linguistic methods for extraction, with international standards for terminology and emerging standards for ontology provides a useful baseline for further exploration of our method.

Acknowledgements

The work presented has partially supported by: EPSRC (SOCIS: GR/M89041), EU (SALT: IST-1999-10951 and GIDA: IST-2000-31123), and ESRC (FINGRID: RES-149-25-0028). Certain aspects of the work have been validated in ongoing work of the international standards community (through ISO TC 37).

References

1. Ahmad, K., Tariq, M., Vrusias, B. and Handy, C. 2003. “Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains”. In Sebastiani, F. (ed.): *Proceedings of ECIR'03*. LNCS-2633. Springer Verlag, Heidelberg (2003) 502-510.
2. Ahmad, K and Rogers, M. 2001. “Corpus-based terminology extraction.” In Budin, G., Wright S.A. (eds.): *Handbook of Terminology Management, Vol.2*. John Benjamins Publishers, Amsterdam. 725-760. (2001)
3. Ahmad, K. Pragmatics of Specialist Terms and Terminology Management. In (Ed.) Petra Steffens. *Machine Translation and the Lexicon*. Proceedings of the 3rd Int. EAMT Workshop, Heidelberg, Germany, April 26-28,1993.) Heidelberg (Germany): Springer. pp.51-76. 1995.
4. Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P. and Shadbolt, N. 2003. “Automatic Ontology-Based Knowledge Extraction from Web Documents.*”* *IEEE Intelligent Systems*, 18(1), 14-21.
5. Aston, G. and Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.
6. Buchanan, B. G. and Shortliffe, E. H. 1984. *Rule-Based Expert Systems*. Addison Wesley.
7. Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press, Avon, Great Britain
8. Efthimiadis, E.N. 1996. “Query Expansion”. In: Williams, M.E., (ed.) *Annual Review of Information Systems and Technology* (ARIST),.31, 121-187

9. Faure, D. and Nédellec, C. 1999. “Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM”. *Lecture Notes in Computer Science*, ISSN: 0302-9743, Volume 1621. Springer-Verlag Heidelberg
10. Faure, D. and Nédellec, C. 1998 “ASIUM: Learning subcategorization frames and restrictions of selection”. In Y. Kodratoff, (Ed.) *10th Conference on Machine Learning (ECML 98) -- Workshop on Text Mining*, Chemnitz, Germany.
11. Gillam, L. (Ed) *Terminology and Knowledge Engineering: making money in the financial services industry*. Proceedings of a workshop at the 2002 conference on Terminology and Knowledge Engineering (TKE).
12. Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, USA
13. Gruber, T. 1993. “A translation approach to portable ontologies”. In *Knowledge Acquisition*, 5(2), 199-220
14. Guarino, N., Masolo, C., and Vetere, G. 1999. “ONTOSEEK: Content-Based Access to the Web.” *IEEE Intelligent Systems*, 14(3), 70-80.
15. Hearst, M. 1992 “Automatic Acquisition of Hyponyms from Large Text Corpora”. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.
16. ISO 704. 2000 *Terminology work – Principles and methods*. ISO, Switzerland
17. ISO 1087-1. 2000. *Terminology work -- Vocabulary -- Part 1: Theory and application*. ISO, Switzerland
18. ISO 12620. 1999 *Computer applications in terminology - Data categories*. ISO, Switzerland.
19. ISO 16642. 2003. *Computer applications in terminology – Terminological mark-up framework*. ISO, Switzerland.
20. Jackson, P. 1990 *Introduction to Expert Systems*. Second edition. Addison-Wesley Publishers Ltd.
21. Jing, Y., Croft, W.B. 1994. “An Association Thesaurus for Information Retrieval”. In Bretano, F., Seitz, F. (eds.) *Proceedings of the RIAO’94 Conference*. 146-160. CIS-CASSIS, Paris, France
22. Lenat, D.B. 1995. “Steps to Sharing Knowledge”. In: Mars, N.J.I. (ed.) *Toward Very Large Knowledge Bases*. IOS Press, Amsterdam
23. Lenat, D.B. and Guha, R.V. 1990 *Building large knowledge based systems*. Addison-Wesley, Reading Massachusetts
24. Luger, G.F., and Stubblefield, W.A. 1993 *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Second Edition, London: Addison-Wesley
25. Maedche, A. and Staab, S. 2003 “Ontology Learning”. In S. Staab & R. Studer (eds.) *Handbook on Ontologies in Information Systems*. Springer.
26. Maedche, A., Motik, B., Stojanovic, L., Studer, R. and Volz, R. 2003. “Ontologies for enterprise knowledge management”. *IEEE Intelligent Systems*, 18(2), 26-33.
27. Maedche, A. 2002. *Ontology Learning for the Semantic Web*. The Kluwer International Series in Engineering and Computer Science, Volume 665, ISBN: 0792376560
28. Maedche, A. and Volz, R. 2001 “The Ontology Extraction and Maintenance Framework Text-To-Onto”. Workshop on Integrating Data Mining and Knowledge Management. California, USA
29. Mikheev, A. and Finch, S. 1995 “A Workbench for Acquisition of Ontological Knowledge from Natural Text”. In *Proceedings of the 7th conference of the European Chapter for Computational Linguistics* (EACL'95), 194-201. Dublin, Ireland. 1995.
30. Navigli, R., Velardi, P. and Gangemi, A. 2003 “Ontology Learning and Its Application to Automated Terminology Translation”. *IEEE Intelligent Systems* 18(1), 22-31.
31. Noy, N.F. and Hafner, C.D. 1997 “The State of the Art in Ontology Design: A Survey and Comparative Review” in *AAAI*, Fall 1997, 53-74.
32. Oard, D.W. 1997 “Alternative approaches for cross-language text retrieval”. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence.
33. Ratnaparkhi, A. 1996. “A Maximum Entropy Part-Of-Speech Tagger”. In *Proceedings of the Empirical Methods in Natural Language Processing Conference* 133-141.
34. Russell, S. and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ
35. Salton, G. 1971. “Experiments in Automatic Thesauri Construction for Information Retrieval”. In *Proceedings of the IFIP Congress*, Vol. TA-2, 43-49. Ljubljana, Yoguslavia.

36. Schröder, H. 1991. "Linguistic and Text Theoretical Research on Languages for Special Purposes. A thematic and bibliographical guide". In H. Schröder (Ed.) *Subject-oriented Texts: Languages for Special Purposes and Text Theory*, 1-48. Berlin & New York: Walter de Gruyter & Co..
37. Smadja, F. 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics*, 19(1), 143-178. Oxford University Press.
38. Sowa, J.F. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA.
39. Sowa, J.F. 1984. *Conceptual Structures: Information Processing in Mind and Machine* Addison-Wesley
40. Tariq, M., Manumaisupat, P., Al-Sayed, R. and Ahmad, K. 2003. "Experiments in Ontology Construction from Specialist Texts". *Proceedings of EUROLAN Workshop: Ontologies and Information Extractio*n, Bucharest, Romania.
41. Zipf, G.K. 1949 *Human Behavior and the Principle of Least Effort*. Hafner, New York.