[From: *Translating and the Computer 14*. Papers presented at a conference… 10-11 November 1992 (London: Aslib 1992)]

TERMINOLOGY MANAGEMENT: A CORPUS-BASED APPROACH[1]

Khurshid Ahmad* & Margaret Rogers**
*Department of Mathematical & Computing Sciences & **Department of Linguistic & International Studies, University of Surrey, Guildford, Surrey, GU2 5XH, U.K.

The notion of terminology management embraces term capture, term elaboration, term storage, term retrieval, term updating, and term dissemination. Arising from our work in 'managing' terminology for translators, we outline some aspects of terminology management in the context of a descriptive text-based approach. An integrated set of software tools covering all aspects of the terminology management cycle is briefly described. Methods are discussed for improving the degree of support given to the user in the identification and elaboration of terms from text.

TERMINOLOGY MANAGEMENT

The majority of translators specialise in particular subject domains in their professional work, forming a crucial link in the chain of multilingual communication at many different points: between experts of the same and different domains, between experts and technicians, between experts and hobbyists, between marketing specialists and consumers, between a company and its shareholders, and so on. In so doing, translators require a familiarity not only with the domain, but also with the linguistic variation which is associated with each level of scientific and technical communication. One of the features which varies is that of terminology - the vocabulary of a domain which is uniquely associated with that domain and which evolves, matures and becomes obsolescent (or 'dies') with it.

Information in specialist technical dictionaries - monolingual and bilingual - is frequently inadequate for translators' needs in a number of ways. For example, information on situational context (who is speaking to who?) and on linguistic context (how is this term typically used?) occurs rarely, if at all; if indicated, semantic relations between terms in the same field are often buried in the lexicographic symbols of cross-references; and the treatment of highly productive compounds is problematic. Consequently, most translators compile their own 'glossaries', often still on paper (e.g. index cards), or, if in the electronic medium, in ordinary text files, or possibly, in a database or one of the proprietary terminology data bases (Hainitz and Pownall (1) (2)). The information which they store is hard won and often gleaned from disparate paper-based sources such as mono- and bilingual specialist dictionaries, encyclopaedia, text books, journal articles, newspaper articles and notes taken during conversations with experts. Whichever medium the translator has chosen for his or her glossary, paper or electronic, information needs to be located, sifted, copied and, finally, represented in such a way that it is reusable on future occasions. Optimally, the translator needs to ensure that this information is accurate, linguistically informative in a context-dependent way, consistently represented, and crucially, retrievable according to need. Present commercially available computerised

terminology storage systems do not match these needs for the cycle of what can be called 'terminology management'.

Faced with the ever-increasing specialisation of knowledge, how are translators to cope with the resulting terminological explosion as well as their own wide range of terminological needs? In this paper, we report on a methodology, that is, methods, techniques and (computer-supported) tools or programs, which will enable a terminologist or a translator to manage terminology resources, particularly terminology data banks and specialist texts where terms are actually used. The methodology encompasses term capture, term elaboration, term storage, term retrieval, term updating, and term dissemination. The techniques are drawn from corpus linguistics, conventional (normative) terminology theory and practice. The computer-supported tools help in the organisation, dissemination and upkeep of the terminology resources, term banks and specialist corpora. The objective is to provide an integrated support system for the translator or the dedicated terminologist which achieves optimal interaction between human user and machine in the utilisation of a collection of real texts, i.e. a 'corpus', for the compilation of special-purpose vocabularies.

In this paper we describe results largely based on our participation in the Translator's Workbench Project (TWB). TWB is a project funded by the European Commission ESPRIT II Programme (European Strategic Programme for Research and Development in Information Technology). The first phase of the three-year project has been successfully completed (1989-92). TWB-II, a follow-up two year project (1992-94) intends to bring the results of the research project to the 'market place'. The aim of the project is to develop and integrate a set of computer-based multilingual text processing tools for language professionals, particularly translators. The following organisations are participating in the project: Olivetti Office/TA Triumph Adler AG (Germany), Mercedes-Benz AG (Germany), L-Cube (Greece), Siemens AG (Germany), Siemens SA (Spain), Fraunhofer Gesellschaft (Germany), University of Surrey (United Kingdom), University of Heidelberg (Germany), Universitat Politecnica de Catalunya (Spain), University of Stuttgart (Germany). The Translator's Workbench will provide an integrated set of software tools which will help to ensure the conversion of documents to and from a number of EC languages in a grammatically, stylistically and terminologically correct and consistent manner. Translators will be provided with multilingual text processing facilities, including a term bank and term bank building tools, grammar, style and spelling checkers, semi-automatic translation help systems, and remote access to a machine translation system (METAL) and to other term banks (e.g. EURODICAUTOM).

CORPUS-BASED TERMINOLOGY

A corpus can be simply defined as a collection of naturally-occurring texts, nowadays in machine-readable form. Corpora vary in their design, i.e. according to the criteria by which the texts are selected. These criteria include: spoken/written; whole texts/samples; size, text type, time span, language/s, language variety/ies, and, if not a general-purpose corpus, domain. While texts may be stored 'raw' without metalinguistic analysis (or 'annotation') of any kind, various types of annotation are

possible from structuring with SGML (Standard Generalized Markup Language) to full grammatical tagging. In these cases, processing may take place at a more abstract level, such as according to word class. *As* a structured archive or record of communications about a particular domain, a corpus, even without annotation, is a potentially rich source of evidence about the terminological problems which translators encounter, not least those of variation.

Terminological variation has a number of dimensions: examples include 'horizontal' and 'vertical' (3), geographical or regional, company-specific, orthographic, and diachronic. For example, horizontal variation in terminology distinguishes one domain from another. Less salient is vertical variation, which often distinguishes levels of communication according to interlocutors, or loosely, level of expertise (e.g. expert - technician - workshop - layperson).

As recorded in the textual archive of any domain, such variation can be viewed from different perspectives. For the terminologist concerned with standardisation (the elimination of synonymy and the reduction of homonymy) in the interest of unambiguous professional communication, variation is something to be eliminated, or at least reduced. For the linguist, variation is a natural phenomenon, to be investigated, described and explained in a sociolinguistic framework. But for the translator, terminological variation is a problem which has to be solved; translators must deal with terms in their full variation and in the context of running text. This variation cannot in the case of translators be dismissed as 'usage' and therefore of no interest. In fact, such variation is of central interest, and texts, particularly if they are structured in a corpus, are a valuable source of information about many kinds of variation.

Leech (4) provides a useful overview of the development of corpus linguistics. Early examples can be found in the work of structuralist linguists such as Zellig Harris (comprehensively reported in Harris (5)). The data-driven approach was consistent with the contemporary linguistic orthodoxy preceding the mentalist revolution initiated by Harris's student Chomsky in the mid-1950s. Chomsky's competence-performance distinction led to a rejection by some linguists of language samples as unrevealing about human knowledge of language, or 'competence'. Nevertheless, despite the change in prevailing ideology, two major corpus projects were conceived in the late 1950s/early 1960s: by Quirk in the UK (the Survey of English Usage) and Francis and Kucera of Brown University in the USA. The London-Oslo/Bergen (LOB) corpus mirrored the Brown Corpus for British English. These early corpora each contained one million words. Brown and LOB were comprised of 500 text extracts of 2000 words from a range of 'genres'. Thirty years later, 100 million words has been cited as a goal by emerging projects (the Data Collection Initiative of the Association for Computational Linguistics for American English; the British National Corpus).

The corpus-based analysis of natural language has many potential applications, including lexicography, language teaching and learning, machine translation, text critiquing, text synthesis, and the creation of linguistic databases. The lexicographical application has become well known through the Collins Cobuild Project (Sinclair (6)), which has resulted in a range of general-purpose dictionaries, including collocations

and phrasal verbs, and a range of English Language Teaching books, based on the evidence of the Birmingham Collection of English Texts. Other major English-language publishers such as Longman and Oxford University Press now also work with corpora. While general-purpose lexicographers have traditionally worked descriptively, starting their analysis from the linguistic sign (the word), the tradition of terminological analysis predominant in Europe has been concept-based and normative (Wüster (7); Felber (8); Picht and Draskau (9)), reflecting the concerns of technologists and scientists for standardised terminologies. The use of corpus-based evidence in special-purpose lexicography, certainly in the European tradition, has therefore not yet received general acceptance, since it is in the first instance word-based and descriptive. However, since the terminological work carried out for the Translator's Workbench Project is primarily focussed on the translator as the end user, then in our work at Surrey we have adopted a text-based approach. The University of Surrey automotive engineering corpus is trilingual (English; German; Spanish) and is structured according to five text types and three sub-domains (defined according to need by the end user of the terminology which was produced from the evidence of this corpus, Mercedes-Benz AG Language Services Department) (current size: 831,553 words).

However, special-purpose multilingual corpora can provide evidence not only on linguistic variation for the translator, whose work is text-based, but also for the terminologist, whose objectives may be normative. If a corpus is well-designed, then the user will be able to select both the range of texts to be processed and the method of processing. For purposes of standardisation, for instance, a frequency count of synonyms can provide useful distributional evidence indicating statistically preferred terms.

Other special-purpose corpora are currently being added to the original automotive engineering corpus (e.g. hydrology; knowledge engineering; linguistics). The corpora are managed using a dedicated software tool which automatically assigns headers to new texts, and allows the user to define a hierarchical structure, add new texts accordingly, and count words for all or defined parts of the corpus:

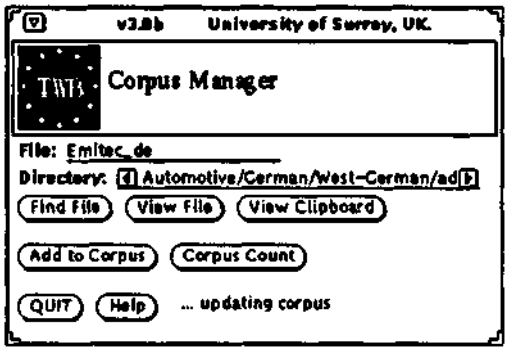

Figure 1: Corpus Manager Main Window.

The Corpus Manager is part of the MATE system (Machine-Assisted Terminology Elicitation)[2], to which we return below.

The paper-based resources to which translators traditionally refer in their search for terminological information are now becoming increasingly available in electronic form and accessible through a personal computer using CD-ROM technology. Collections of such texts - encyclopaedia, journals, newspapers, and so on - may be regarded as a kind of *ad hoc* corpus. But this material, a wealth of potential evidence for the solution of terminological problems, needs to be processed efficiently and purposefully. In other words, translators need software tools to perform various processing operations to extract terminological information from these texts, and to record this information for future use.

COMPUTERISED TERMINOGRAPHY

In this section we describe some of the functionality of the MATE system (Machine-Assisted Terminology Elicitation) developed at the University of Surrey (Holmes-Higgin and Griffin (10) and Holmes-Higgin and Ahmad (11)). MATE is an integrated toolset which covers all stages of terminography - or special-purpose lexicography - from term identification to publication in customised formats. The current toolset comprises: Customiser (for setting defaults including some which determine the automatic generation of administrative and codified data for selected term bank record fields), Corpus Manager (for managing the input of texts into the corpus), KonText (for processing text). Term Browser (for browsing the database), Term Refiner (for editing term bank data), and Term Publisher (for high-quality publishing). Two additional demonstration tools are the Natural Language Query and the Intelligent Query systems. Each tool can be called individually or through the MATE toolbox:

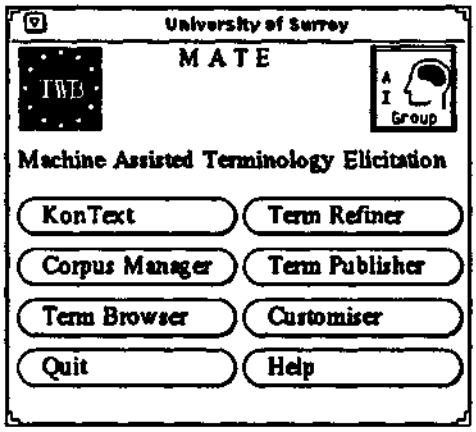

Figure 2: MATE toolbox

MATE, which is written in QUINTUS-PROLOG, was developed under UNIX on SUN-SPARCstations. The user interface was written using ProWindows, and the term bank data is stored in ORACLE, a proprietary relational database management system (RDBMS). A version of MATE is currently being ported to a PC environment under DOS using Windows-3 and C++, and the term bank data is stored in a COMFO-BASE, an RDBMS

designed for PC systems and marketed in Europe by Siemens. MATE can, in principle, be interfaced with any proprietary database.

In the rest of this section, we focus on three of the MATE tools: KonText, Term Refiner and Term Browser.

Term Capture and Elaboration

For a terminologist compiling a systematic terminology of a particular domain, the first phase of data capture involves the identification of potential terms from texts in the domain-specific corpus. The value of textual material is also acknowledged within the normative concept-oriented approach of Wüster's Vienna School: the processing of selected 'evaluated' texts as valuable raw material is recommended as a source of terms and other terminological data (such as definitions) (Felber (8); Picht and Draskau (9)). For the translator, the task of data capture will vary according to the nature of the comprehension and production problems encountered in the source language or target language texts respectively. Source language problems may include establishing or clarifying the meaning of a term, checking the status or authenticity of a term, and so on. Target language problems include establishing the collocational behaviour of a term, checking its default grammatical features, or checking its equivalence (e.g. through the use of parallel texts).

The tool for data capture in MATE is KonText:

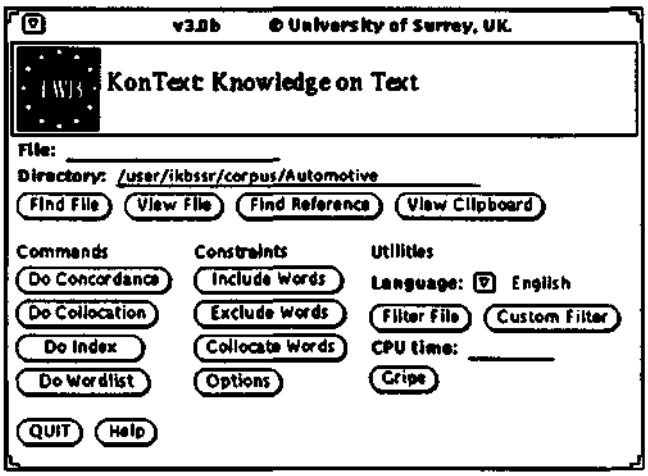

Figure 3: KonText Main Menu

KonText allows four basic operations to 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 line in source text); collocation (a list of the co-occurrence 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).

The relevant place in the original text can be recalled in a supplementary window by clicking on the line reference. Constraints may be set on these searches. For instance, 'Exclude Words' can be used to compile lexica to exclude closed class grammatical words, which are never terms. 'Include Words' can be used to restrict the search to a limited number of specified words.

The Options button allows various parameters for processing to be set. These include: the size of the context window for concordances; upper/lower frequency limits; sensitivity to hyphens in compounds; whether hyphenated words at the end of lines should be concatenated; whether punctuation should be processed or ignored; whether numbers in text should be processed or ignored; case sensitivity; sorting preference; printer name.

KonText has a limited ability to process texts that have been marked up with SGML. SGML markers can be used within the 'Include Words' and 'Exclude Words' constraints in the same way as words. KonText will include or exclude text for a given SGML marker until its end marker is found. This facility makes it possible to distinguish parts of texts which may exhibit characteristics untypical of running text, such as figure and table legends, headings, examples, bibliographic references, and so on.

The issue of 'what is a term of the domain?' is of less immediate relevance to the translator than to the terminologist. The translator's queries are motivated by the set of 'terms' occurring in the source language text. The question is not whether this 'term' is a legitimate or preferred term of the domain, but whether its meaning can be determined and a translation equivalent identified. The translator will therefore approach the corpus with a specific problem related to a particular term or set of terms. The terminologist, on the other hand, must first determine what is a term of the domain for the intended user group of the terminology which is being compiled. The elaboration of terms is of interest to both translator and terminologist. By elaboration we understand the provision of further terminological data such as definitions, contextual examples, grammatical information, collocations, sense relations, and foreign language equivalents. The whole cycle of terminographical work, i.e. the development of a term bank, has been modelled according to principles of software engineering. Four consecutive phases have been identified:

acquisition:	conceptual organisation of the domain
	creation of a corpus
	identification of terms
representation:	linguistic description of the term
explication:	definition
	contextual example
deployment:	sense relations
	cross-linguistic equivalence

The successful execution of each phase is delineated with clearly identifiable data (see Ahmad et al (12)).

Semi-automatic identification and elaboration of terms

One of the important differences between a specialist text and a general text is that of the distribution of linguistic tokens, that is words and word combinations, in the two kinds of text. One can also distinguish between the different genres of text by examining the distribution of linguistic tokens. We believe that this unevenness in distribution can be exploited for identifying terms. Consider some distribution statistics. Word distribution statistics for general language corpora indicate that the first 50 most frequent words are closed class words. However, an examination of specialist text corpora has shown that for a variety of domains, including automotive engineering, artificial intelligence, mammography and urban drainage, there are generally between five to ten open class words, usually terms, among the 50 most frequent words in these corpora (see Ahmad and Davies (13)). We have developed a simple comparison criterion to compare the relative frequency of occurrence of a word in a representative general corpus, for instance the LOB corpus or its equivalent contemporary corpus, with that of a specialist text corpus. The initial results are encouraging (see Ahmad et al (14)). The MATE system is currently being extended to incorporate this comparison facility such that the system will be able to produce a list of 'potential terms'.

The provision of further support to the MATE user in the semi-automatic identification of terms and of various sense relations including synonyms, hyponyms, and meronyms is progressing. Since neither the notion of word class (in an untagged corpus) nor that of sense relation is directly computable, our strategy has been to focus on the lexical environment of terms and their relations in running text, by investigating the linguistic patterns which form typical contexts for their occurrence. Terms, which are typically nouns, mostly occur in the environment of closed class words and punctuation marks. The identification of sense relations is being explored with a similar method, using a catalogue of 'probes'. For instance, *a type of, a species of, classed as, breed of, manner of,* and so on, were considered as equivalent to *a kind of,* the typical diagnostic frame for hyponyms. (see Ahmad and Fulford (15)). The results of this study have enabled us to specify another extension to the MATE system whereby the system will 'find' potential sense relations between terms. Subsequently, MATE will present its findings to the terminologist for use as possible elaboration data for a given term. An initial implementation of this extension has proved to be useful.

Term Storage and Retrieval

The use of databases to store and retrieve terminological data (term banks) has been the principal application of computers in terminography to date. In most implemented term banks, the data associated with each term is normally stored in separate fields of the term record with no connection between fields or between records. Given that linguistic data is highly interrelated (e.g. terms related to other terms - synonyms; homonyms; meronyms; variants; foreign language equivalents; terms embedded in text - definitions; contextual examples), the danger of entering inconsistent or contradictory data is considerable. The use of a relational database management system, as in the Translator's Workbench, has some advantages in that it allows links to be made between data items. For instance, in the Surrey term bank

definitions are shared between synonymous terms, so that any new synonym which is entered will automatically acquire the same definition/s. And links can be automatically created between the synonyms of terms which are entered as foreign language equivalents (see Holmes-Higgin and Ahmad (11)). More recently, interest has shifted to the possibilities offered by artificial intelligence for the representation of conceptual information. Such schemata include: semantic networks, frames and predicate logic, and their various derivatives and extensions, as well as schemata which are hybrids of two or more other schemata. Conceptual Graphs, a semantic network-oriented schema originally proposed by Sowa (16), have been used to represent terms on a computer system in order to help the user of a term bank visualise and explore relations between terms (see Ahmad and Hook (17)). This type of representation is much richer than the links in a relational database management system, since the system itself is able to infer relations between new terms and stored terms, given certain types of relation.

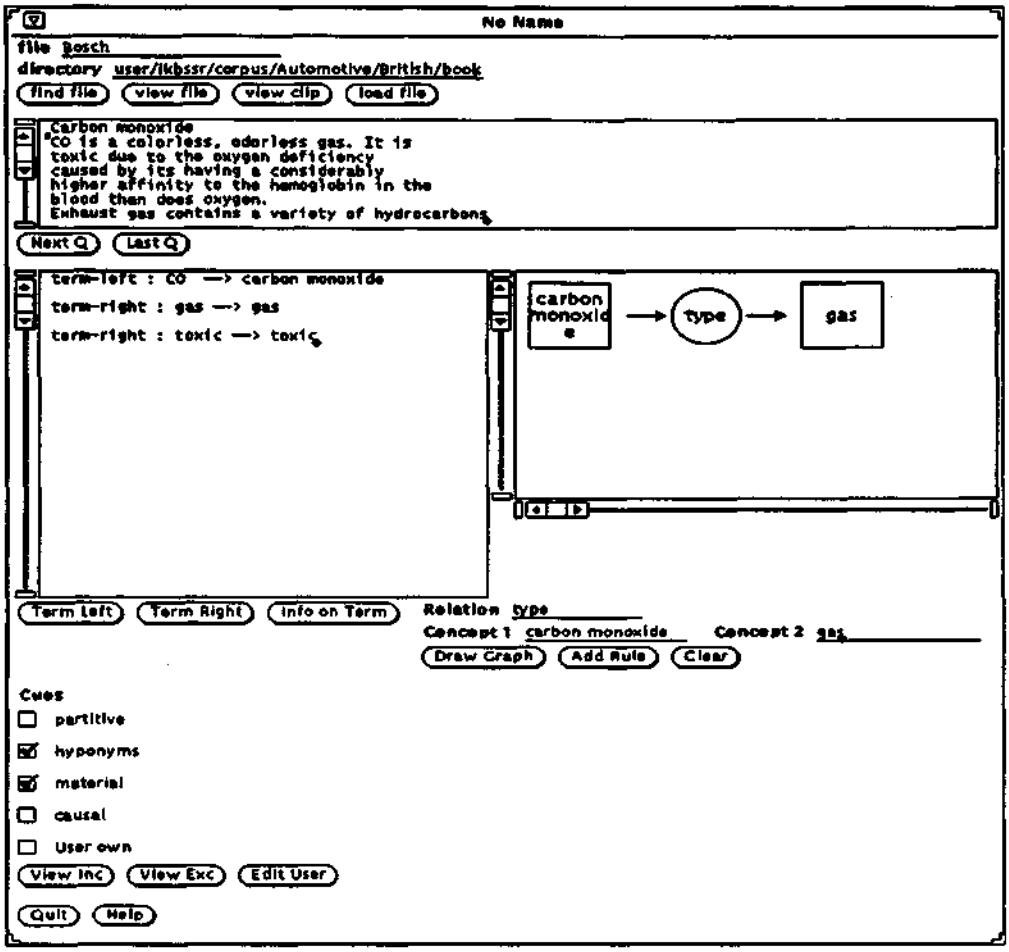

Figure 4: Output from a program for creating Conceptual Graphs for animating semantic relations in text. (The Conceptual graphs are in the 'window' on the right hand side: the 'concepts' are in square boxes, e.g. *carbon monoxide* and *gas* and the 'relation', like the hyponomy relation *type,* are inscribed inside the circle).

The MATE tools Term Refiner and Term Browser are used to input and modify data and to view the contents of the term bank in different subsets and configurations. Figure 5 shows the main window of Term Refiner, illustrating the kind of information which can be selected by the user. In fact, the subset of record fields to be displayed can be selected by the user, as well as the navigation paths between records.

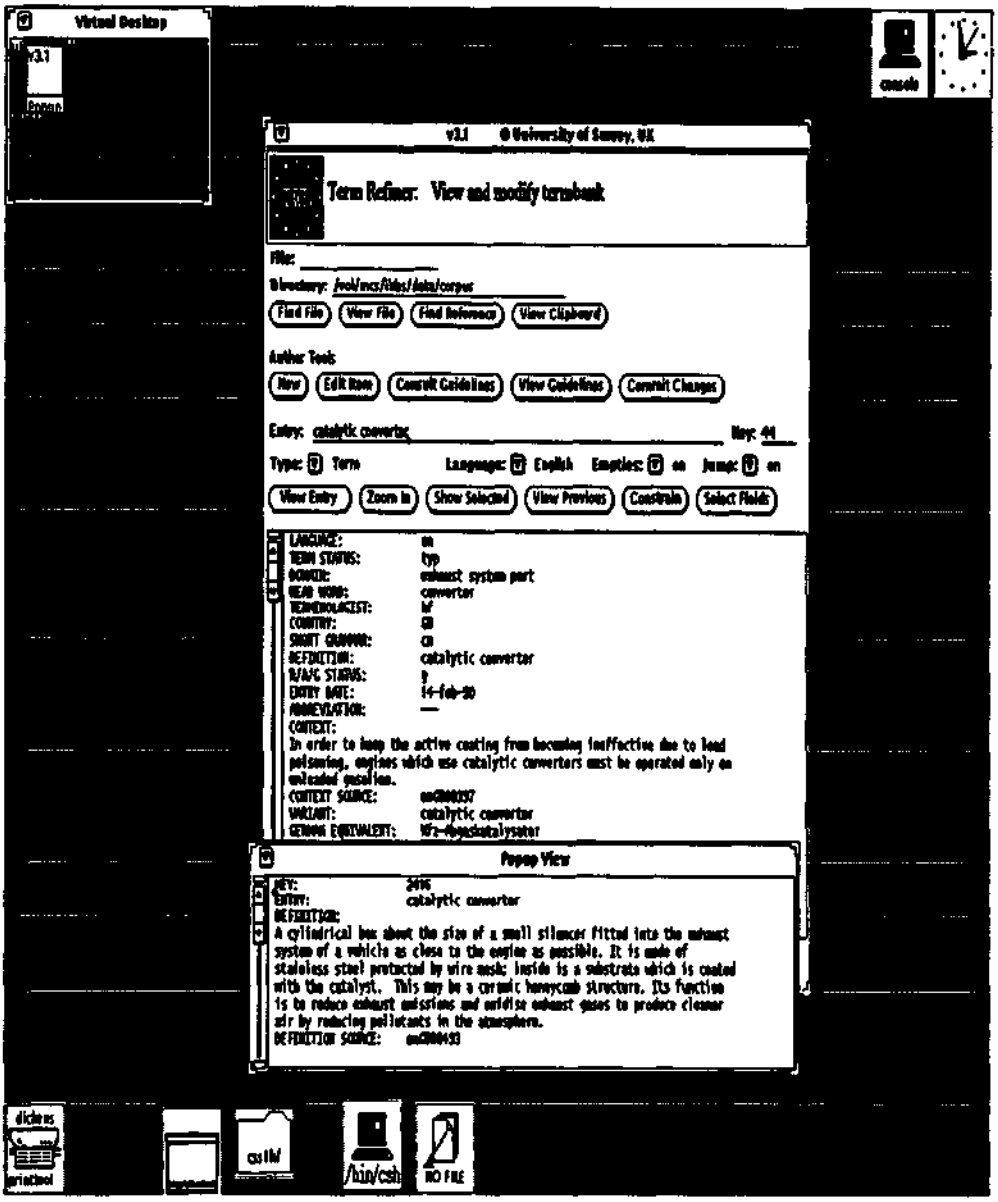

Figure 5: Term Refiner Main Window

The interface in Figure 5 (which shows the data stored for the builder of the term bank) must be distinguished from the retrieval interface for the term bank. The term bank interface has been implemented by another TWB partner, and is reported elsewhere (van Hoof and Mayer (18)). The trilingual term bank is multidirectional in

terms of source and target language, and allows five points of entry to the term bank structure for the user of the term bank: entry term; abbreviation; collocation; synonym; variant.

CONCLUSION

Corpora are a rich source of evidence for both general language and special language research. Their comparison is additionally informative, especially for the investigation of special language, where general language norms can be treated as a base-line. But for this evidence to be fully, consistently, and purposefully exploited, appropriate software tools are necessary. Text-processing tools currently available tend to be simple concordance tools with unfriendly human-computer interfaces. The MATE toolset, which we have outlined briefly in this paper, provides an integrated set of processing and storage tools for the whole cycle of terminology management. We are at present seeking to extend its functionality to provide a higher quality of support to the user. If corpora are to be used in a commercially viable way in the various applications to which they have direct relevance, including both general-purpose and special-purpose lexicography, then tools of the kind offered in the MATE toolset offer a promising way forward.

REFERENCES

1. Hainitz, R & Pownall, T., 1989, 'The ITI/TINOS Terminology Survey', Professional Translator and Interpreter. No.3.

2. Hainitz, R & Pownall, T., 1990, 'The ITI/TINOS Terminology Survey', Professional Translator and Interpreter. No.3.

3. Fluck, H-R., 1985, Fachsprachen, Einführung und Bibliographie. München: Francke. 3rd edition.

4. Leech, G., 1991, The state of the art in corpus linguistics'. In: K. Aijmer & B. Altenberg (eds.) English Corpus Linguistics. London & New York: Longman, 8-29.

5. Harris, Z., 1991, A Theory of Language and Information: A Mathematical Approach. Oxford: Clarendon Press.

6. Sinclair, J., 1991, Corpus. Concordance. Collocation. Oxford: OUP.

7. Wüster, E., 1985, Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie. 2. unveränderte Auflage; The LSP Centre, UNESCO ALSED LSP Network. The Copenhagen School of Economics.

8. Felber, H., 1984, Terminology Manual. Paris: UNESCO and Vienna: INFOTERM.

9. Picht, H. & Draskau, J., 1985, Terminology: An Introduction. Guildford: University of Surrey.

10. Holmes-Higgin, P. & Griffin, S., 1991, Machine Assisted Terminology Elicitation User Guide. University of Surrey.

11. Holmes-Higgin P and Ahmad, K. 1992, Knowledge Processing 8: The Machine Assisted Terminology Elicitation Environment: Text and Data Processing and Management in Prolog. University of Surrey. CS Report No. CS-92-11.

12. Ahmad, K., Davies, A., Fulford, H., Holmes-Higgin, P. & Rogers, M., 1992, 'Creating Terminology Resources'. In: M. Kugler, K. Ahmad, G. Heyer, M. Rogers, G. Thurmair (eds.) Multilingual Documentation and Communication. Final report of the ESPRIT Project 2315 The Translator's Workbench'. March 1992, 61-75.

13. Ahmad, K., and Davies, A., 1992, Knowledge Processing: 7. Methods of Corpus Linguistics & Terminology Extraction. University of Surrey. CS Report No. CS-92-10.

14. Ahmad, K., Davies, A., Fulford, H. & Rogers, M., 1992, 'What is a Term? The semi-automatic extraction of terms from text'. Paper presented at a conference held at the University of Vienna, Institut für Übersetzer- und Dolmetscherausbildung, Translation Studies - An Interdiscipline, 8th-12th September 1992.

15. Ahmad, K. & Fulford, H., 1992, Knowledge Processing 4: Semantic relations and their Use in Elaborating Terminology. University of Surrey. CS Report No. CS-92-07.

16. Sowa, J., 1984, Conceptual Structures: Information Processing in Mind and Machine. Reading, MASS: Addison-Wesley Publishing Co.

17. Ahmad, K., and Hook, S., 1992 Knowledge Processing 10: Conceptual Graphs and Term Elaboration Explicating (Terminological) Knowledge. University of Surrey. CS Report No. CS-92-13.

18. van Hoof, A. & Mayer, R., 1992, 'Special Language Resources: Termbanks, Cardbox'. In: M. Kugler, K. Ahmad, G. Heyer, M. Rogers, G. Thurmair (eds.) Multilingual Documentation and Communication. Final report of the ESPRIT Project 2315 The Translator's Workbench'. March 1992, 49-59.

FOOTNOTES

1 This work was carried out as a part of the Translator's Workbench projects (ESPRIT 2315; ESPRIT 6005) funded by the Commission of the European Communities.

2 The chief systems designer of MATE is Paul-Holmes Higgin.