KITES
(Knowledge Based Integrated Terminology System)

Patricia Thomas, Anne Wilson and Anne Judge

Brief Outline of the KITES Project

The principal aim of KITES, (*K*nowledge-Based *I*ntegrated *TE*rminology *S*ystem) is to incorporate a term bank within the framework of an integrated electronic office system, particularly in a European multilingual environment, to help translators and subject specialists. The KITES project is an attempt to develop a term bank based on principles of Terminology and Linguistics, including those of Lexicology, and augmented by the use of Artificial Intelligence. This work involves investigating the application of Artificial Intelligence (AI) techniques and methods, especially the use of knowledge-based techniques, to the storage of terms and their constituents.

This paper is a report on the progress achieved since 1985 so that others wishing to set up term banks and knowledge bases adapted to their particular needs may benefit from our experience. The emphasis of the paper will be on terminological rather than computational aspects.

Term Banks

There are various types of term bank used to store terminological data; some are word-based, such as EURODICAUTOM, while others may be concept-oriented (e.g. DANTERM). The Surrey term bank is an organised collection of specialised terms and includes the administrative, linguistic and conceptual data required to identify each term, as well as extra data such as examples of the term in context, classification and bibliographic references. The term and its relevant data are presented in a record format structured in fields, with storage and retrieval being the key factors.

The Surrey record format provides some 30 separate fields divided into the three sections mentioned (Fig. 1). These fields can be linked to the MATER structure. The administrative fields include such information as the origin of term, its key number, entry date and update, the domain and the identity of the terminologist responsible for it etc. The linguistic fields include the *term*, which describes a particular *concept*, together with any grammatical and phonetic information which give additional information about the term. Fields which help to elucidate the concept include synonyms (Field 16), antonyms (Field 24) and the *definition* (Field 20), in which may be found conceptual links with other terms, such as ontological and logical relations, as they are known in terminology. It is to the *analysis* of the concepts in the *definition* that our efforts are being directed; in other words, the record is concept-oriented.

The record format established initially aimed at a data structure which would provide comprehensive information suitable for all domains. The terminologists working on the five subject domains which will be listed in the third part of this paper have completed those fields of the record format which they deemed most relevant for the needs of translators and subject specialists, and have analysed the extent to which these fields fulfilled their requirements. To give an example, the terminologist working in the area of Information Technology considered that the inclusion of 'in-house' jargon and archaic terms would not be relevant but was surprised to find subsequently how many terms were used 'in house' or had already been superseded.

While storage in such a record format is ideal for *atomic* data such as instances of one word equalling one concept in different languages at a given time, it presents major problems with regard to the definition. A definition, although describing one concept, will inevitably mention others. Access to these other concepts, however, is not possible in a relational data base management system (RDBMS), as text fields are inadequately catered for by its record system which can only retrieve large portions of text-(hyphen) no search and retrieval can be carried out *within* those texts. Computational techniques are therefore being devised to break down the definition into its conceptual characteristics so that information may be retrieved selectively. This means that the structure of the various types of definition needs to be investigated further and AI techniques deployed to convert the 'text base' into a knowledge base.

Figure 1 The SURREY record format

The SURREY record format is shown below. It has 29 fields, and can be used to represent terms in three languages. The DBMS, *ORACLE*, allows us to extend the number of key fields to 200. The important distinctions between the SURREY record format and others are the use of conceptual links (9) and the inclusion of phonetics.

Location	Field Name	Display Name
0	KEY	No: of term
1	ORIGIN	Place of origin of term
2	POOL	Pool (sub-division of domain)
3	ORIGINTR	Terminologist's initials
4	DATE	Entry date
5	UPDATE	Update
6	UPDTR	Updater
7	SUPDT	Source of update
8	UPDTXT	Text of update
9	SUBJECT	Domain or subject field
10	LCCODE	2-letter language and country code (as per ISO)
11	ENTRY	*Entry term*
12	SORIGIN	Source (3 letters, e.g. ISO, BSI)
13	STYPE	Source type (3 letters, e.g. DIC=dictionary; JOU=journal)
14	SNUMBER	Source number (listed in separate computer file)
15	SPAGE	Source page
16	SYNONYMS	Synonyms (Latin equivalent, jargon)
17	ABFORM	Abbreviations (initialisms, acronyms, cuttings)

18	CONTSYN	Contextual synonyms
19	DEPTERM	Deprecated terms (archaic, slang)
20	DEFINITION	Definition (includes characteristics of term)
21	BIBREF	Bibliographic references
22	CONTUSAGE	Context/usage (complements definition)
23	CONCEPOS	Conceptual links (position of terms in system of concepts)
24	ANTONYMS	Antonyms (where applicable)
25	RELTERM	Related term(s)
26	GRAM	Grammatical note
27	CLASSIF	Classification
28	COPYRIGHT	Copyright
29	PHON	Phonetic forms
30	LCCODE2	Second language and country code (2 letters each)
31	EQUIV	Equivalent term
32	SYNONYM2	Synonyms
33	DEPTERM2	Deprecated terms
34	SCOPE	Scope note (on degree of equivalence between first and second lang)
35	BIBREF2	Bibliographic reference
40-45		Third language as for second
50-55		Subsequent languages may be added

Subject Domains

The Surrey term bank at present contains three types of specialised subject domain: highly taxonomic subjects, such as virology and sound insulation; subjects which start their life as highly taxonomic, yet become less so through popular usage, e.g. automotive engineering (AE) and information technology (IT) and subjects which are less taxonomic and more difficult to equate from language to language, such as law and administration. The three types of domain were initially represented bilingually, with British English featuring as source language in each language pair. The second and subsequent languages are European (French, German, Spanish, Norwegian and Italian). With each foreign language equivalent, scope, synonyms and bibliographic references are provided, along with grammatical and phonetic notes where necessary.

Virology and sound insulation are two rapidly growing domains which have evolved comparatively recently and which have wide international application. The terminology is determined by the consensus of experts at international meetings. The stress is on technical use, with terms being generally consistent and English the most widely used language of communication. Administrative terminology, on the other hand, stems from the requirements of national institutions and their officers, and has developed over a long period of time in a monolingual environment.

Because of their non-taxonomic nature, these terms and their foreign language equivalents are difficult to standardise; in these instances information such as definition and scope notes plays a crucial role. Legal terminology, where it is basically mercantile, is less difficult. Information technology and automotive engineering terms need special terminology treatment because they primarily evolve in enclosed technical and commercial environments where extensive abbreviation is often used. Due to the pressures of commercial competition, terms may evolve rapidly through general usage, sometimes with surprising results; the term 'to boot' the computer, for example.

A problem which has long been of concern to lexicographers is where to draw the line between domains, and between Language for Specific Purposes (LSP) and Language for General Purposes (LGP). A term may still be considered technical even though it has entered popular usage. The problem with these overlap areas is how they should be represented in a term bank. The 'part/whole' relationship can be stored; differences in use may come under 'scope', but homonyms in different subject areas, where properties of inheritance may not apply, could result in the repetition of storage (Ross 1988). This is one area where considerable research is necessary.

Term Banks and Computer-Aided Lexicography

There are several areas in which both a term bank and a knowledge base can be used as 'lexicomputerate' tools. Specialised dictionaries can be produced which may be available on paper or on computer. There are three important advantages to these dictionaries. The first is that in rapidly evolving fields, they may be regularly and frequently updated: for instance, the Macmillan's Directory and Dictionary of Animal, Bacterial and Plant Viruses (1989), although a traditional dictionary from the point of view of layout, has been compiled on computers to allow for easy transmission of data between authors and editor, editor and printers, and for ease of access and updating. The authors of the dictionary have been advised on principles of terminology when compiling their definitions so that a formula has been adopted for each definition which gives hierarchical relations, intrinsic and extrinsic characteristics, and so on. Secondly, the development of CD-ROM techniques allows a vast quantity of data to be stored in a very small space. Thirdly, dictionaries no longer need to be ordered alphabetically but may be structured conceptually in the manner of a thesaurus; this is particularly useful where the name of a concept does not readily come to mind. There are some good examples of the latter already in print, such as the CAMBRIDGE ILLUSTRATED THESAURUS OF COMPUTING SCIENCE, published in 1984.

Another instance of the effectiveness of term banks is their use in storing up-to-date terms which have been extracted by compiling word frequency lists from recently published texts. Using statistical techniques, the subject domains of technical texts can be identified from these frequency lists (Lyne 1984; Yang 1986; Thomas 1988). Concordances are run on text in order to extract terms from it. The examples thus extracted provide contextual illustrations of the terms which are invaluable for both monolingual and multilingual users. The structure of the COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY produced by the University of Birmingham and Collins Publishers bears witness to the effectiveness of this revolutionary style of dictionary compilation (Sinclair, 1987). On a smaller scale, the Word Lists of French neologisms taken from current popular journals, with comments, and published at intervals by the University of Aberystwyth in Wales, is a useful tool for translators of general language texts (Bremner 1986-).

The Role of Terminology in Knowledge Representation

Concordances can also be helpful in the construction of knowledge bases. A knowledge base is that part of an expert system in which the knowledge pertaining to the subject is stored, i.e. for our purposes, that which is found in the *definition* of the term. One of the most common means of storage is by *frames*. Frames are organised hierarchically, with links between frames being explicitly stated. Fig. 2 shows a mammal frame in which may be found, among others, a feline frame; under this might be a cat frame, then a frame for Siamese cats and under this a frame for my cat. In other words, a frame-based system attempts to store its knowledge in a fashion that reflects the conceptual structure of the domain: objects (in this case, cats) are hierarchically structured, with inheritance playing a key role (i.e. my cat will automatically inherit all the properties of Siamese cats). Inferences may also be drawn. Associative terms (i.e. RELTERM; Field 25, e.g. lions and tigers) are indicated, and generic relations and equivalence are covered. There are clear links here with terminology.

Figure 2

Inheritance:

mammal
|
feline
|
cat
|
Siamese cat
|
my cat

Related terms: lion, tiger, leopard

Conclusion

Term banks need not be multilingual, since they are involved primarily with the storage and dissemination of knowledge. However, given the importance of accurate translation in communication, they are an ideal medium for the incorporation of equivalents in different languages. Verification of terms (in source and target languages) and their definitions frequently requires reference to a number of other sources, particularly to experts in a domain. Since term banks are concerned with the storage and dissemination of knowledge (as well as with the standardisation of terms), they offer an ideal opportunity for providing one system for the translator, as well as other potential users, containing source language and target language terms, together with the relevant "knowledge". While a term bank, as mentioned, stores its facts in a purely atomic way, there is no link between representation and deployment. By linking term banks to expert systems, knowledge can be represented so that it can be deployed to solve problems in a given domain. The use of concordances may provide primary data for the construction of a so-called 'intelligent' term bank: data are no longer treated as atomic but are seen in context, and context is the first step in providing links between domain objects, whether they be viruses, cats, or parts of a car. These links, if fully implemented, will give a multi-faceted view of the concept and thus obviate the need for definitions or, to put it another way, they will *form* the definitions: any term is described via its relationships with (or differences from) other terms: the full representation of these relationships defines the term. The KITES Project aims to provide such a multi-functional, cost-effective system to fulfil the requirements of lexicographers, terminologists, mono- and multilingual users, and yet provide data in a workable format for inclusion in an expert system.

References

Cited Dictionaries

CAMBRIDGE ILLUSTRATED THESAURUS OF COMPUTING SCIENCE 1984. Cambridge: Cambridge University Press.

COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY (CCELD) 1987. J. Sinclair et al. (eds.) London and Glasgow: Collins.

Other Literature

Bremner, G. 1986 — present. Aberystwyth Word Lists. University of Aberystwyth.

Calzolari, N. and E. Picchi. 1988. 'Acquisition of semantic information from an on-line dictionary' in *Proceedings of the 12th International Conference on Computational Linguistics.* Budapest: John von Neumann Society for Computing Sciences.

Lyne, A.A. 1984. 'On rooting out words with inflated frequencies in word-counts of specialized registers' in *LEXeter '83 Proceedings.* Tübingen: Max Niemeyer Verlag.

Ross, C. 1988. *British Term Bank Project — The Heriot-Watt Experience.* UNESCO ALSED-LSP Newsletter. Copenhagen School of Economics.

Thomas, P. 1988. *Analysis of an English and French LSP: Some comparisons with English general text corpora.* UNESCO ALSED-LSP Newsletter. Copenhagen School of Economics.

Yang, H. 1986. 'A new technique for identifying scientific/technical terms and describing science texts' in *Literary and Linguistic Computing* 1: 93—103.