Pragmatics of Specialist Terms: The Acquisition and Representation of Terminology

Khurshid Ahmad

Artificial Intelligence Group, Department of Mathematical and Computing Sciences, University of Surrey, Guildford, Surrey, GU2 5XH.

Abstract. The compilation of specialist terminology requires an understanding of how specialists coin and use terms of their specialisms. We show how an exploitation of the pragmatic features of specialist terms will help in the semi-automatic extraction of terms and in the organisation of terms in terminology data banks.

1 Introduction

The word terminology has at least two senses. According to the first, the more popular definition, terminology refers to a systematically collected and organised vocabulary of technical terms used in a particular field, subject or science. In the second, and perhaps less well-known definition, terminology is the study of nomenclature or the science of terms. This science deals with the coinage, establishment, collection, usage and obsolescence of terms of specialist domains. The coinage-to-obsolescence cycle can be construed as a life-cycle, involving stages of inception, growth, maturity and the ultimate redundancy of a term or indeed collections of terms. Terminology science is the study of this life-cycle.

The terms of a specialist domain are one of the domain's key resources: a resource which has to be nurtured and conserved, a resource which has to be managed. The management of terminology deals with practical questions related to the entire life-cycle of terms of a specialist domain, involving a critique of methods, tools and techniques used in the identification, collection, verification/validation, standardisation, and storage in, as well as retrieval from, a specially designed data management system - a terminology data bank or term bank.

The major bottleneck in acquiring terminology is the dependence of the terminologist on the availability of articulate specialists; the terminologist's ability to find and comprehend the written output of a growing plethora of highly specialised domains is being increasingly taxed. This output is perhaps the single most important resource for identifying and elaborating specialist terms. Specialists rely heavily on written text for describing new ideas, consolidating current theory and practice, and censoring outdated concepts and traditions. The uses and abuses of ideas, practice, theories and traditions are almost always related to the uses and abuses of terms of the domain.

The claim that terminology is a science of terms can be challenged by semanticists, epistemologists, information scientists, grammarians, lexicographers and text
linguists. Semanticists would argue that since terminology deals with the semantics of linguistic expressions, then it really is a branch of semantics dealing with *semantic fields* in restricted domains. Epistemologists can argue that since terminology deals with issues of knowledge primitives, then it really is a part of epistemology. Information scientists can claim that their forays into classification are really what terminology is all about. Grammarians, such as Zellig Harris (1990), have discussed issues related to terminology under the rubric of *sublanguages*. Harris, together with Kittredge and Lehrberger (1982), argues that the collocation patterns observed in scientific discourse, including compound terms and idiosyncratic clauses and sentences, can be understood in the structuralist paradigm pioneered by Harris.

Lexicographers will argue that terminology is really dealing with pragmatics: terms are used by an identifiable discourse community, much like a dialect used in a geographically identifiable area or a sociolect used by a socially identifiable group of people, and, provided we can systematically keep track of the pragmatic data, problems of terminology are almost indistinguishable from those of lexicography. Last but not least, work in text linguistics, particularly that by de Beaugrande and Dressler (1981), includes a spirited discussion of terminological issues.

Michael Halliday has used the notions he has developed in systemic linguistics together with his predications about language as a 'social semiotic' (Halliday 1978), to explore the historical relationships between science, language and literature (Halliday and Martin 1993). Halliday's explorations are of significant import for terminology science in that he has commented on the language used by scientists and has argued that technical terms play an important role in the creation of 'a discourse of organised knowledge'. Focusing on the language of science text books used in schools, and with one eye on what Halliday and Martin perceive as the 'alienating' and 'anti-democratic' nature of twentieth century science, these authors have looked at scientific discourse from the times of Chaucer, through Newton, right up to the language of modern-day science, with exemplars from non-linear dynamics (popularly known as *chaos theory*), cognitive psychology, material sciences, geography and biology. For Halliday et al. the uniqueness of scientific language lies in the 'lexicogrammar' or *wording* of the language.

The scientist adapts and innovates upon the lexicogrammatical resources of his or her first language, or that of the second (or third) language if he or she does not use the first language for scientific writing, for expressing abstract and concrete ideas, for simplifying and summarising complex facts, for explicating and exemplifying, for contradicting and reinforcing, for categorising and stating exceptions to rules. And, in the execution of all these complex tasks - tasks which are executed through the medium of written text - scientists coin new terms, suppress the use of some extant terms, restrict or expand the scope of terms, borrow from other disciplines and from other languages. Terminology, as a collection of vocabulary items and as a science of terms, plays a central role in scientific endeavour.

Given the strategic importance of terminology in technical and scientific documentation and in human and machine translation, it must be gratifying for terminologists that semanticists, philosophers, text and systemic linguists, lexicographers, and information scientists are also interested in terminology. However, while the concerns of a terminologist are similar to those of all the above,
the terminologist also has practical considerations of terminology management to address.

We would like to argue that terminology management can be regarded as a study of special languages from the point of view of the users, especially of the choices they make, the constraints they encounter in using such a variety of language, and the effect their language use has on the other participants in an act of communication. The definition of terminology management adopted here is a paraphrase of Crystal’s definition of pragmatics (1992:217). The central problem of terminology management is to understand and deal with the pragmatics of specialist terms and the pragmatics of specialist texts. But what of terminology management now?

Much of the business of terminology management revolves around the maintenance of terminology data banks. There are international standards (ISO R/1087, 1969), national guidelines (BSI, DIN etc.), and company policies on how to collect and organise data about terms that are already in possession of terminology data bank managers. Terms in existing data banks are stored using a template record format. Typically, the record format, a format developed by the standards organisations or autonomous enterprises, specifies slot names for the data related to a typical term; some of these slots are mandatory and others are optional. A record format can have as few as 20 such slots or as many as 60. Essentially, these slots can be grouped under four major groupings, each group referring to one aspect of the terminology life cycle: acquisition, representation, deployment and dissemination (Table 1):

| Slot Group    | Exemplar Slots                                      |
|---------------|-----------------------------------------------------|
| Acquisitional | headword, terminologist’s name, date of entry       |
| Representational | grammar, foreign language, equivalent, language variety |
| Deployment    | definition, contextual example                       |
| Dissemination | synonyms, antonyms                                   |

Table 1. Slots used in terminology record formats

Existing term banks do not as a matter of course focus on usage-related data. However, such data would not only be of help to many users of term banks, but would also help in the efficient organisation of term banks. Terminology entries in a typical term bank do contain deprecated terms, and categories like currency, frequency of use, regional or geographic variation and style are generally avoided. The absence of such information is usually due to the terminologist’s concern about the introduction of polysemy in term banks by the consideration of these pragmatic categories. But more of this later.

By contrast, general-language dictionaries usually contain some pragmatic data. More recently, the methods and techniques of computer-based corpus-oriented lexicography have allowed lexicographers to compute frequency and variance data, for instance, with a speed and accuracy that was not possible only a decade ago (Sinclair 1987, 1991). Landau (1989) has outlined usage categories in general-language dictionaries devised for providing pragmatic information together with the typical values associated with each of the categories (see Table 2).
In this paper, it will be shown that the pragmatic categories are relevant to terminology management, like currency, frequency of use, regional or geographic variation. Whilst others, including style (functional variety or register) and status (items 8 and 9 in Table 2) are to be defined more precisely for specialist terms. The 'slang' category is relevant to terminology in that slangs of specialist terms are used in describing complex artefacts, particularly in populist use like advertisements and reportage in mass-circulation print media and in conversations on the factory floor. 'Restricted' and 'insult' categories are not directly applicable to specialist texts. Nevertheless, one can trace the development of a (scientific) theory sometimes by the censorship applied by a specialist domain community and the ways in which certain terms are used in a rather pejorative manner.

The prevalent situation in terminology, as far as pragmatic issues are concerned, can be summed up as follows. Terminologists are concerned with ensuring that scientists and technologists are able to communicate amongst themselves, and with others, in as unambiguous a manner as possible. Much of the discussion in prescriptive terminology is focused on the establishment of monosemy with mononymy and a reduction in polysemy and the elimination of synonymy. As far as a terminology is concerned, once a term has been coined in a natural language, then all regional and societal varieties of that language should endeavour to use the same term. The frequency of usage is not relevant to some in the terminology community, rooted as they are in the traditions of the standards/normative organisations, in that a scientific term, for them, is approved or rejected by standards organisations. The question of whether a certain discourse (domain) community uses a term frequently, rarely, or never, is not really an issue for them. Deprecated terms are entered in traditional term banks usually to persuade a translator, technical writer or other members of a specialist group that these terms are to be avoided, if possible.

The two major areas of terminology management which we claim can benefit from exploring and exploiting pragmatic features are term identification, and the organisation/dissemination of terms. In the area of term identification, the pragmatic features can be used for identifying new terms, or neologisms, of a specialism from the text archives of the specialism, including learned journals, books, popular articles, newspaper reportage and advertisement material associated with the specialism. A concordance of the specialist archive is produced together with the frequency of all words in the archive. The results of this concordance are compared with a
diachronically compatible corpus of general-language texts. Such a contrast is a very simple way of identifying terms as we will show later, and the concordancing of terms is by far the quickest way of establishing its potential lexicogrammatical characteristics. Such a procedure expedites the identification and elaboration of neologisms. Considering that building a term bank is a complex and labour intensive task, that each term can cost up to £10.00 to identify and elaborate, and that a typical domain in science and engineering may require its community to be familiar and fluent with around 2500 to 4000 terms, any saving in term identification will make the difference between having a good quality term bank and having no term bank at all!

In the area of terminology organisation and dissemination, pragmatic features can be used to organise and disseminate the terminology of the domain, particularly by partitioning a term bank into a mother term bank, containing, for example, frequent terms, terms in the national variety of the documentor's/translator's language, and daughter term banks that may contain, for instance, less frequent terms, other regional varieties, lesser-used synonyms, and so on. The expectation here is that the documentor or translator will probably be familiar with terms in the mother term bank, but less familiar with the terms in the daughter term bank. We have estimated that a fully elaborated term bank of, say, 4000 terms can demand over 50MB of data; thus a pragmatic partitioning of term banks will make their availability on smaller computer systems more feasible.

Sections 2 and 3 of this paper deal with the acquisition of terms. Section 4 with the representation of terms using pragmatic criteria. Section 5 concludes the paper.

Section 2 contains a discussion of the quantitative difference between specialist texts of a given domain and a representative sample of general-language texts: these differences are crucial for our methodology for semi-automatically acquiring terms from a corpora of specialist texts. Problems related to the design of special-language corpora are discussed in the light of the fact that these texts, like learned journals, books, advertisements, etc., are written for different audience groups. Section 3 describes how the methodology has been effected through the use of a corpus-based terminology (and lexicography) management system - System Quirk. The exploration of the pragmatics of specialist texts will help in the acquisition of terms. Section 4 describes how the pragmatic features of a term can be exploited to partition terms on the basis of an empirical notion of how familiar a typical term bank user may be with a given term.

2 Terminology in Text and its Acquisition

The inextricable link between knowledge and language, a source of continual and heated philosophical debate, is of considerable import to terminology science and practice, particularly in view of the rapid and massive increase in the volume of literature related to science and technology. This increase has been accompanied by a new genre of writing, specialist writing, a genre that uses and adapts and innovates on the general language to create the special-language literature^3. Exemplars of specialist writing include the familiar learned-journal text type, the textbook and technical manual texts. But specialist writing also includes the popular
science and hobby magazine text types and the newspaper feature and specialist reportage. Furthermore, as most specialisms are involved in trading goods and services, then texts such as advertisements and brochures are also a part of the text archive of a domain.

2.1 The 'Weirdness' of Special-language Texts

A distinctive feature of specialist writing is the profusion of scientific terms in this genre. These are single words or multi-word phrases and are usually coined to express an idea, a fact, or a phenomenon with a view on the part of the person coining the term to the new term being adopted by his or her peers. Scientists show considerable dexterity in coining terms, borrowing sometimes from the dominant language of science and technology, for example, depending on the historical time, from Latin, Greek, Arabic (for example, atom, nucleus, vitality, chemistry, zero) and now English. Sometimes acronyms are created (cf. laser - acronym for light amplification through stimulated energy radiation), whilst at other times scientists draw upon contemporary literature (for instance, quarks in James Joyce's Finnegans Wake have found their way into elementary particle physics).

Once a term has been accepted, the scientific community enforces its use with considerable vigour and with equal vigour censors certain terms. Recent literature in linguistics, c. 1960-70, avoided the use of the terms behaviour, reinforcement, etc. and instead gave preference to language acquisition device, transformation, and so on.

The coinage and the subsequent use of terminology involves Platonist conceptualisations on the one hand and the use of Aristotelian sense-related data on the other: scientific writing is an archive of such conceptualisations and sense-related data. And, if we were to understand the mechanisms and processes that contribute to good scientific writing, we would not only be able to communicate better, but also be able to discuss deeper questions related to knowledge in a more objective manner.

At the linguistic level, scientific writing can be distinguished not only by the profusion of scientific terms, a large majority of which can be classified as nominal expressions, but by the preponderance of agentless passives (Svartvik 1966), the marked nominalisation of verbs (Halliday and Martin 1993:64), and by the very low frequency of personal pronouns. It is at the lexical level that the text types in this genre can be distinguished from other text types in general language, like newspaper editorials, short stories and novels, personal letters, and so on.

The most frequently-occurring words in any text belong to the closed-class category. Closed-class words have a largely or wholly grammatical role and include articles, pronouns, prepositions, auxiliary and modal verbs, and conjunctions. The membership of the closed-class word-family is fixed or limited; neologisms are not normally added. The open-class category, on the other hand, comprises words whose membership is in principle indefinite or unlimited. New words and senses are continually being added to this set as new inventions, ideas, and so on, emerge. But there are differences in the distribution of these categories in general language and special language texts.
A comparison of the word frequency between a general-language corpus, say, for example, the Lancaster/Oslo Bergen (LOB) Corpus of British English (c.1961), and a specialist corpus of automotive engineering texts compiled at the University of Surrey (see Section 2.3 for details), shows the difference in behaviour of open-class items in the two corpora (see Table 3). The six most frequent words in both corpora are the same closed-class words, comprising over 15% of the total words in each corpus. The first ten words in both corpora are still closed-class words and comprise just under 25% of the total words of each corpus: this, indeed, is a large number, about 1/4 million for the LOB corpus and about 90,000 words for the Surrey corpus.

| Surrey's automotive engineering corpus |
|---------------------------------------|
| 369,751 words                         |

| Word Form | Rank | Relative Frequency (%) | Word Class |
|-----------|------|------------------------|------------|
| the       | 1    | 7.15                   | closed     |
| of        | 2    | 3.34                   | closed     |
| and       | 3    | 2.36                   | closed     |
| to        | 4    | 2.23                   | closed     |
| in        | 5    | 2.10                   | closed     |
| a         | 6    | 1.92                   | closed     |
| is        | 7    | 1.33                   | closed     |
| for       | 8    | 1.09                   | closed     |
| with      | 9    | 0.86                   | closed     |
| on        | 10   | 0.75                   | closed     |
| as        | 11   | 0.68                   | closed     |
| be        | 12   | 0.66                   | closed     |
| are       | 13   | 0.64                   | closed     |
| by        | 14   | 0.62                   | closed     |
| that      | 15   | 0.62                   | closed     |
| emission  | 16   | 0.59                   | open       |
| this      | 17   | 0.58                   | closed     |
| at        | 18   | 0.56                   | closed     |
| engine    | 19   | 0.56                   | open       |
| vehicle   | 20   | 0.51                   | open       |
| system    | 21   | 0.48                   | open       |
| car       | 22   | 0.48                   | open       |
| catalyst  | 23   | 0.46                   | open       |
| it        | 24   | 0.45                   | closed     |
| which     | 25   | 0.44                   | closed     |

| Lancaster/Oslo-Bergen corpus          |
|---------------------------------------|
| 1,013,737 words                      |

| Word Form | Rank | Relative Frequency (%) | Word Class |
|-----------|------|------------------------|------------|
| the       | 1    | 6.74                   | closed     |
| of        | 2    | 3.53                   | closed     |
| and       | 3    | 2.75                   | closed     |
| to        | 4    | 2.64                   | closed     |
| a         | 5    | 2.23                   | closed     |
| in        | 6    | 2.10                   | closed     |
| that      | 7    | 1.12                   | closed     |
| is        | 8    | 1.10                   | closed     |
| was       | 9    | 1.05                   | closed     |
| it        | 10   | 1.04                   | closed     |
| for       | 11   | 0.92                   | closed     |
| he        | 12   | 0.89                   | closed     |
| I         | 13   | 0.75                   | closed     |
| as        | 14   | 0.72                   | closed     |
| with      | 15   | 0.71                   | closed     |
| be        | 16   | 0.71                   | closed     |
| on        | 17   | 0.69                   | closed     |
| his       | 18   | 0.62                   | closed     |
| at        | 19   | 0.60                   | closed     |
| by        | 20   | 0.57                   | closed     |
| not       | 21   | 0.54                   | closed     |
| had       | 22   | 0.54                   | closed     |
| this      | 23   | 0.52                   | closed     |
| but       | 24   | 0.49                   | closed     |
| from      | 25   | 0.46                   | closed     |

Table 3. A contrastive analysis of a special-language and a general-language corpus

Now, if we look at the 25 most frequent words in Table 3, it turns out that for the LOB corpus, the first 25 words are closed-class words and comprise over 30% of the total words of the corpora. However, whilst the first 25 words in the special-language corpus also comprise just over 30% of the texts, there are at least six words that are open class, all nouns, comprising over 3% of the total corpus. All six of these open-class words are potentially key terms of the domain: emission, engine, vehicle, system, car and catalyst. These nouns are as frequent as the important closed-
class words, like *by*, *not*, *had*, etc., in the general-language corpus. The first two open-class words in the LOB corpus do not appear until rank 53 (the reporting verb *said*) and rank 62 (the common noun *time*) respectively.

If we compare the relative frequency of the first six open-class words from the 25 most frequent words in the automotive engineering corpus with their relative frequency in the LOB Corpus, we find that the co-efficient of the relative frequency is some guide to the quantitative differences between special-language texts and general-language texts. The very high values of this co-efficient for some words indicates that, perhaps, these words are used almost exclusively, say, for automotive engineering. The high-frequency open class words identified in Table 3 have a large co-efficient of relative frequency ranging between 16 and infinity. Terms such as *emission* and *catalyst*, together with related terms like *autocatalyst*, *converter*, and *hydrocarbon(s)*, have zero frequency in the LOB Corpus, but a finite frequency in the automotive engineering corpus; hence, the co-efficient of weirdness for these terms, when computed by comparison against the LOB corpus, is infinity (Table 4a).

| Word       | Surrey Automotive Engineering corpus (369,751) | Lancaster Oslo-Bergen corpus (1,013,737) | Co-efficient of Weirdness |
|------------|-----------------------------------------------|------------------------------------------|---------------------------|
|            | Absolute Freq. | Relative Freq. (%) | Absolute Freq. | Relative Freq. (%) | (b/d) |
| autocatalyst| 27              | 0.0073%          | 0              | 0.0000%           | Infinity |
| car        | 1,790           | 0.4841%          | 272            | 0.0268%           | 18.04 |
| catalyst   | 1,700           | 0.4598%          | 0              | 0.0000%           | Infinity |
| control    | 1,517           | 0.4103%          | 199            | 0.0196%           | 20.90 |
| emission   | 2,194           | 0.5934%          | 0              | 0.0000%           | Infinity |
| engine     | 2,083           | 0.5634%          | 70             | 0.0069%           | 81.58 |
| hydrocarbon| 140             | 0.0379%          | 0              | 0.0000%           | Infinity |
| hydrocarbons| 290            | 0.0784%          | 0              | 0.0000%           | Infinity |
| system     | 1,795           | 0.4855%          | 298            | 0.0294%           | 16.51 |
| vehicle    | 1,884           | 0.5095%          | 20             | 0.0020%           | 258.27 |

Table 4a. The preponderance of open-class words in special-language literature (Figures in columns 'b' and 'd' have been rounded up)

In contrast, for the most frequently occurring closed-class words, like *the*, *of*, *and*, *to*, *a* and *in* comprising just under 20% of the LOB Corpus and the automotive engineering corpus, the co-efficient of relative frequency is close to unity (Table 4b). For other closed-class words, like *we*, *what*, and *would*, for example, the co-efficient of relative frequency is far less than unity: it appears that scientists in particular and specialists in general tend to 'suppress' the use of certain closed-class category words (Table 4b). This suppression is as much an idiosyncrasy of the specialist texts as is the preponderance of nominals in such texts: a kind of weirdness, a departure from the norm, a departure from the general language of everyday usage.
Following Malinowski’s (1935) comments on the language of magicians in remote tribal communities, we refer to the co-efficient of the relative frequency of words in special-language text and in general-language texts as the co-efficient of weirdness. This weirdness may be said to give a distinctive texture to scientific writing and in some cases helps the scientists to report the unusual and the esoteric, whilst in others it makes such text rather opaque to a layperson.

2.2 A Typology of Scientific and Technical Writing

The most obvious source of special-language terminology of a domain is the archive of learned texts of the domain: learned papers and advanced text books are amongst the best and most frequent examples of this type of text. Here, one can find the terms in their earliest phase of their life-cycle, that is, when they have just been coined by a scientist and are being introduced to the specialist community for the very first time. Articulate scientists elaborate on their neologisms and are also frequently involved in elaborating neologisms created by others. Elaboration involves either extending or restricting the scope of a neologism.

Advances in science and technology are also reported in popular science magazines, hobby journals and in the news bulletins of learned societies. Another important variant of the scientific writing genre is found in the language used in technical manuals, imperative in style and containing many verb-initial sentences in English. The popular science and the technical manual text are an important source of terminology: these texts may not contain as many neologisms as, say, the learned texts, but may allow access to more mature and more established terms.

Scientific and engineering artefacts are bought and sold, and the vendors of such artefacts rely on advertisements in the news media, in popular magazines, and,
sometimes in learned journals. The advertisement text is yet another source of terminology: here we may find popularised and deprecated forms of terms: *a honeycombed* three-way catalytic converter in learned journals becomes catalytic converter or autocatalyst in popular magazines and in quality newspapers, and is reduced even further to cat con or cat in advertisements.

Public interest in the impact of science and technology means that (quality) newspapers carry editorial comments and news reports on science and technology topics. Apart from the terminology found in newsworthy stories - for instance, the discovery of a cure for a certain type of cancer, the collision of meteors with planets - much of the terminology in the newspaper genre of texts relies on mature terminology. News journalists, on the whole, tend to elaborate the terms they use: acronyms are expanded and display boxes in the middle of the text contain salient terms and their definitions.

One can identify at least six different types of texts produced by or with the help of a scientific or technological community: learned journals and advanced text books and popular science texts, texts designed to inform, and in the latter case, texts designed to entertain as well; technical manuals and introductory text books, texts designed with pedagogic or instructional purposes in mind; and newspaper and advertisement texts, texts that require imagination on the part of both the reader and writer, a kind of semifictional account of a scientific discovery or an artefact.

Specialist texts, like general language text, can be generally divided into two broad (sub-) genres: informative texts and imaginative texts. The designers of the Brown, the LOB and to an extent the other more modern corpora, distinguish between these types of text. The informative genre contains the non-narrative text, texts like essays, government documents, and the learned genre, though this genre was for all intents and purposes popular science texts. The imaginative text was essentially narrative text and comprised fiction, covering the spectrum from detective fiction to science fiction and from adventure fiction to romantic fiction.

Indeed, there are many more text types produced by the specialist community. For example, Sager, Dungworth and McDonald (1980) have discussed 100 different text types, including memos, patents, and so on. Newer text types appear regularly: the poster presentation in learned conferences, essentially mounted displays that require the presence of the author at certain fixed times; written texts containing voice annotations; electronic-mail communications between the members of a specialist domain.

### 2.3 Design of Special-language Corpora

Corpora designed for lexicographical purposes are designed according to author, title, date of publication, length of the text, nationality of the author, language variety, and so on. The literature on corpus-based lexicography includes discussions on how to select the texts so as to minimise the selectors' bias: a distinction is made between a randomly-selected corpus of texts and a corpus that contains deliberately selected texts (by random selection it is meant that texts are chosen from a catalogue of books in print at random). The LOB and Brown corpora are randomly selected corpora, whereas the Birmingham Collection of English Text was selected by a group of linguists,
literature experts and lexicographers (Renouff 1987). The Longman Corpus of Contemporary English is a hybrid corpus in that one-half of the corpus is randomly selected whilst the other half was selected deliberately (Summers 1991). Corpus linguists vary in their choice of the length of texts: the LOB and Brown corpora contain 500 texts of approximately 2,000 words; the Longman corpus has an upper limit of 40,000 words and the Birmingham collection contains texts in their entirety.

Most general-language text corpora of English are contemporary in that the focus is on text published in the second half of the twentieth century. Language variety, especially for monolingual corpora, is sometimes an important consideration. Corpora of English texts are clearly labelled in terms of the national origins of the various texts, like American English, British English, Australian English and Indian English.

Recall that it has been argued that a specialist text corpus would be a good source of neologisms and would be equally effective in showing which terms of the domain are preferred by the domain community and which ones are out of favour. Our focus is on obtaining pragmatic data related to the terms of a domain. This requires that we have a certain pragmatic balance in our corpus. In part, such a balance is achieved by having at least texts of the six different text types mentioned above comprising the informative and the imaginative texts of a specialist domain.

Pragmatic balance can also be achieved in part by making sure that texts used to build a specialist corpus are current: diachronic studies of language apart, there would be little point in looking for neologisms in older texts! Innovations in science and technology seldom have well-defined national or regional boundaries: pragmatic balance requires that the specialist corpus contain texts in two or more national varieties of a given language used by a discourse community. An up-to-date corpus of scientific and technical English texts should contain British, American and Australian English texts. Furthermore, as more and more scientists join the innovative fray and use English as a second language, for instance, Indian English, Euro-English, it is important to incorporate other English variants in a pragmatically-balanced specialist-corpus.

It must be noted that there is a whole range of continuously-growing text for general language, while the number of texts available for a specialist domain is bound to be smaller. In other words, the choice of a terminologist working with special-language text corpora is more limited than that of a lexicographer working with general-language text corpora.

The corpus-minded terminologist, unlike his or her lexicography counterpart, cannot make a decision on the size of the text he or she is going to use: LOB corpus designers restricted the size of text to 2,000 words primarily because of memory restrictions on computer systems; Longman Contemporary Corpus designers restrict the size of the text to 40,000 words. The terminologist cannot easily pass judgement on which text fragment to choose; perhaps there are more terms in the introduction, or perhaps they are in the middle of the text, or perhaps in the conclusion, or the author of a text may have even decided not to use certain terms. The weirdness of specialist texts, that is, the profusion of nominals and suppression of some closed-class words, coupled with the fact that one has more texts to choose from in general language means that the terminologist has little choice but to use full texts.
Lexicographers can pick and choose texts on the basis of the authors of the texts. These authors are well known in the sense that some would have been reviewed in the press, and, indeed, the lexicographer can select texts purely on his or her subjective judgement of the texts. The lexicographer is in a position to select texts for building lexica, because he or she can appreciate, analyse and understand general language texts without much help from others. However, such freedom of choice is not available to a terminologist who can neither appreciate nor analyse/understand specialist texts in the way a domain expert may do. Seldom does a terminologist have any formal knowledge of a specialist domain, and indeed, even if he or she knows something about a domain, it is quite probable that the next terminology assignment for a terminologist would be in a specialist area of which the terminologist knows very little. The terminologist either has to rely on a domain expert for selecting terminologically relevant texts or just has to take any text that becomes available.

2.4 Pragmatic Balance and the Design of Specialist Corpora

The Surrey automotive engineering corpus is one of the ten special-language corpora developed at the University of Surrey for building term banks and for studying special-language texts. Roughly the same number of words are available for German and for Spanish. These texts contain a mixture of British and American English texts, otherwise all others are British English texts. Table 5 contains the details of the English texts in the automotive engineering corpus.

| Text Type               | Total No. of Texts | Total No. of Tokens | Average | Maximum | Minimum |
|-------------------------|--------------------|---------------------|---------|---------|---------|
| Popular Science         | 6                  | 19,700              | 3,283   | 5,607   | 731     |
| Journals                | 50                 | 203,079             | 4,062   | 19,785  | 154     |
| Manuals                 | 7                  | 36,027              | 5,147   | 8,116   | 2,027   |
| Books                   | 5                  | 31,105              | 6,221   | 14,430  | 755     |
| Advertisements (inc. Brochures) | 38        | 42,749              | 1,125   | 2666    | 184     |
| News Reportage and Features | 24              | 37,091              | 1,545   | 5,884   | 206     |
| Grand Total             | 130                | 369,751             | 2,844   | 19,785  | 154     |

Table 5: Details of the English section of the Surrey's Automotive Engineering Corpus

These texts were collected specifically for the purpose of building a multi-lingual terminology data bank, containing terms for emission control and anti-lock braking systems in two varieties of English (British and American), Spanish and German, for use in translation and in technical documentation. The objective of building such a specialised term bank was to initiate and evaluate the efficacy of corpus-based methods and tools in terminology management: this effort was sponsored by the European Commission's ESPRIT Programme (Ahmad et al 1992, 1993).

An analysis of the distribution of words in the different text types of the corpus shows that there was a degree of uniformity in their distribution so far as the co-efficient of weirdness is concerned. Table 6 shows that despite the differences in text types, authors of specialist texts, whether they are scientists, technical authors, specialist correspondents for newspapers and popular magazines, or advertising copy writers, show similar tendencies in using some categories of words more frequently than others when compared with general-language texts.
Table 6 shows that all six types contain, on average, 9-10% words that do not exist in a contemporary corpus of English, like the Longman Corpus: the co-efficient of weirdness for such words is infinity. Again, over 10% of the words are used at least 10 times, or as much as 100 times, and in some cases 5,000 times, more in a specialist text than in a general-language text. Furthermore, Table 6 shows that over a third of words in specialist texts are used at least 1.5 times more, and in some cases as many as 10 times more, than in general-language texts. These words are likely to be closed-class words and a linked set of open-class words. There is only a small proportion of words, totalling about 12%, that have roughly the same frequency distribution in the automotive engineering corpus as in the Longman Corpus: the co-efficient of weirdness of such words, generally determiners and modal verbs, is one. But pronouns and a number of nouns of everyday use like society, people, and child are avoided almost completely in science and technology texts, and, surprisingly, in the advertisement, news reportage and features, and popular science genres. (Note that the genre manual has its own characteristic behaviour, though still within the norm of the corpus as a whole, and shows a preponderance of open-class nominals and suppression of closed-class items like pronouns, etc.)

### 2.5 A Note on the Choice of General-language Corpus for Computing Weirdness

The computation of the co-efficient of weirdness depends upon the availability of frequency information related to a relevant corpus of general language. By relevant we mean a general-language corpus that will help in the extraction of hitherto unidentified terms, particularly neologisms. Once used in science and technology, terms which have some impact on our everyday lives become absorbed quite quickly in general
language. The LOB corpus does not have any instances of words like emission, catalyst, catalytic converter and electronic. However, all these words can be found in the Longman-Lancaster Corpus and the dictionaries that are associated with these corpora. These once-neologisms still occur with much lower frequency in contemporary general language corpora than, say, in a specialist text. Thus, a relevant general-language corpus should be as near contemporaneous as possible for identifying neologisms. The computation of the co-efficient of weirdness, through the use of a contemporaneous corpus, will help to avoid spending too much time investigating a potential term which may already have had some currency.

Note that the general-language corpus used for comparison presented in Table 6 was the more contemporary Longman-Lancaster Corpus. The bulk of the text in this 20 million word corpus was published during 1968-85; the smaller Lancaster-Oslo Bergen Corpus (1 million words), which was compiled in 1961, is more outdated, the bulk of the text having been published during the 1940s and 1950s.

In computing weirdness, we also looked at the question of genre balance in the general language corpus. Therefore, we focused on a corpus that has as wide a coverage of the written genre as possible and attempts were made to minimise selectors’ bias: recall that half of the Corpus is randomly selected and the other half was the choice of a panel. Longman-Lancaster, mainly a British English Corpus, has coverage of American English also. And, despite the upper limit of 40,000 tokens on all texts, Longman-Lancaster contains excerpts from long narrative texts together with the kaleidoscopic newspaper texts. Above all, Longman-Lancaster is available to all academic institutions at a reasonable price.

3. Acquiring Terminology from a Text Archive

A text-based approach to terminology can be of help to a terminologist not only in semi-automatically identifying terms, through the computation of the co-efficient of weirdness, but also by making available other kinds of text-derived data. By viewing text fragments, containing keywords-in-context (KWIC), a terminologist can deduce a variety of syntactic, semantic and pragmatic details that can be found much more easily than by manual scanning methods, introspection or interrogation of domain experts.

Using such methods, we have been able to create term banks in at least ten different subject fields, ranging from drug addiction to automotive engineering, and from information technology, including artificial intelligence, to environmental protection. This work has been supported by 'System Quirk'\(^{10}\) (Holmes-Higgin & Ahmad 1992), an intelligent terminology and lexical development system.

3.1 System Quirk: A Terminology and Lexicography Support System

System Quirk is essentially an integrated set of programs or software 'tools' for examining and extracting relevant material from evidence sources, such as an organised special-language text corpus, and for creating, deleting, modifying and maintaining a reference source such as terms in a term bank. There are tools for dealing with each phase of a term's life-cycle.
The literature in corpus linguistics, particularly in corpus-oriented lexicography, contains descriptions of software tools that are used for gathering data about lexicogrammatical properties of words. Leech has discussed the need for developing at least three different types of software tools that comprise a 'sophisticated computational environment' for retrieving data from a corpus and for processing linguistically the corpus itself. Leech's specification includes (i) general-purpose data retrieval tools, (ii) tools to facilitate corpus annotations at various levels, and (iii) tools to provide interchange of information between corpora and lexical and grammatical databases (1991:22-23).

System Quirk contains general-purpose data retrieval tools and tools for exchange of information between corpora and lexical (and terminology) databases. (System Quirk does not contain any corpus annotation tools, but it is capable of importing and exporting texts encoded in SGML format and terminology in a number of terminology interchange formats1). In addition to the text (and term) analysis tools and corpus and term bank organisation tools (see Table 7 for details), System Quirk contains the so-called visualisation tools. These tools can be used for elaborating a term, selectively browsing a text corpus or corpora, and tools for visualising the inter-relationships between terms. The visualisation tool has some facility for deducing new facts from old, through the use of knowledge representation formalisms, and a facility for identifying semantic relations based on linguistic cues. Table 7 below shows the functional characteristics of the tools:

| Analysis Tools | Lexical/Term Analysis |
|----------------|-----------------------|
| Text Analysis  | Relationships with other lexical items, Foreign Language Equivalents. |
| Concordance, Collocation, Statistical Analysis, Term identification. | |

| Organisational Tools | Organisational Tools |
|----------------------|----------------------|
| Corpus Organisation  | Term Bank Organisation |
| Classification and Representation of full text units. Organisation along pragmatic lines. SGML mark-up. | Creation, maintenance, and quality control of term banks. Accessing other term banks. TIF mark-up. |

| Visualisation Tools | Visualisation Tools |
|---------------------|---------------------|
| Selective Explication | Illustrative Explication |
| Access within and across corpus. Goal-oriented browsing. Selectional constraints on fragments. | Selection of illustrative text fragments - contextual examples. Use of semantic nets for illustrating inter-term relations. Publishing tools. |

**Table 7**: Functional characteristics of the System Quirk toolbox

### 3.2 Text Analysis and Term Identification

The work-horse of System Quirk is the concordance and text analysis (sub-)program KonText. KonText is used in the first stage of terminological work following document selection, i.e., it is used to search the corpus, or sections of it, initially in order to capture term candidates.
Four basic operations can be performed on the texts selected: concordance (an alphabetical list of all the words in a text shown together with their context and reference to lines in the source text); collocation (a list of the co-occurrences of specified terms within sentence boundaries); wordlist (an alphabetical or frequency-sorted list of words); word index (as wordlist with references to lines in source texts). These operations can be refined and customised by a series of further options (A detailed discussion has been presented in Ahmad and Rogers forthcoming).

The KonText module can be used to compute the co-efficient of weirdness to identify terms. The computation is dependent on the four tasks performed by KonText. First, the relative frequency of each word form in the selected special-language texts is computed. Second, KonText provides the choice of a general-language corpus. Third, the relative frequency of the word forms in the selected special-language texts divided by the relative frequency of the same word forms in a general-language corpus specified by the user is computed (known as ratio). And, fourth, a facility to define the ratio level is given. The results for one such computation involving the execution of all the four tasks are shown in Figure 2.

Computing the ratio of word forms in special-language and general-language texts allows a provisional distinction to be made between general-language open-class words on the one hand, and special-language open-class words on the other hand, i.e. term candidates. For instance, in Figure 2 below, the word form company is shown occurring 18 times in the selected information technology text, while the word form configuration occurs only once. When the ratios are computed, however, we see that they are <10 and >10,000 respectively, indicating configuration as a term candidate, but not company. The particular ratio level at which term candidates can be optimally identified needs to be further explored in relation to a number of factors including the type of general-language corpus used as a comparison, and the type of domain covered by the special language.
4 Term Pragmatics and Terminology Organisation

During a translation session or a technical writing session, a typical translator or technical writer may need only a limited subset of data from the record format of a term in a term bank: a term and its foreign language equivalent, for instance. Furthermore, the translator or technical writer only needs to consult terms which he or she seldom or never uses: some terms may never be looked up, others could be consulted very frequently. The organisation of a term bank should, in principle, reflect this need of its end-users.

Our intention is to identify and implement a machine-assisted procedure that will use four pragmatic properties of a term, that is, language variety, style, frequency of the term in a text archive, and the age of the term, to partition a term bank into two parts: a mother term bank and its associated daughter term bank. The former containing terms with which technical writers or translators may be very familiar, and the daughter term bank containing terms with which the technical writer or translator may not be familiar.

Using these four pragmatic properties, an *ad hoc* score can be computed automatically, whereby each property contributes a partial familiarity score and all the partial scores

![Table](image)

Fig. 2. Typical results from an information technology corpus showing all single word forms with a co-efficient of weirdness in excess of 10,000
are added up to give a total familiarity score for each term. For instance, terms that occur very frequently in a given specialism, were coined in a widely-recognised geographical/regional variety, and have been in use over a considerable period of time in popular science texts (or newspaper texts) would not be looked up frequently by translators or terminologists. Such terms will have a high familiarity score. On the other hand, terms that are 'unfamiliar', in that they had been coined recently, reported in an esoteric learned journal by an author who is more familiar with a lesser used variety of, say, English, will have a low familiarity score.

The 'familiarity' scores can be used as a basis for dividing a term bank into the mother and daughter components, such that the access to daughter terms is quicker than, say, the mother terms. Later, in Section 4.2 we report on a software tool, Lexicon Distiller, a component of System Quirk, that can be used to divide a term bank into mother and daughter components.

4.1 Computing the Pragmatic Familiarity Score for a Term

We will show how the partial familiarity score for each of the four pragmatic factors associated with a given term can be computed automatically. An arbitrary scale ranging from +2 to -2 has been assigned to each of the properties. The partial score +2 indicates that for a given pragmatic property the term is familiar and a score of -2 means that on that particular pragmatic basis the term is unfamiliar.

For instance, a term would be assigned to the mother term bank if the total familiarity score is +8, that is, each of the four partial familiarity scores is +2, and to a daughter term bank if the familiarity score is -8. For intermediate scores, a policy can be laid out by the administrator of a given term bank and a threshold familiarity score can be assigned: terms above the threshold, any positive number between 0 and 8, will be assigned to the mother term bank and below that threshold can be assigned to the daughter term bank.

Language Variety Terminology collections do acknowledge regional variations: silencer in British English is equivalent to muffler in the American English variety. This type of information is included in the record format for most typical term banks. The terminologist or technical writer will, perhaps, be more familiar with terms in the native language variety, or some varieties that are closer to the terminologist's or translator's variety. In Figure 3a we show how a partial familiarity score can be computed, if it is assumed by a term bank's administrator that the term bank users are British English speakers:

| Partial Familiarity Scores | +2 | +1 | 0 | -1 | -2 |
|----------------------------|----|----|---|----|----|
| Language Variety           | British | American | Australian | Indian | Japanese |
| English                    | English | English | English | English | English |

Fig. 3a. Partial Familiarity Score based on language variety

Regional variants are usually noted in the record formats of extant term banks.
**Style Register** Translators and technical writers usually deal with 'learned' texts, including learned papers and text books, and with technical manuals. Such texts are characterised by their formal style. The *formality* shows itself *inter alia* in the usage of standardised terms, in the avoidance of variants, and in the absence of popular equivalents of standardised terms. However, popular science material, advertisements and brochures, contain popularised terms: for example, *cat* for *catalytic converters*.

Figure 3b shows how partial familiarity scores can be computed for individual terms in a term bank. The assumption here is that the term bank user usually translates learned text or technical manuals.

| Partial Familiarity Score | Style Corruption/ Register | Popular | Deprecated | Abbreviation | Standardised |
|---------------------------|---------------------------|---------|-------------|--------------|--------------|
|                            |                          |         |             |              |              |

**Fig. 3b. Partial Familiarity Score based on style**

Most term bank record formats contain some *stylistic* data for a number of terms.

**Frequency of occurrence** Some terms are more frequent in the text of a specialist domain than others: in genetics and in neurology, the term *cell* will be amongst the most frequent, in that the term is usually a part of a compound nominal, like *Purkinjee cell, Pyramidal cell* or may appear as one of its morphological variants, e.g., *cells* (plural), *cellular* (derived adjective). The same is true of other *carrier* terms, like *nucleus* in nuclear physics; *force, stress, strain* in material sciences, and so on.

One can argue that if a term, single or complex or multiword, occurs more frequently than other terms in, say, a corpus of special-language texts, then the translator or technical writer will either know it already or will subconsciously memorise the elaboration of such a term. In contrast, the terms that occur with very low frequency in a special-language text archive, for example, terms used in the reporting of a new device or a novel concept, will be the terms with which the translator or technical writer will not be familiar.

Recall that the most frequently occurring words in a specialist corpus (and in a general-language corpus) are the closed-class words, and that these words have a relative frequency of between 1-5%. In special-language corpora, the closed-class words are followed by terms, usually nouns or noun compounds, which have a relative frequency of less than 1%. Following the closed-class words and the more frequent terms are terms that have lower frequency, ranging between 1 in 1000 (relative frequency of 0.001%) to one in 10 million words (relative frequency of 0.00000001%).

Now on the basis of relative frequency of words in a given corpus, a partial familiarity score can be computed. This computation involves taking the logarithm of the relative frequency and adding the number four (4) to the logarithm. Thus, any term occurring with a frequency of 1 in 100 (relative frequency of 0.01 or 10⁻²) in scientific
notation), will have a partial familiarity score of +2 and a term which occurs with a frequency of less than or equal to 1 in 10000 (relative frequency of 0.0001 or $10^{-4}$) will have a partial familiarity score of 0, and if the frequency of the word is less than or equal to 1 in 1 million (a relative frequency of $10^{-6}$ or less) then the partial familiarity score will be -2. Figure 3a shows how the partial familiarity score is related to the relative frequency of a term.

The results of this computation is rounded up to the nearest integer value. Furthermore, if the result of the computation is greater than +2, then a value of +2 is used. Similarly, if the result is less than -2, then a value of -2 is used.

It should be noted here that since most term banks are developed without reference to a text corpus it is not usual to note the frequency of a term in typical record formats used by existing term banks.

**Life-cycle details** Term banks are, or rather ought to be, updated at regular intervals and new terms added at a greater frequency than, say, to a conventional general-language dictionary. The progress of a scientific discipline is characterised not only by the introduction of new terms, but also by a gradual obsolescence of established terms. Translators and documentors will be more familiar with terms that are established and in use for 10 or more years. However, terms that have entered a specialism more recently, for instance in the last decade or so, have to be established even though their use may not be widespread. Well established terms can then be assigned a higher partial familiarity score, and less established terms can be assigned lower partial familiarity scores, and be stored in the daughter term bank. Figure 3d shows how partial familiarity scores can be computed by looking at the age of a term:

Existing term banks do not directly record the 'age' of a term. However, a note is made of when the term was entered into a term bank.

**4.2 Towards a 'Pragmatic' Basis for Organising Term Banks**

As mentioned above, we will regard a term with a score of -8 as very unfamiliar and a
term with a score of +8 as very familiar, with a gradable scale in-between. Consider the following automotive engineering terms, shown in Table 8, together with their familiarity scores that may be assigned for, say, a British English speaking translator or documentor: the figures in brackets show our devised score based on the polar scales mentioned above.

| Term        | Relative Frequency in Text | Language Variety | Style Register | Life-cycle History/ Age of the term | Familiarity Score |
|-------------|---------------------------|------------------|----------------|------------------------------------|-------------------|
| catalyst    | 0.46 (+2)                 | British English (+2) | Standardised (+2) | Well-established (+2)                  | 8                 |
| catalytic converter | 0.00108357 (+1)            | British English (+2) | Standardised (+2) | Maturing                             | 5                 |
| silencer    | 4.7719E-05 (-1)           | British English (+2) | Standardised (+2) | Well-established (+2)                  | 5                 |
| cat         | 0.00011831 (0)            | British English (+2) | Popular (-1)     | Maturing                             | 1                 |
| muffler     | 0 (-2)                    | American English (+1) | Popular (-1)     | Well-established (+2)                  | 0                 |

Table 8. Familiarity points score, shown in bold italics, for a number of terms from the automotive engineering (AE) domain. (Frequency data from Surrey's AE corpus)

Suppose now that we choose a familiarity threshold of +6; terms that have a score of less than 6 will be stored in the daughter term bank, and all other terms will be stored in the mother term bank. Thus, terms like catalyst, catalytic converter and silencer will be stored in the mother term bank, and terms like cat and muffler in the daughter term bank. The arguments relating to the life-cycle history of a term need to be looked into much more carefully in future research.

Lexicon Distiller, a recent addition to System Quirk12, helps in exploiting the pragmatic properties of a term for assigning it to a mother or to a daughter term bank (Figure 4).
The Lexicon Distiller allows options for choosing a variety of filters for assessing a familiarity points score based on the pragmatic properties of a term such as its language variety, style register, frequency of usage, and its life-cycle history or age. The administrator of a term bank has the option of selecting, or filtering, a certain number of terms to a new term bank. Once the user of Lexicon Distiller has worked on creating part or whole of a daughter term bank, then the administrator can store them in a daughter term bank (see Figure 5).

5 Conclusions

The above discussion of the pragmatics of specialist terms, particularly its role in extracting terms from texts and efficiently organising the terms, shows the effectiveness of corpus-linguistic methods and techniques in terminology. We feel that such methods and techniques are in some sense more relevant to terminology than, say, to lexicography. Questions relating to the authenticity of terms can be settled by seeking the opinions of experts, and since these opinions are more frequently expressed in text, then special-language text is arguably a more important source, and perhaps sometimes the only source of evidence for the existence or non-
existence of a term. The representativeness of texts, a vexed issue in lexicography and language studies, is a less ideologically-sensitive issue in terminology, since special-language texts do not usually involve subjective questions related to economic and social class-differences.

Terminology can benefit from being viewed and treated in the context of pragmatics, corpus linguistics and other paradigms in linguistics and in philosophy dedicated to the study of text for investigating problems related to knowledge of language and knowledge of matters scientific and technological.

The discussion in this paper was based on extracting and elaborating terms in English. Corpus-linguistic methods and techniques appear as equally effective for related languages like Dutch, and typologically different languages like Welsh (Ahmad and Davies, in press).

Acknowledgements

The author would like to thank Dr. M Rogers for her comments and criticisms on this paper, Paul Holmes-Higgin and Syed Sibte Raza Abidi who wrote the Lexicon Distiller program and Andrea Davies who helped in the debugging of the program and helped with the layout of this paper. Caroline McInnes was as usual patient with my handwriting and reminded us all of the various deadlines: her efforts deserve many thanks. Petra Steffens used her wit and editorial charm to persuade the author to produce this paper almost in time, for which I am very grateful.

References

Ahmad, Khurshid & Davies, Andrea E. (Forthcoming) "Weirdness" in Special-language Text: Welsh Radioactive Chemicals Text as an Exemplar. Internationales Institut für Terminologieforschung Journal.

Ahmad, Khurshid & Rogers, Margaret A. (Forthcoming) 'The analysis of text corpora for the creation of advanced terminology databases'. In (Eds.) Sue-Ellen Wright and Gerhard Budin The Handbook of Terminology Management. Amsterdam and Philadelphia: John Benjamins.

Ahmad, Khurshid & Rogers, Margaret A. (1992) Translation and Information Technology: The Translator's Workbench, in ReCALL, No. 6: 3-9.

Ahmad, Khurshid, Davies, Andrea, & Rogers, Margaret A. (1993) The Contrastive Analysis of Special-language and General-language text: Towards Automatic Term Identification. University of Surrey: Guildford. CS-93-12.

Aijmer, Karin and Altenberg, Bengt (Eds.) (1991) English Corpus Linguistics: Studies in Honour of Jan Svartvik. London and New York: Longman Publishers.

de Beaugrande, Robert & Dressler, Wolfgang (1981) Introduction to Text Linguistics. New York: Longman.

Crystal, David (1992). A Dictionary of Language and Linguistics. Oxford: Blackwell.
Geertz, Clifford (1988) Work and Lives - The Anthropologist as Author. Stanford (California, USA): Stanford University Press. (Also marketed in the UK under the imprint of Polity Press/Blackwell Publishers, Oxford).

Halliday, Michael A.K. & Martin, James R. (1993) Writing Science: Literary and Discursive Power. London and Washington D. C.: The Falmer Press.

Halliday, Michael A. K. (1978) Language as Social Semiotic: the social interpretation of language and meaning. London: Edward Arnold.

Harris, Zellig (1990) A Theory of Information and Knowledge. Oxford: Oxford Univ. Press.

Heisenberg, Werner (1963) Physics and Philosophy - The Revolution in Modern Science. London: George Allen and Unwin.

Heisenberg, Werner (1983) Encounters with Einstein (And Other Essays on People, Places and Particles). Princeton (New Jersey, USA): Princeton University Press.

Holmes-Higgin, Paul & Ahmad, Khurshid (1992) Knowledge Processing & Machine-Assisted Terminology Elicitation. Guildford: University of Surrey. CS Report No. CS-92-14.

ISO/R 1087 (1969) The Vocabulary of Terminology. Geneva: International Organization for Standardization.

Kittredge, Richard & Lehrberger, John (Eds.) (1982) Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin: de Gruyter.

Landau, Sidney I. (1989) Dictionaries: The Art and Craft of Lexicography. Cambridge (UK): Cambridge University Press. (First published 1984).

Leech, Geoffrey (1991) 'The state of the art in corpus linguistics'. In (Eds.) Karin Aijmer and Bengt Altenberg (1991) pp 8-29.

Malinowski, Bronislaw (1935) Coral Gardens and their Magic 2. London: Allen and Unwin.

Renouff, Antoinette (1987) 'Corpus Development'. In (Ed.) John McH. Sinclair (1987) pp 1-40.

Roda, Roberts P. (1992) 'Exploring the frontiers of documentary research in the translation process'. In (Ed.) Edith F. Losa. Frontiers: Proceedings of the 33rd Annual Conference of the American Translators Association. (November 4-8, 1992, San Diego, California). Medford (New Jersey, USA): Learned Information, Inc., pp 361-375.
Sager, Juan C., Dungworth, David and McDonald, Peter F. (1980) *English Special Languages*. Wiesbaden: Brandstetter Verlag.

Sinclair, John McH. (Ed.) (1987) *Looking Up*. London: Collins.

Sinclair, John M. (Ed.) (1991) *Corpus. Concordance. Collocation*. Oxford: Oxford University Press.

Summers, Della (1991) *Longman/Lancaster English Language Corpus. Criteria and Design*. Unpublished ms.

Svartvik, Jan (1966) *On Voice in the English Verb*. Mouton: The Hague.

Footnotes

1 It has been estimated that up to 40% of the content of 'unabridged unilingual general dictionaries' is devoted to specialised terminology (Roberts 1992). However, general-language dictionaries usually include only the well-established terms, especially those terms that may appear in school text books.

2 Computer scientists, including software engineers and the artificial intelligence community, are involved in the creation of *ad hoc* term collections: a data dictionary used in the specification and design of a software system contains substantial amounts of unstructured knowledge, and those involved in knowledge engineering get much closer to building term banks that comprise the building blocks of the so-called knowledge bases of expert systems. This is not to argue that the computer scientists should use currently available term banks as these data banks have been designed for translation or technical documentation. But the data available in such term banks can be of considerable use for building data dictionaries and for building fact bases of a knowledge base (Ahmad 1993).

3 Scientific writing, a sub-genre of specialist writing, is on its way to becoming a well-established genre that has its own publications and, indeed, like general literature has its own prizes for distinguished writing.

4 The literature concerning terminology is sharply divided into two major groups. For one group, concepts are the primary source of terminology, where conceptual hierarchies and part-whole relationships are the organising frameworks for building collections of terminology. For the other group, sense-related data is of paramount importance and the imposition of the hierarchies and part-whole relations perhaps mitigates against the creative use of language. This debate has its roots in the Platonist versus non-Platonist basis of knowledge and of language.

5 See, for instance, observations on the language of physics by Werner Heisenberg (1963, 1983), and on the language of anthropology by Clifford Geertz (1988). These authors discuss philosophical issues connected to their discipline by looking at the vocabulary and the discourse structures of the texts of their respective disciplines.
Very similar results are also found in more modern corpora such as the Lancaster/Longman Corpus (c. 1900-1985) and the Birmingham Collection of English Texts (c. 1965-1985).

News bulletins, as opposed to learned journals, carry news about the members of the society and news about those who can or who have had influence on the society, etc.

The real reduction is the use of the drawing of a cat's eye in advertisements for automobiles that are fitted with catalytic converters.

The exceptions include the Helsinki Corpora of Historical English that for obvious reasons contain texts produced during different periods of this millennium.

System Quirk is written in Quintus-Prolog, a logic programming language available on a SUN-SPARC station, running the UNIX operating system. Paul Holmes-Higgin, together with Stephen Hook, Stephen Griffin and Syed Sibte Raza Abidi, developed this program for the ESPRIT-sponsored Translator's Workbench Projects (1989-1992; 1992-94) and MULTILEX (1990-93).

Corpus annotation is important for automatically analysing corpora with a view to determining the grammatical texture of tokens, texts and indeed text corpora. Grammatical tagging, a popular method of annotating texts, relies on the foreknowledge of the grammatical properties of all the tokens that may exist in a text. Our concern is identifying a neologism, and if a neologism has just been coined and reported in literature, then perhaps little would be known about its grammatical properties.

Lexicon Distiller was written by Syed Sibte Raza Abidi and Paul Holmes-Higgin.