Specialist knowledge and its management

K. Ahmad and L. A. Miles

ABSTRACT

A method for systematically analysing text is outlined for use in the acquisition of specialist knowledge. Such knowledge can typically be engineered in the knowledge bases of hydroinformatic systems. A synthesis of work in the corpus-based studies of language and in the literature on Language for Special Purposes is presented. This synthesis forms the basis of semi-automatic analysis of texts for extracting terms, elaborating terms and identifying heuristics.

Key words | knowledge acquisition, knowledge management, language engineering, stakeholder participation, urban wastewater management

K. Ahmad (corresponding author)
Artificial Intelligence Group,
Department of Computing,
School of Electronics,
Computing and Mathematics,
University of Surrey,
Guildford,
Surrey GU2 5XH, UK
Tel.: +44 1483 259322;
Fax: +44 1483 259385;
E-mail: *K.Ahmad@surrey.ac.uk*

L. A. Miles
Artificial Intelligence Group,
Department of Computing,
School of Electronics,
Computing and Mathematics,
University of Surrey,
Guildford,
Surrey GU2 5XH, UK
and
HR Wallingford,
Howbery Park,
Wallingford,
Oxon OX10 8BA, UK

INTRODUCTION

Hydroinformatic systems are being put to many different uses in the management of the aquatic environment. From policy makers through pressure groups to the general public, hydroinformatic systems and their outputs are being used and scrutinised by a diverse set of people with a range of interests and objectives, and with greatly differing levels of understanding. Different stakeholders interpret the data and information related to a given aquatic environment differently. A number of case studies in hydroinformatics, where expert systems were used, show that the knowledge of hydraulic engineers, hydrologists and kindred persons was made explicit and subsequently represented within, and retrieved from, knowledge bases. Hydroinformatic systems of the future will be required to pool the interpretations of different stakeholders.

Knowledge bases are typically developed by *knowledge engineers* who themselves are not usually familiar with the area of application, and this in itself can cause problems or delays. For the effective management of knowledge, and for the much discussed *knowledge society* to come about, there would appear to be a need for supporting *infrastructure*. This infrastructure should include methods and computer programs to support the development of knowledge bases and to assist users in understanding and employing the contents of these knowledge bases.

Philosophers have been discussing the emergence of knowledge and its organisation under the rubric of *epistemology*, *ontology*, and, latterly, *philosophy of science*. These subjects, full disciplines in their own right, broadly deal with how human thought influences human action. Epistemology deals with the structure of knowledge, and here we have discussions on taxonomies, for instance. Ontology deals broadly with the 'essence of being'. Philosophy of science is a specialisation of philosophy itself and here, instead of looking at the whole range of human knowledge, philosophers focus on scientific and technical knowledge. There are heated debates in philosophy about the link between knowledge and language. Essentially, what is being discussed is whether or not I can articulate what I, as a human being,

know and, conversely, is what is being articulated the same as what is known.

Language can be viewed as a symbol system, highly developed and extremely complex, which is used to create, disseminate and censor knowledge. The literature on knowledge management (see, for example, Nonaka & Takeuchi 1995), directly or indirectly, deals with communications within an organisation; this communication involves the use of language, mathematical and scientific symbols, graphics and speech. The increasing availability of *digital libraries*, comprising computer-accessible texts, and *terminology databases*, comprising specialist terms, has meant that information about specialist domains is more readily available. Digital libraries contain a range of textual material from learned journals by individual publishers to national libraries like Librarie de France and the British Library, and from technical manuals to policy documents. Terminology databases contain terms of a specialist domain together with the definition of the terms and their equivalents in other languages; these databases are owned/sponsored by organisations like the WHO and federal governments like the bilingual Canadian federal government.

The ready availability of documents and terminology resources will aid a number of the so-called *knowledge workers* (Drucker 1998) which include researchers, knowledge engineers, and planners working for and on behalf of a variety of stakeholders. The contents of these documents are processed by the knowledge workers to extract facts, rules, and rules about rules (metarules) of a given domain for a particular stakeholder. The knowledge workers use their linguistic abilities to analyse documents and use their knowledge of language to consult terminology resources.

There have been a number of developments in cognitive psychology and in artificial intelligence that have helped researchers in these fields to study the nature and function of expertise and the behaviour of experts. A number of methods and techniques have been developed to acquire experience-based knowledge from experts and this has been referred to in the growing body of knowledge as *knowledge acquisition* (Gaines & Boose 1988). Key methods and techniques include brainstorming the experts and introducing protocols for interviews for extracting heuristics. Again, these methods and techniques exploit human linguistic abilities to document knowledge.

The knowledge workers, it appears, deal with two kinds of knowledge: one which is very well documented and commented upon by the expert community, knowledge which is available in textbooks and classic research papers and monographs, together with the highly structured knowledge available in terminology databases; and the other kind includes knowledge which is still emerging and apparently lacks the consensus of the domain community. A number of authors denote this distinction in a variety of terms: *public* and *private knowledge*; *explicit* and *tacit knowledge*. Our concern is how these two kinds of knowledge are articulated through the medium of natural language. In this paper we attempt to demonstrate how methods in linguistics and in information extraction can be used to acquire knowledge from public knowledge sources, comprising well-documented and largely explicit knowledge, like digital libraries and terminology databases. We then demonstrate how these methods can be used to extract knowledge from 'private' knowledge sources, comprising hitherto undocumented and implicit knowledge, like, for example, (transcripts of) interviews in which experts describe how they solve problems.

Natural language is essentially a human phenomenon that has all the human traits of creativity and ambiguity. Those who study language point out that, despite the creativity and ambiguity inherent in language-based communication, language itself can be regarded as a *system*: 'a network of patterned relationships constituting the organisation of language', a network which can be measured in 'units of language'. Language can be described at different levels of generalisation. Languages have vocabularies and the level of individual words is referred to as the *lexical level*. The next level is that of grammar where the discussions vary but some linguists talk about categories and processes. This level can be further subdivided into two: at one level we can view language purely in terms of the structure or form of words (how we make plurals from singulars, how we describe tenses, how words change), referred to as the *morphological* level, and the second level is that of *syntax*, which deals with the rules that exist for governing the ways

in which words are combined to form sentences. More complex levels of language involve how we infer the meaning of words and sentences, the *semantic* level. The more abstract and, perhaps, abstruse level of linguistic description is where different language users—the stakeholders in our terminology—make choices about the use of language: the *pragmatic* level.

There are two important developments in the study of language that are relevant for acquiring knowledge from documents and from experts. The ambiguity and creativity that characterises natural language, especially the language of everyday use, is carefully controlled in texts produced by specialist communities. This does not mean to say that the specialist communities are not creative, but that they are more strict in the use of language than that encountered in everyday speech and writing. Linguistic literature refers to language of everyday use as *language for general purposes* (LGP) and specialist languages as *language for special purposes* (LSP). Perhaps one novel aspect of our work is that we utilise observations about LGP in order to extract terms in LSP text: specifically at the lexical and morphological levels. LSP texts are differentiated in terms of their use and this distinction includes expert-to-expert communication (journal papers, for example), expert-to-novice communication (technical manuals) and expert-to-layperson communication (government regulations, popular science articles). We also use this distinction, at the pragmatic level, for acquiring knowledge.

The other important development in the study of language is that of the use of large samples of speech and text acquired from users of language. Linguists organise large samples of written text and recorded speech and have developed protocols to minimise the bias of the builder of such collections—known in the literature as *text and speech corpora*. The name of this method of study of language is *corpus linguistics* (see, for example, Sinclair 1991; Stubbs 1996): a discipline dedicated to studying the preferences of language users as evidenced from text corpora, through techniques based on the statistics used on the frequency of these units. These techniques can be used for emergent units—new words or new usage of existing words—as well as by analysing text corpora; however, caution is essential here in that new words or new usage is, by definition, less frequent and statistics of small numbers have to be dealt with carefully. The key conclusion of some workers in this area is that judgments about the various structural levels and units of language correlate with the frequency of the occurrences of these units and levels. Therefore, if a word is very frequently used in preference to others, then that tells us something about the structure of language at the lexical level; spelling variants in British and American English texts is a good example here (Americans prefer *organization* and *burnt*, and the British *organisation* and *burned*, for instance) and another example is the use of a certain set of terms which indicates the preferences of a specialist community.

By studying the absolute and relative frequency of occurrence of words in corpora representative of LSP and LGP it is possible to gather evidence for the existence of structures within LSP and then to exploit these in the acquisition of knowledge from text. Lexical levels and grammatical levels can be studied by looking at the distribution of lexical, morphological and syntactic units within a corpus. Semantics of texts can be studied by making inferences based on lexical and grammatical units. Pragmatics can be studied by making comparisons between texts with different intentions and target audiences.

AUTOMATIC KNOWLEDGE ACQUISITION FROM TEXT

Corpus linguists have pioneered the notion of a *reference corpus*: a collection of texts published within a specified time period, for a varied readership and in different genres of publishing. These genres include fiction writing, newspaper writing, popular science, religious writing and others. This collection is selected carefully by a group of authors, linguists, literature experts, educationalists, scientists and others. This peer group *samples* the writing of a period in different genres to study how a language is used and how it has changed. The peer group may be focused on publishing learner dictionaries or language workbooks, on publishing reference grammars, and latterly on producing spellcheckers. Various national

and transnational initiatives have led to the creation of reference corpora; the educational publishers Longman have, in co-operation with the University of Lancaster, created the 30 million word Longman/Lancaster Corpus of Contemporary (*c*. mid-20th century) English (Summers 1993); UK universities and dictionary publishers have created the 100 million word British National Corpus having sampled texts largely published in the late 20th century (Aston & Burnard 1998). The reference corpora implicitly contain some of the knowledge about English language of everyday usage.

Our method of acquiring knowledge from text is as follows. In the first instance, a collection of machine-readable texts about the domain in question is gathered. Knowledge which is already publicly available, primarily because it has been documented in textbooks, learned journals, guidance notes (for managing assets, for instance) or legislation relating to the aquatic environment, is regarded as the primary source of knowledge. The word 'primary' is used in the sense of a progenitor; a source from which successive generations of knowledge are derived, even if it sometimes means negating the contents of the primary sources. This text collection of publicly available primary knowledge material will be referred to as the *mother corpus*. The mother corpus is regarded as a representative sample of the knowledge of the domain. Frequency of occurrence of terms in this corpus might, as we show later, be used as a guide to the key concepts of a given domain. Conditional sentences (IF_THEN type) in this corpus might, as we also show later, be used as candidate heuristics, or rules-of-thumb, in this domain.

Knowledge which is mainly experience-based is typically undocumented, and which is in 'the head of the experts' may be regarded as the progeny of the documented knowledge. For instance, an interview with an expert who deals with the planning or operation of an urban wastewater management system will reveal that he or she uses roughly the same stock of terms as can be found in the mother corpus, but with much greater or much lesser frequency as a way of emphasising his or her special interest within a specialism. There is also a good chance that the heuristics in textbooks or learned journals have been used by our expert but with more or less restrictions on the use of the heuristics, again to emphasise his or her expert view of the specialisation. Transcripts of a brainstorming session will reveal how a (small) group of stakeholders individually behave like our exemplar expert above. A corpus of texts comprising problem-solving interview transcripts, brainstorming sessions and other textual artefacts of implicit knowledge will be referred to as the *progeny corpus*.

A progeny corpus will not have the same representative status as the mother corpus, in that the former collection of texts have gone through a consensual process similar to the various checking processes a text goes through from manuscript to final published article undertaken by editors, referees, colleagues and others.

The reference corpora of everyday language, the mother corpus of a discipline, and the rather ephemeral, but nevertheless unique and useful progeny corpora, can be used to study how a language is used in general (a Language for General Purpose, or *LGP*, corpus), and how scientists and technologists use language to communicate concepts and artefacts of their specialism (a Language for Special Purpose, or *LSP*, corpus).

LSP AND SPECIALIST KNOWLEDGE: A CASE STUDY

Literature on knowledge acquisition deals extensively with the so-called knowledge acquisition bottleneck—the difficulties knowledge engineers face in extracting knowledge from experts (Boose 1992). Typically knowledge engineers have difficulty with the terminology of their problem domain. The terms of the domain, including the names of objects and artefacts comprising the domain, and their interrelationship, as found in rules and heuristics for example, do pose particular problems. Knowledge engineers end up making lists of terms *by hand* and seldom consult, or are aware of, the existence of terminology databases that may exist in their domain of interest. For a knowledge engineer without access to a verified and validated term database, the answer may be to ask the experts. In order to facilitate this consultation, it may be prudent for the knowledge engineer to analyse the

available texts and academic papers in advance, in order to extract a set of *candidate terms* for presentation to the experts. Discussion in the recent literature on specialist languages as to the most appropriate methods to use to facilitate the extraction of candidate terms has taken place (Ahmad & Rogers 2001). It will suffice to say here that specialist terms are used much more frequently in LSP than may be the case in language for general purposes (LGP). In addition, the probability of finding two or more words used contiguously is relatively high in LSP. We show how statistical techniques, based upon the frequency of usage of single words and the probability of words occurring together, can be used in identifying candidate terms and these have been employed in our most recent work.

Mother and progeny corpora for wastewater management (WWM)

The University of Surrey has been involved in a number of projects for building hydroinformatic systems for sewerage rehabilitation planning, for water-distribution network control, for managing algae growth, for licensing abstraction of river water, and for safe design of sewer systems (see Price *et al*. (1998) for details). More recently the University has developed a system for facilitating the planning of urban wastewater management projects in developing countries (Miles 2001).

These hydroinformatic systems comprised expert systems working in conjunction with either a simulation model or an expert system working together with a digital text library. The urban wastewater management system is a web-based program based on planning rules as incorporated in EU guidelines.

Surrey mother corpus of WWM. Over the years a corpus of published texts has been created at Surrey dealing mainly with the planning and design of water management systems and with related legislation. The mother corpus contains over 1.2 million words which were sampled from learned texts, full texts and abstracts, from journals (31 in number), legislation and statutes (8 texts), advertisements for public service (8 texts), official notices (31 texts), newspaper reports (7 texts) and random samples from advanced textbooks (6 texts) and academic textbooks (11 texts). These texts were published on the whole during 1970–1995.

Surrey progeny corpus of WWM. Each of the hydroinformatic systems involved an interview-based knowledge acquisition session. In this session an expert was asked to answer questions, and to think aloud, about specific problems in water management or wastewater management. During the past 15 years, the University of Surrey has conducted interviews with over 25 experts and verified and validated the interview contents by other experts who were advising on the development of Surrey's WWM hydroinformatics systems. Surrey's WWM progeny corpus contains over 50,000 words.

The texts included in the mother corpus have been analysed at the lexical level, at the syntactic level, at the semantic level and at the pragmatic level. By revealing structures in the text and enabling us to present candidate terms to the experts, the mother corpus thus allowed us, as knowledge engineers, to communicate more effectively and to better understand the domain, therefore facilitating the process of knowledge acquisition. Expert interviews have been carried out to gain a comprehension of the domain, and have been transcribed for further analysis. *A progeny corpus* has thus been created. A contrastive analysis comparing the mother corpus and the progeny corpus has been useful in ensuring that that the analysis of the mother corpus has revealed all of the important domain terminology. Where interviews have covered new territory and new terminology is revealed, we begin to deal with the tacit knowledge discussed earlier.

LSP of wastewater management: the lexical level

At the lexical level, we have assessed the distribution of words within the mother corpus. The relative frequency of each word (calculated as the quotient of the absolute frequency of the word and the total number of words in the corpus) has then been compared with the relative frequency of that word in a general language corpus, the reference corpus, in this case the British National Corpus (BNC), comprising over 102 million words (Aston & Burnard 1998). Where the frequency in the mother corpus

Table 1 | A comparison between the relative distribution of key words in the wastewater management mother corpus and the general language British National Corpus

	water	sanitation	discharge	sewer	catchment	abstract	effluent	drain	filter	pollute
Frequency	9,454	1,794	932	625	551	308	284	90	88	8
SuMoC	0.00758	0.00144	0.000747	0.000501	0.000442	0.000247	0.000228	7.22E-05	7.06E-05	6.41E-06
BNC	4.77E-04	7.09E-6	1.11E-05	3.58E-06	3.88E-07	3.84E-05	1.27E-06	2.07E-05	8.74E-06	2.91E-07
Freq. ratio	15.9	203	67.5	140	1,140	6.44	180	3.49	8.08	22

Table 2 | A comparison between the relative distribution of compound words in the wastewater management mother corpus and the general language British National Corpus

	groundwater	borehole	watercourse	overflow	stakeholder
Frequency	286	284	247	243	153
SuMoC	0.00029	0.000228	0.000198	0.000195	0.000123
SuPoC	4.92E-07	5.97E-07	1.94E-07	1.46E-06	9.76E-08
Freq. ratio	590	469	1,020	134	1,260

has been found to be significantly greater than that in the BNC, the word has been deemed a *candidate term*. We will call the Surrey mother corpus *SuMoC* and the Surrey progeny corpus *SuPoC*.

There are some interesting differences between our mother corpus and the British National Corpus. The ten terms in Table 1, extracted purely by virtue of their frequency, refer to key notions within the domain of urban wastewater management. The term *sewer*, for example, appears 140 times more frequently in our mother corpus than it does in general language use. As such, we would be inclined as knowledge engineers to investigate (and thus elaborate upon) the meaning of the term. The results indicate that the approach is merited as it will not only extract terms like these which are familiar to us, but will also reveal terms which are new to us but nevertheless represent important notions and therefore warrant further investigation.

Also of importance, and indicative of specialist language use, are *compound terms*, methods for the extraction of which are discussed in Ahmad (2001). The British National Corpus does contain some of those compound terms which are found in the field of wastewater management (some examples being *groundwater* and *borehole*) but these are used much more frequently in specialist language (Table 2 gives some of the frequencies derived from our corpus).

Key similarities and differences in the distribution of words and terms between the mother corpus and the progeny corpus are evident. A comparative analysis reveals that the ten most frequently occurring words in both corpora are all closed class words such as determiners, conjunctions, prepositions, pronouns or modal verbs. In the mother corpus these ten words comprise almost 27% of the total text; in the progeny corpus the equivalent figure is just under 24%. The first ten open class words (nouns, adjectives and adverbs) covered a much smaller percentage of the total text (3.2% and 2.9%, respectively). The four sets are shown in Table 3. One immediately notices that the candidate terms identified in the mother

Table 3 | The relative and absolute frequencies of the ten most frequent words in the mother and progeny corpora, together with the frequencies of the first ten open class words in the two corpora

	Mother corpus (1247,179 words): the ten most frequent closed class words									
	the	**of**	**to**	**and**	**in**	**a**	**or**	**for**	**be**	**is**
Frequency	83,873	58,698	39,567	35,573	27,763	24,269	17,270	16,648	16,177	14,709
SuMOC	6.73E-02	4.71E-02	3.17E-02	2.85E-02	2.23E-02	1.95E-02	1.38E-02	1.33E-02	1.30E-02	1.18E-02
BNC	6.12E-02	3.06E-02	2.50E-02	2.79E-02	1.87E-02	2.20E-02	3.97E-03	7.51E-03	5.91E-03	9.08E-03
Freq. ratio	1.10	1.54	1.27	1.02	1.19	0.89	3.48	1.77	2.20	1.30
	Mother corpus: the ten most frequent open class words/candidate terms									
	water	**authority**	**section**	**act**	**order**	**land**	**paragraph**	**local**	**person**	**urban**
Frequency	9454	6573	5022	3450	3241	2682	2585	2455	2235	2097
SuMoC	7.58E-03	5.27E-03	4.03E-03	2.77E-03	2.60E-03	2.15E-03	2.07E-03	1.97E-03	1.79E-03	1.68E-03
BNC	4.77E-04	9.31E-05	8.14E-05	1.86E-04	3.09E-04	1.94E-04	1.73E-05	1.84E-04	2.28E-04	3.39E-05
Freq. ratio	15.89	56.61	49.51	14.89	8.42	11.08	119.65	10.71	7.85	49.56
	Progeny corpus (242,416 words): the ten most frequent closed class words									
	the	**to**	**of**	**and**	**a**	**in**	**is**	**that**	**it**	**you**
Frequency	13,593	7,050	6,742	6,017	5,616	4,352	4,340	3,689	3,238	3,133
SuPoC	5.61E-02	2.91E-02	2.78E-02	2.48E-02	2.32E-02	1.80E-02	1.79E-02	1.52E-02	1.34E-02	1.29E-02
BNC	6.09E-02	2.51E-02	3.06E-02	2.80E-02	2.19E-02	1.87E-02	9.09E-03	1.10E-02	1.13E-02	6.03E-03
Freq. ratio	0.92	1.16	0.91	0.89	1.06	0.96	1.97	1.38	1.19	2.14
	Progeny corpus: the ten most frequent open class words/candidate terms									
	water	**flow**	**model**	**system**	**area**	**pipe**	**results**	**licence**	**file**	**time**
Frequency	1,019	884	824	813	809	652	569	548	483	450
SuPoC	4.20E-03	3.65E-03	3.40E-03	3.35E-03	3.34E-03	2.69E-03	2.35E-03	2.26E-03	1.99E-03	1.86E-03
BNC	4.77E-04	5.31E-05	1.04E-04	4.07E-04	1.30E-04	2.69E-05	9.59E-05	5.54E-06	4.76E-05	1.74E-03
Freq. ratio	8.81	68.74	32.69	8.23	25.69	100.00	24.50	407.94	41.81	1.07

Table 4 | The relative frequency of keywords and their morphological variants in the wastewater management corpus and the wastewater management progeny corpus. The lemmas are shown in bold in the first column

	Surrey Mother Corpus (SuMoC)			Surrey Progeny Corpus (SuPoC)		
	Frequency	Rel. freq.	Weirdness (Mother/BNC)	Frequency	Rel. freq.	Freq. ratio (transcript/mother)
pollute	8	6.41E-06	22	1	4.13E-06	0.64
pollution	948	0.00076	94.3	80	0.00033	0.43
pollutants	182	0.000146	501	3	1.24E-05	0.08
pollutant	141	0.000113	Infinite	1	4.13E-06	0.04
polluting	80	6.41E-05	220	2	8.25E-06	0.13
polluted	53	4.25E-05	25.7	3	1.24E-05	0.29
polluter	14	1.12E-05	116	0	0	0.00
polluters	4	3.21E-06	Infinite	2	8.25E-06	2.57
drain	90	7.22E-05	3.49	24	9.90E-05	1.37
drainage	1,764	0.00141	286	253	0.00104	0.74
drains	103	8.26E-05	23.6	37	0.000153	1.85
draining	57	4.57E-05	13.1	19	7.84E-05	1.72
drained	46	3.69E-05	3.27	14	5.78E-05	1.57

corpus are strongly influenced by the language of the legal profession, whereas the progeny corpus includes discussion of modelling: *model*, *results* and *files* are key terms. A comparison of the different registers or types of text within the mother corpus is also given.

LSP of wastewater management: the morphological level

An inspection of the morphological variance of the key terms used in the two corpora reveal some interesting characteristics in the way scientists and technical experts use language. Let us look at the morphological variants of the lemma *to pollute*. The term *pollute* is used as a verb and as such it has its inflexional variants *polluted*, *polluting* and *pollutes*. In addition, and perhaps most importantly, we see the derivational variants of pollute, particularly the noun *pollution* (inflexional variants preserve grammatical category but derivational variants are of a different class). The same is true of the verb *to drain*. Again, the key variant is the nominalisation of the verb, *drainage* (see Table 4).

Halliday & Martin (1993) have remarked that verbs are regrammaticised in scientific discourse into nouns in order to create 'things' which can be observed and experimented with. Hence, 'is polluted' often becomes something like 'pollution is found in' and 'it drains' becomes something like 'drainage occurs'. Only 141 instances of the verb pollute being used are found in the

Table 5 | The relative frequency of the lemma sewer and its morphological variants in the wastewater management corpus and the wastewater management progeny corpus

	Surrey Mother Corpus (SuMoC)			Surrey Progeny Corpus (SuPoC)		
	Frequency	Rel. freq.	Weirdness (Mother/BNC)	Frequency	Rel. freq.	Freq. ratio (transcript/mother)
sewer	625	0.000501	140	436	0.0018	3.59
sewerage	590	0.000473	2,440	43	0.000177	0.37
sewers	391	0.000314	104	382	0.00158	5.03
sewered	17	1.36E-05	Infinite	5	2.06E-05	1.51

mother corpus (and inspection reveals that in many instances 'polluted' is used as an adjective rather than in the past tense) whilst the nominalisation of the verb, 'pollution', is found 948 times. The phenomenon is particularly pronounced in written (as opposed to spoken) scientific discourse (as is evidenced in the right hand column of Table 4, which gives the relative frequency of each term in the progeny corpus in relation to its frequency in the mother corpus).

LSP of wastewater management: the semantic level

If we accept that the key variants of lemmas which are used primarily as verbs are their nominalisations, then it can be argued that the key variants of lemmas which are used primarily as nouns are their plurals. Referring as they do to numbers of objects, plurals can be considered semantic variants. The plurals of the lemma which refer to important notions within a domain are more prominent in special language than they are in general language use. The phenomena is evident in Table 4 with regard to the lemma drain (used as both a verb and a noun) and is evident in Table 5 when we look at the lemma *sewer*.

The plural of *sewer* is used over 500 times more frequently in the progeny corpus than in general language use. The fact that plurals are often used to denote classes of objects and events in scientific discourse may account for this. In fact, no fewer than 18 classes of sewer are referred to in total within the progeny corpus (see Table 6).

Further analysis upon the collocation of words reveals that, where terms are found to collocate frequently, they often represent important objects within a domain. This is particularly true when these collocations are referred to in the plural as well as the singular. If a plural is present but

Table 6 | The different classes of sewer referred to within our progeny corpus

Class of sewer	Frequency of occurrence	Class of sewer	Frequency of occurrence
tank sewers	19	arch sewers	2
brick sewers	12	foul sewers	2
trunk sewers	9	core sewers	2
entry sewers	8	collector sewers	2
interceptor sewers	6	surcharge sewers	1
task sewers	6	major sewers	1
pipe sewers	5	minor sewers	1
storm sewers	4	house sewers	1
line tank sewers	4	surface water sewers	1

Table 7 | Collocating terms extracted from the progeny corpus which occur in plural

Collocating terms	Frequency	Plural	Frequency
water resource	4	water resources	48
line tank	27	line tanks	34
structural problem	2	structural problems	22
tank sewer	16	tank sewers	19
concrete tank	12	concrete tanks	14
hydraulic problem	5	hydraulic problems	14
storage tank	7	storage tanks	12

only referred to once or twice, then, as might be expected, the terms are not so likely to represent important objects (see Table 7). Collocating terms which have been referred to in the plural, automatically extracted from the progeny corpus, are shown together with their frequencies of occurrence.

Even a short consultation with an experienced practitioner will reveal that the classes of sewer referred to previously can be further divided into those that are defined by their cross-section, those that are defined by what they carry and so on. Collections of candidate terms such as these, then, can be used to begin to build up domain ontologies. Thus, through eliciting knowledge about term databases such as these, derived purely by virtue of the frequency of occurrence of terms, and through investigating how those terms and their morphological variants are used in context, we gain a better understanding of the domain in question.

LSP of wastewater management: the pragmatic level

We can study the structure of specialist texts at the pragmatic level by comparing registers of different types of text against one another. In Table 8(a, b), a comparison between the collection of journal papers held in our mother corpus and the legislation in the same corpus shows that quite different terms are important. Our journal papers are concerned with *software, systems, design* and *drainage*, whilst the legislation is built around *authority, section, order* and *schedule*. Note that, in both cases, the most frequently occurring 100 words make up over half of the respective corpora, in particular, that over one quarter of both corpora are made up of the first ten most frequently occurring closed class words. The legislation register is also identified by the prominence of words such as *any* (the universal quantifier), *shall* and *may*.

EXTRACTING INFORMATION AND PROCESSING KNOWLEDGE IN OUR CORPORA

Elaborating terms

The key point about the way in which patterns can be found at different levels in particular types of text is that they can help us to understand them and in turn to develop computer programs which are able to analyse them. Texts can be understood on a number of levels; whether lexical, syntactical (and morphological), semantic or pragmatic. Only humans are able to understand texts on all of these levels, but by beginning to exploit those patterns which emerge in the analysis of texts, we can start to extract information from them. It is possible that in several years' time techniques and technologies will be available to enable computers to understand texts at all of these levels, thereby permitting the automatic elicitation of knowledge from text. For now though, and as a first step, the characteristics of language for specialist purposes (LSP) can be exploited to enable us to extract candidate terms. Looking at how these candidate terms occur in context can help in their elaboration and can reveal classes of objects and therein help to build up the structure of a domain. The knowledge acquisition process can be seen to be an iterative one. Once key terms have been extracted they can be fed back into the process as search terms in order to assist with their elaboration, effectively 'informing' the knowledge acquisition process. Fragments of knowledge, such as rules and semantics, can subsequently be extracted and these in turn can lead the

Table 8(a) | The 100 most frequently occurring word tokens in the journals register of SuMoC (67,443 words in total)

Most frequently occurring word tokens in the journals register of our mother corpus (67,443 words in total)	% of corpus	Cumulative % of corpus
the, of, and, to, a, in, is, for, be, that	26.68%	26.68%
as, on, with, are, or, by, **software**, **system**, this, **water**	6.03%	32.71%
systems, can, **design**, it, from, **drainage**, such, **flow**, an, **data**	4.14%	36.85%
knowledge, at, **model**, has, have, used, **time**, **urban**, which, **simulation**	3.10%	39.95%
method, these, was, **modelling**, **engineering**, there, **wallingford**, will, **information**, been	2.33%	42.28%
pipe, **network**, may, use, **number**, more, **models**, **analysis**, not, **surface**	2.05%	44.33%
development, **rainfall**, **hydroinformatics**, each, using, **uk**, then, **based**, **research**, **runoff**	1.77%	46.10%
expert, **conduit**, **cost**, also, through, where, **procedure**, new, if, **networks**	1.59%	47.70%
other, **computational**, **hydraulic**, **methods**, its, **computer**, **management**, than, therefore, user	1.48%	49.17%
storm, were, **equations**, **flows**, **given**, their, all, **level**, some, **control**	1.38%	50.56%

Table 8(b) | The 100 most frequently occurring word tokens in the legislation register of SuMoC (158,795 words in total)

Most frequently occurring word tokens in the legislation register of our mother corpus (158,795 words in total)	% of corpus	Cumulative % of corpus
the, of, to, in, or, a, and, any, for, by	31.44%	31.44%
be, **authority**, shall, this, is, as, that, under, section, which	9.79%	41.23%
with, c, **water**, on, **such**, **act**, above, b, may, **schedule**	6.41%	47.65%
an, **order**, **paragraph**, **respect**, **land**, not, it, **other**, **provisions**, from	4.26%	51.91%
person, **subsection**, **purposes**, made, **works**, **state**, **relation**, if, **part**, **flood**	3.22%	55.12%
where, **defence**, **drainage**, **application**, are, **waters**, **secretary**, has, **notice**, **power**	2.71%	57.83%
means, have, s, at, **relevant**, **subject**, **powers**, so, **local**, **licence**	2.35%	60.19%
out, **provision**, **amp**, make, been, **functions**, **committee**, r, sub, sch	1.93%	62.11%
consent, **compensation**, **member**, **virtue**, **within**, **chapter**, **charges**, those, ii, **effect**	1.69%	63.80%
area, **orders**, **below**, **article**, **specified**, he, into, **scheme**, **case**, etc	1.51%	65.31%

knowledge engineer to investigate the relationships within the domain further.

Extracting heuristics from texts?

Once the terms of interest in a specialist domain have been extracted and elaborated upon to indicate the pertinent objects and concepts within that domain, knowledge engineers must begin to look at the relationships between these objects in order to build up a model of the domain and the problem solving behaviour employed by experts. Of particular importance with regard to the extraction of knowledge from LSP texts are semantic relations. Cruse (1986) discusses a wide range of these relations and distinguishes between several different types of hierarchical relation such as hyponymy, taxonomy and meronymy. He also considers various kinds of synonymy and antonymy and introduces the idea of 'diagnostic frames' for some of these relations. These essentially consist of particular phrases which denote a particular relation. To provide an example, it can be seen that the phrase 'X is a kind of Y' denotes the relation of hyponymy. There are a number of ways in which these frames can be expressed within natural language, and the phrases 'X is a type of Y' or 'X is a species of Y' are examples. Ahmad (2001) has shown that knowledge can be extracted from a text corpus by taking account of these diagnostic frames and formulae.

A great many knowledge based systems use *production rules* to store and retrieve aspects of problem solving knowledge, especially rules-of-thumb or heuristics. Defined as a procedural response triggered by a pattern, production rules are commonly structured in the format:

IF a pattern is matched **THEN** schedule a procedure for execution

Various names are used within the knowledge based systems literature to refer to the IF part of such a clause. The antecedent part, the condition part, and even the left hand side of the rule are common. Accordingly, the THEN part of the clause is referred to as the consequent part, the action part and the right hand side of the rule. Though in some cases the words 'if' and 'then' may actually be used within texts to point to candidate rules or heuristics, this is not always the case. A set of semantic cues, or words which might be used to identify these heuristics, have been identified (see Table 9).

Table 9 | Cues which might be used to identify heuristics

affect	as a rule	as long as	assuming
because	customarily	due to	effect of
generally	hypothesis	if	if then
in general	therefore	precondition	premise
provided	proviso	reason	regularly
rule of thumb	seldom	so that	to ensure
typically	unless	usually	when
normally	ordinarily		

Computer programs can be used to search for such cues to help find candidate rules. The following candidate rule is a typical example, extracted from the transcripts of interviews carried out recently in which domain interface groups consulted experts about the conceptualisation of urban areas for the purpose of modelling sewerage networks. The semantic cues which led us to the candidate rule are shown in the left hand column of Table 10.

A second example shows a rule taken from the transcripts of the interviews conducted for the development of the SafeDIS system. An expert is discussing assumptions which can be made about the characteristics of sewers for modelling purposes (see Table 11).

Although rules such as these can be found within technical manuals, textbooks and so on, many are not well documented and are often passed on from practitioner to practitioner and from expert to novice by word of mouth. It is sometimes only when experts are interviewed and asked to explain how they solve a given problem that such heuristics are articulated, and so the analysis of verbatim transcripts of such interviews can reveal them for the first time.

Table 10 | A typical candidate rule, extracted by searching for the semantic cues on the left

Semantic cues	Section of transcript	Paper knowledge base rule
if **then** **therefore** **normally** **if** **then**	So, **if** you had a town which was immediately downstream of another town, if it was a hundred miles, [**then**] you would assume, or you might be able to say, or be able to prove that the river had actually recovered to a certain state, and **therefore** you could assume a boundary condition at the upstream end of that town to be **normally** true, to be able to treat them as separate towns, and **if** they are ten miles apart **then** of course you could not treat them as separate towns, they are just one town. You have to think in terms of are they linked or not linked for the system you are concerned with.	**IF** two urban areas lie on the same river less than 100 miles apart **THEN** their combined effect upon water quality should be considered

Table 11 | A second candidate rule, identified by the semantic cues, *generally* and *typically*

Semantic cues	Section of transcript	Paper knowledge base rule
generally **typically**	I've already mentioned problems of roughness of estimation etc. Now, we have a lot of sewers and I'm sure most other old cities and towns do, where they're **generally** described as egg-shaped, they're not conforming to any predefined pattern exactly, they're very **typically** a flat stone invert which might be 250–300 mm wide and which you can walk in. It might have trapezoidal dry stone walls, effectively.	**IF** old sewers are described as egg-shaped **THEN** modellers should be aware that they may well have a trapezoidal cross-section, with rough stone walls

AFTERWORD

The acquisition and elaboration of relevant terminology is often considered to be the most problematic stage of knowledge acquisition, but a systematic approach is rarely adopted in response. Here some techniques from the field of special languages and corpus linguistics have been proposed as solutions. By studying the structures which are found at the four linguistic levels in language for specialist purposes, we can begin to see how computer programs might be used to extract information from free text. We have shown that the process of extracting terms and rules relating these terms can be facilitated by computer programs and, indeed, the experience of developing the hydroinformatic systems at Surrey suggests that the process of building knowledge based systems might be eased considerably.

QUIRTE: *Quir*kiness of Specialist *Te*xts

The differences in the lexical and morphological texture of the specialist texts is important for our method. Essentially, we rely on the fact that general language texts are dominated by closed class words (determiners, prepositions, conjunctions and modal verbs) in that these words are amongst the most frequent. This domination of closed class words is evident in both the mother corpus

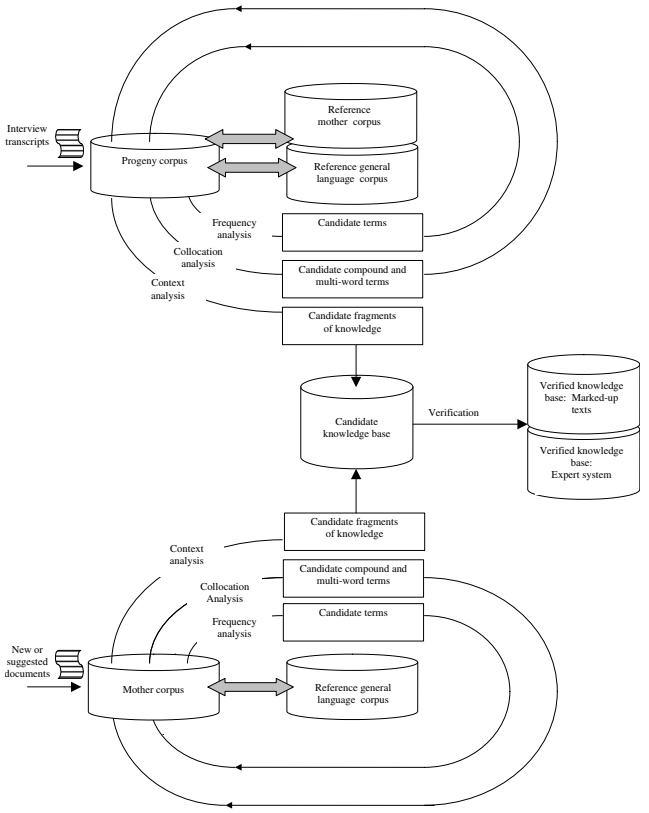

Figure 1 | QUIRTE: a knowledge acquisition method.

and the progeny corpus, but we see key open class words, generally nouns, indicating names of concepts and artefacts, for example, amongst the *most* frequent words. We also note that these frequent words are used more in their plural form as well as the basis for forming adjectives and verbs. The key words also form the nucleus around which new compound words are formed. The general language corpora are impoverished in compound words as these words are used to provide texture to specialist texts.

One can argue that the lexical and morphological texture give a *quir*kiness to specialist *te*xts (QUIRTE). Our method of contrasting general language texts with specialist texts, and specialist texts among themselves, relies on QUIRTE – the name of our method (see Figure 1).

Local knowledge and its incorporation in knowledge bases

Unlike many other specialist domains, urban wastewater management is fundamentally linked to the geographic location to which it is applied. The climate, physical and human geography, economy and history of the town or city to which a technology is to be applied are all as important as the technology itself. There are as many 'best management practices' as there are towns and cities, and as such the knowledge of the domain is continually revised and refined as practitioners apply it to new towns and cities. In addition, though wastewater management technologies themselves cannot be said to be rapidly evolving in the same way that, say, communications technologies are, the entire framework within which urban wastewater management takes place (from policy and legislation through to such things as the urban planning procedure and public attitudes towards the environment) is of sufficient complexity to ensure that knowledge of the overall domain is constantly shifting and being revised.

Where domains are rapidly evolving, it is necessary for the means by which knowledge is managed to be dynamic. If it is acknowledged that knowledge management is key to the effective and sustainable management of the aquatic environment, and that linguistic resources such as terminology databases and digital libraries are important tools for knowledge management, then a method for the automatic generation of terminology databases and hypertext documents such as that outlined is of value.

A note for the future? Stakeholders and a language-informed hydroinformatics system

Some aspects of the knowledge of specialist domains may be tacit and will reside only in the heads of individuals, while other aspects may be explicit, documented and held within textbooks, journal papers and other texts of the domain. Methods for the acquisition of knowledge from individuals have been established to support the development of knowledge-based (expert) systems. Contemporaneously, techniques for the extraction of knowledge from text have been developed by practitioners of computational linguistics. Here we have shown how the two can support one another.

The knowledge engineer uses techniques from computational (and corpus) linguistics to extract terminology and fragments of knowledge from existing text documents. By studying the extracted knowledge, the knowledge engineer becomes more familiar with the domain and its terminology. The extracted knowledge is then discussed in interviews with experienced practitioners, thereby acting as a prompt before being used as the basis of interviews to elicit tacit knowledge from these individuals or groups of individuals.

These interviews are themselves transcribed and, together with any additional text documents suggested by the interviewee, are analysed using the method employed previously to look for new terms and to look for heuristics typically articulated using IF_THEN constructs. New documents are compared first with a general English language corpus and then with the mother corpus to reveal emergent knowledge. The method is one which can be partly automated, thereby facilitating the creation and ongoing development of knowledge bases for specialist domains.

The management of the aquatic environment involves many different stakeholders. This is particularly true in densely populated urban areas. These stakeholders, of various backgrounds and with different levels of understanding of the various disciplines which contribute to urban wastewater management, are demanding to take part in the associated decision making processes. For these reasons, and because decisions are often made on the basis of the results of complex computational models, the domain is one particularly in need of a common knowledge base. By making clear the important objects and concepts within the domain and by making transparent the relationships between them, such a knowledge base would provide the means to assist members of the public and practitioners of disparate disciplines alike in understanding and making a contribution to the decision making process. The method presented here for the acquisition of knowledge will facilitate the construction of such a knowledge base and may therefore provide the means to bring about more inclusive procedures for the management of the aquatic environment.

REFERENCES

Ahmad, K. & Rogers, M. 2001 The analysis of text corpora for the creation of advanced terminology databases. In: *The Handbook of Terminology Management* (ed. Wright, S. E. & Budin, G.), pp. 725–760. John Benjamins, Amsterdam.

Ahmad, K. 2001 The role of terminology work in artificial intelligence. In: *The Handbook of Terminology Management* (ed. Wright, S. E. & Budin, G.), pp. 809–844. John Benjamins, Amsterdam.

Aston, G. & Burnard, L. 1998 *The BNC Handbook: Exploring the British National Corpus*. Edinburgh University Press, Edinburgh.

Boose, J. H. 1992 Knowledge acquisition. In *Encyclopaedia of Artificial Intelligence*, 2nd edn (ed. Shapiro, S. M.) pp. 719–742. John Wiley and Sons, New York.

Cruse, D. A. 1986 *Lexical Semantics*. Cambridge University Press, Cambridge.

Drucker , P. F. 1998 The coming of the new organization. In: *Harvard Business Review on Knowledge Management*. pp. Harvard Business School Press, Boston, MA.

Gaines, B. R. & Boose, J. H. 1988 *Knowledge Acquisition for Knowledge Based Systems*. Academic Press, London.

Halliday, M. A. K. & Martin, J. R. 1993 *Writing Science: Literary and Discursive Power*. The Falmer Press, London.

Miles, L. 2001 Knowledge management and environmental management. *PhD thesis*, University of Surrey, Guildford, unpublished.

Nonaka, I. & Takeuchi, H. 1995 *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, Oxford.

Price, R. K., Ahmad, K. & Holz, P. 1998 *Hydroinfomatics concepts. In Hydroinformatics Tools* (ed. Marselak, J., Maksimovic, C., Zeman, E. & Price, R.), pp 47–76. Kluwer, Dordrecht.

Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Stubbs, M. 1996 *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*. Blackwell, Oxford.

Summers, D. 1993 The Longman/Lancaster corpus: Criteria and design. *Int. J. Lexicog.* **6** (3), 181–208.