KNOWLEDGE MANAGEMENT FOR TERMINOLOGY-INTENSIVE APPLICATIONS: NEEDS AND TOOLS

Ingrid Meyer
Artificial Intelligence Laboratory, Dept. of Computer Science
University of Ottawa, Ottawa, Canada K1N 6N5
IXMAL@acadvm1.uottawa.CA

ABSTRACT

This paper addresses the problem of how to provide support for the acquisition, formalization, refinement, retrieval - in other words, for the *management* - of the knowledge required for producing high-quality terminology. This problem will become increasingly significant as term banks evolve into knowledge bases. Knowledge management for terminology-intensive activities is complicated by two factors: 1) the importance of encyclopedic as well as lexical-semantic knowledge, and 2) the wide spectrum of working environments in which terminological activities can be carried out, from terminology as a *distinct specialization* at one end of the spectrum, to terminology as practised in *document-production* at the other. In the first two sections of the paper, we briefly analyze each of the two complicating factors. In the third section, we describe the terminological support that is currently available and under development in a knowledge management tool called *CODE*, which is being used to build a prototype, knowledge-based term bank called *COGNITERM*, designed to be useful across a spectrum of terminology-intensive environments.

1. LEXICAL-SEMANTIC AND ENCYCLOPEDIC KNOWLEDGE IN TERMINOLOGY

Terminology is the practical discipline concerned with *describing* and *naming* concepts in specialized domains. The data produced by this process is, increasingly, stored in data bases known as *term banks*. Since concepts are the starting point for all practical terminology work, and since concepts are the building-blocks of knowledge, it follows that terminology is a very knowledge-intensive activity: *describing* concepts involves acquiring knowledge about their characteristics, and *naming* concepts involves matching conceptual characteristics with linguistic forms (i.e. terms). In fact, *terminology* is somewhat of a misnomer: most fundamentally, it is not the study of "terms", but rather of the knowledge conveyed by the terms.

Given the crucial role of knowledge in terminology, one needs to address the question of what kind of knowledge to include in term banks. In the related discipline of lexicography, the same question, formulated in relation to dictionaries and lexicons, has resulted in the long-standing debate about differences between lexical-semantic (i.e. linguistic) and encyclopedic (i.e. world, extra-linguistic) knowledge (1). One viewpoint has been that dictionaries and encyclopedias should be conceived as distinct entities - hence the apothegm "dictionaries are about words, encyclopedias are about things". Meaning-Text (MT) Lexicography (Mel'cuk 1988a/b), for example, makes a strict distinction between lexical-semantic and encyclopedic knowledge on the basis of semantic features: those which are necessary and sufficient (in the mathematical sense) to a definition are lexical-semantic, while those that are superfluous are encylopedic, and banned from

definitions. A contrasting point of view, expressed for example by McArthur (1986:pp. 102-109) is that for certain purposes, it may be useful to produce a hybrid of dictionary and encyclopedia - in other words, an encyclopedic dictionary. McArthur proposes that the dictionary-encyclopedia relationship be seen as a continuum rather than a dichotomy, and proposes the term *micro-lexicography* to designate the activity dealing with "the world of words...", and the term *macro-lexicography* to designate the activity which "shades out into the world of things and subjects, and centres on compendia of knowledge...".

Depending on the lexicographic framework and the intended user of the dictionary, a strict differentiation between lexical-semantic and encyclopedic knowledge can be not only theoretically interesting, but also practically relevant: in the MT framework, for example, it underlies virtually all aspects of lexicographic methodology. Learners' dictionaries, on the other hand, whether they are aimed at learners of the mother tongue (i.e. children's dictionaries) or learners of a foreign language, typically feature encyclopedic characteristics, such as pictures and a rich supply of examples, that are intended to supplement definitions. Furthermore, the definitions themselves may include information that exceeds the bounds of "necessary and sufficient".

The lexical-semantic *vs.* encyclopedic debate is pertinent to terminology since this discipline is closely linked to lexicography in purpose and method. In our view, terminology is clearly macro-lexicographic (to use McArthur's term) in orientation: term banks must include not only necessary and sufficient information about concepts, but also a certain amount of encyclopedic information as well. The following are just some of the reasons for our view:

Relationship to specialized domains. Terminology is closely related to the specialized domains of activity whose lexica it describes. This is reflected in the basic organization of term banks according to specialized domains (i.e. subject fields). Until recently, most terminology work was done by domain experts, and the increasing numbers of terminologists who are not domain experts still consider consultation with experts to be crucial to their work. One of the goals of terminology is to provide assistance in the ordering and use of terms within specialized domains. Because of its relationship to languages for special purposes (LSPs), terminology has a need for subject classification and thesaural structure. In other words, it is closely linked to information science, with which it shares tools such as keywords, indexes and thesauri.

The role of term banks as learning tools. Although term banks can be consulted by users with a wide range of domain expertise, by far the bulk of users are not domain experts. The largest user group has always been translators, who consult term banks not only for strictly linguistic information (e.g. part-of-speech, morphology, target-language equivalent), but also for conceptual information (e.g. conceptual characteristics, relations to other concepts), since it is well known that a certain depth of understanding of the domain is necessary to use its terminology correctly. Term banks can be seen as learning tools for the terminologists themselves, for example, when they are assigned a new field in which they have little knowledge, or when they are working in a field that is highly influenced by neighbouring fields with which they are not very familiar. Because of the teaching role of term banks, definitions are often complemented by examples of terms in context, much in the same way that learners' dictionaries are. Like encyclopedias, terminological publications often include pictures and diagrams.

The multilingual aspect of terminology. All the large term banks that currently exist are multilingual, and this tendency will most likely remain in the face of the increasing importance of international communication for trade and knowledge communication. It is well known that establishing lexical equivalence between different languages is often impossible on the basis of lexical-semantic information alone. To take a well-known example from general language, the word *river* is defined in *Webster's Ninth New Collegiate Dictionary* as "a natural stream of water of considerable volume". French, however, distinguishes between flowing bodies of water that empty into the ocean (*fleuve*)

and those that empty into a lake or another flowing body of water (*rivière)* - information that would be considered encyclopedic if one applied the necessary-and-sufficient rule.

The need for multifunctional term banks. In keeping with the increasing emphasis on the shareability of lexical resources in general, term banks will have to aim at meeting the needs of more and more user types, including machines (Freibott and Heid 1990, McNaught 1990, Meyer 1991). Machine uses (e.g. machine translation, expert systems, NL interfaces to databases) will require very large quantities of explicitly represented conceptual information, since they do not possess much of the world knowledge that humans know implicitly.

Because of the important encyclopedic dimension of terminology, we feel that a term bank can be conceived as a kind of knowledge base, and we are currently in the process of designing a prototype *knowledge-based term bank*, called *COGNITERM*, in the Artificial Intelligence Laboratory of the University of Ottawa, Canada. COGNITERM will be constructed using a knowledge engineering tool called *CODE* (Conceptually Oriented Design Environment, Skuce *et al.* 1989), that has already been tested in two terminology-intensive environments, where a number of small knowledge bases (several hundred concepts) were constructed. Before discussing the research in progress for the COGNITERM project (Section 3), we will briefly describe some of the knowledge management needs that our research is aiming to fill.

2. KNOWLEDGE MANAGEMENT NEEDS ACROSS THE TERMINOLOGY SPECTRUM

As explained above, the knowledge management problem in terminology is heightened by the fact that persons doing terminology need to manage both lexical-semantic and encyclopedic knowledge. This problem is further complicated by the fact that terminology is a very heterogeneous discipline, since the naming and description of specialized concepts can be carried out in a wide spectrum of working environments, dictating various types of knowledge management support. At one end of this spectrum is what we might call the most "pure" form of terminology, namely terminology practised *as a distinct specialization.* In this type of environment, we find persons officially designated as *terminologists*, often with professional training and/or certification in terminology (2), following a controlled methodology (3). At the other end of the spectrum we find a much more "casual" form of terminology as it is practised *as a component of document-production.* Here, the naming and description of concepts is carried out at various "links" in a "chain" of activities, which can include product design specification, technical writing (e.g. user manuals), revision, proofreading, translation, management information, etc. Normally, many of the persons involved in these activities have no specialized training in terminology, their methodology can be highly informal, and there may be no centralized repository for the terminological data.

The technology and methodology we are developing for terminology-oriented knowledge management support are intended to be generic enough to be useful across the spectrum of terminology environments. The various knowledge management (KM) needs that characterize the two ends of the spectrum are examined in turn below.

2.1 Terminology as a Distinct Specialization

This type of environment is typified by organizations such as the Department of the Secretary of State of Canada, which has had an official terminology service since 1953,

employing up to 80 staff terminologists preparing up to 4,000 terminology records a week (4). The mission of these terminologists is to facilitate the proper use of terms, in English and French, throughout the Public Service of Canada. To this end, terminologists maintain what is now the largest term bank in the world (about one million database records). They also prepare bilingual glossaries (which are often published) on subject areas requested by clients, and respond to inquiries from clients on specific problems. The terminological data that is collected can be conceptual or linguistic (5): conceptual data includes information such as subject-field labels, synonyms and antonyms, definitions, and equivalents in the second language; linguistic data includes information such as part-of-speech, morphological anomalies, usage labels, and idiomatic expressions. Terminologists in environments such as this one most often work *thematically* (6): in other words, they collect and describe (as exhaustively as practical constraints allow) the specialized terms used in a given field.

The major challenge of terminology is conceptual, not linguistic: terminologists are trained in linguistics and thus are properly prepared for the linguistic dimension of their task; in contrast, they are not normally domain experts, yet they require a substantial amount of expert knowledge in order to do their work. In other words, the major difficulty is pinning down the *meanings* of terms. Compounding their problem is the fact that terminologists can be required to work in several fields simultaneously, or to change fields frequently depending on clients' needs.

In the following paragraphs, we summarize (7) the four components of a terminologist's work in terms of the KM tasks on the one hand, and the roles of this knowledge in the production of terminology records on the other.

2.1.1 Selection of documentation

KM tasks. Before any collection or analysis of terms can occur, terminologists must select the knowledge sources for the project. Given their linguistic orientation, they have traditionally preferred texts as knowledge sources, although the collaboration of experts is also highly valued. Before collecting the documentary corpus, terminologists acquire some general knowledge about the field by doing introductory reading in textbooks, encyclopedia articles, popularizing journals, etc. They begin to familiarize themselves with the general knowledge structures of the field, trying to determine its boundaries, subdivisions, and areas of overlap with other fields (for multidisciplinary fields). Often, at this stage, terminologists will sketch out these "skeletal" knowledge structures in the form of a concept network. They will also make mental or written notes on a number of individual concepts which emerge as being particularly important.

Roles of knowledge. These preliminary KM activities are crucial to the selection of the documentary corpus since they help to clarify the project's scope: a clear idea of the conceptual boundaries of the field helps delimit the range of documentation to be sought. Determining areas of overlap with other fields also helps terminologists establish links with related documentation. When the terminologists are ready to begin the search for the documentary corpus, a clear idea of the major subfields helps them orient their work along a number of documentary "paths", which may be prioritized according to users' needs. The names of subfields, of key concepts, and of the characteristics of these concepts help provide specific points of entry into the documentation. Having a general idea of the hierarchical structure of the field also helps orient the process of documentation selection since terminologists tend to proceed from general to more specific literature.

Once a preliminary corpus is obtained, their general knowledge of the domain provides terminologists with a yardstick for judging its quality. It also helps them classify documentation according to subfield. In multilingual terminology, classification according to subfield is particularly important: to "manage" the large amounts of documentation to be *scanned* (see 2.1.2 below), terminologists very often work on one subfield at a time, in one language and then in the other, before proceeding to another subfield.

Finally, these preliminary conceptual activities provide terminologists with the conceptual framework and basic terminology needed for communicating with librarians and other documentation resource persons, as well as with experts. Communication is particularly important in the case of experts (8), who tend to be very busy: if terminologists have done their "homework", they will be able to direct the conversation in order to elicit the maximum information in a minimum amount of time. "Starting out on the right foot" in this way boosts terminologists' credibility with experts, and increases their chances of convincing experts to remain involved as the project advances.

2.1.2 Establishment of a nomenclature

KM tasks. Once the documentation has been selected, it undergoes a process called *scanning*, i.e. careful reading, with the extraction (9) of potential terms along with their contexts (10). Additional research may be needed for specific problems (e.g. terms not found for concepts identified, terms with inadequate contexts), after which the data is organized by grouping the various instances of a term, noting obvious cases of synonymy, abbreviations, usage labels, etc. Through the scanning process, terminologists begin to analyze (11) the general knowledge structures of the field, fleshing out (whether on paper or in the mind) the skeletal concept network drafted during their background reading. They also begin to analyze the conceptual characteristics of individual terms (i.e. the terms found in the documentation), based on the contexts in which the terms appear.

Roles of knowledge. Drawing on their general understanding of the domain, terminologists begin identifying the lexical items that are specific to their field. This process involves eliminating terms that would constitute "noise" in the terminology, i.e. lexical items that belong to general rather than specialized vocabulary, or terms that do not fall within the established boundaries of the field. As well, terminologists must identify what are known as "silences," i.e. lacunae in the preliminary terminology. As terminologists prepare to finalize the nomenclature (i.e. to determine the terms for which records will be prepared) and decide which contexts will be retained for analysis (2.1.3), the conceptual framework acquired so far will help them continue to communicate about problem areas with documentation resource persons and domain experts.

2.1.3 Preparation of term records

KM tasks. Using the established terminology and associated contexts, terminologists can begin a systematic analysis of terms-in-context. The primary function of this analysis is to determine the meanings of the terms, although it also serves to identify other linguistic characteristics such as part-of-speech, gender, frequency, geographic origin, etc. The conceptual goal at this stage is to achieve the depth of understanding needed to complete the term records. Terminologists carefully analyze the various contexts in which the terms have been found in order to identify a certain number of conceptual characteristics for all concepts. These characteristics will then be compared with those of potentially related concepts (e.g. synonyms, equivalents in the other language) in order to determine those which are necessary for establishing a conceptual match.

Roles of knowledge. The most important application of conceptual analysis is definition construction. If they are attempting the classic intensional (i.e. genus-differentia) definition, terminologists will need to compare the characteristics of a given concept with those of concepts at the same hierarchical level (i.e. with the characteristics of the co-ordinate concepts (12)) in order to determine the distinguishing characteristics (i.e. the differentia in an intensional definition). Relations other than the generic-specific (e.g. whole-part, cause-effect, tool-function) may also be analyzed and reflected in definitions.

Conceptual analysis is also essential to identifying synonyms and equivalents in the second language. Identifying synonyms requires a careful comparison of conceptual characteristics in order to determine that these are indeed identical for the terms in question.

When two concepts differ in only a very few (and not very significant) characteristics, they may be designated as pseudosynonyms (e.g. one concept may have one more characteristic than another, and thus be more specific). Establishing a conceptual match is also crucial to multilingual terminology work, which is complicated by the fact that conceptual structures often do not correspond perfectly from one language to another, resulting in cases of incomplete equivalence. Sometimes there may be no equivalent in the other language at all, resulting in the need to create a neologism (13). In this case, conceptual analysis is essential for determining whether the concept already exists within the current knowledge structures of the target language, and when it does, what its characteristics are (since the concept is so new, its characteristics, and consequently its location within the knowledge structures, may still be fluctuating). In many cases, an existing term will be adopted to designate the new concept, and conceptual analysis of the candidate terms is essential for determining which one possesses the greatest semantic compatibility with the new concept.

2.1.4 Quality control

KM tasks. Quality control can be achieved by two types of activity: *revision* and *updating*. On the one hand, before the project is completed, the various types of information collected by the terminologist are revised by domain experts and other terminologists (e.g. terminologists with experience in neighbouring or related fields, or more experienced terminologists). On the other hand, after the project is completed, a periodic updating of terminology records can occur whenever this is justified by changes and expansion in the domain. Revising the results of a terminology project involves analyzing and discussing specific conceptual problems identified by the experts and/or other terminologists. Periodic updating implies a monitoring of changes in knowledge structures and conceptual characteristics.

Roles of knowledge. To facilitate revision, terminologists need a sound understanding of the domain in order to interpret feedback from experts, and to elicit information on this feedback (e.g. when terminologists do not understand feedback, when the feedback contradicts what the terminologists found, or when experts give conflicting feedback). Regarding updating, a clear understanding of the current state of the knowledge will give the terminologist a basis for comparison when new structures and conceptual characteristics emerge. Conceptual problems increase when a field is particularly large or has complex knowledge structures, or when the field is changing rapidly.

2.2 Terminology as a Component of Document-Production

By *document-production*, we mean a "chain" of writing activities that are carried out from the inception of a product (14) to the production of public (or widely available) written information about this product. The "links" in a document-production chain can be distributed throughout an organization, and the actual "documents" in various states of completion. These documents can include anything from product designers' rough personal notes, to intermediate "current state" documents used to coordinate members of a team, to "official" publications (e.g. technical manuals produced by technical writers), to translations of these official publications.

Although there are usually no officially designated *terminologists* in this type of environment, terminology-intensive activities are pervasive nonetheless: concepts are described and named by persons such as product designers, technical writers, proofreaders, revisers, abstracters, management information specialists, public relations officers, and translators (15). Given the heterogeneity of this type of environment, the terminology-related KM problems are much more complex than they are in the "purer" form of terminology work described above. The following are just some of the issues that contribute to this complexity.

2.2.1 Methodology

Given the variety of people involved in document-production, this kind of environment typically exhibits a lack of consistent methodology for terminology work. This problem is particularly crucial at the early stages of document-production. For example, product designers carry a heavy burden of defining and naming concepts, but have no formal training (and very often, no interest!) in terminology. Terms that are chosen "on the fly" easily become entrenched, even though they may be inappropriate. Normally, this type of environment does not stress a methodology for assuring that terms are clearly described and logically named, nor that the consistent use of approved terminology is enforced.

2.2.2 Coordination between links in the document-production chain

Given the number of people that can be involved in document-production, coordinating the various links in the chain is a fundamental problem. Writers in a given link in the chain may, for example, have trouble understanding what the originator of certain terms actually meant by them. If it is impossible to contact the originators of knowledge personally, the meanings of terms may have to be reconstructed from scant resources. Knowing that a given document will soon be passed on to another link in the chain, documentors are easily tempted *not* to resolve terminological problems that they have inherited, leading to a "pass-my-confusion-onto-the-next-person" phenomenon. Complicating things is that documents do not flow in a one-way direction from inception to finalization; documentors, consequently, can be sent in loops. Common terminological problems resulting from this situation are *inconsistency* (terms being used to mean different things by different people), and *overloading* (terms used in too many different senses).

Coordination is also complicated by the fact that concepts exist at different levels of "clarity" at the various links in the chain: at the initial design stage, they may still be quite fuzzy; by the time they are documented in some kind of "official" text, their conceptual characteristics should be (in principle, at least!) much clearer. From a terminological point of view, this conceptual fluidity means a continuous evolution of concept definitions and names from one link in the document-production chain to the next.

2.2.3 Centralization of terminological data

Most organizations do not maintain centralized repositories (e.g. term banks) of terminological data. When such repositories do exist, they often take the form of informal glossaries that may be out of date, not validated by experts and/or professional writers, and not used consistently throughout the organization. This state of affairs places a heavy onus on the documentor to find out who originated certain terms, what the terms mean, how they should be used in context, how they should be translated, and so on. The lack of centralization of terminological data (particularly conceptual data) is particularly problematic for people who are at the *end* of the document-production chain - for example, the editors, proofreaders, and translators (16): they are the furthest away from the originators of concepts (and have the hardest time accessing these originators); the documents passed on to them are likely to have the greatest number of terminological problems (due to the "pass-my-confusion-onto-the-next-person" phenomenon mentioned above); and finally, these people usually have the least amount of domain expertise (editors, proofreaders and translators are typically language experts, not domain experts). A lack of centralized information about terms is also a drawback for newcomers to a project, since it forces them to acquire knowledge about terms almost from scratch.

3. A GENERIC TOOL FOR TERMINOLOGY-ORIENTED KNOWLEDGE MANAGEMENT

As has been argued elsewhere (e.g. Ahmad *et al.* 1989, Czap and Nedobity 1990, Meyer and Paradis 1991, Parent 1989, Skuce and Meyer 1990a/b, Wijnands 1989), the knowledge management problems of terminology are not unique to this field. Rather, they are general problems of knowledge engineering that are now receiving extensive attention in the literature of AI. The AI research group at the University of Ottawa, Canada, has over the past few years developed a generic knowledge engineering tool called *CODE* (Conceptually Oriented Design Environment, Skuce *et al.* 1989), which is written in Smalltalk and runs on a Macintosh, 386 or UNIX platform. CODE can be described as a generic knowledge manager, designed to assist any person (including the non-expert) faced with the task of acquiring, formalizing, refining and accessing the knowledge structures of a specialized domain. CODE allows the user to construct a knowledge base which describes concepts in frame-like units called CDs (concept descriptors) that are normally, though not necessarily, arranged in inheritance hierarchies.

CODE has been tested in two terminology applications: a bilingual vocabulary project at the Department of the Secretary of State of Canada (Meyer and Paradis 1991, Skuce and Meyer 1990a/b) and a software documentation project at Bell Northern Research, the Canadian counterpart of Bell Labs (Skuce 1991). These two environments correspond to the two ends of the terminology spectrum described above. Based on what we learned during these experiences, we are now enhancing the system's terminological support in a new version of CODE (Version 4), expected to be operational in late 1991. Concurrently with system development, we are using CODE to build a prototype bilingual term bank, called COGNITERM, with a rich, highly structured and easily accessible knowledge component. In a nutshell, this term bank can be described as a hybrid between a traditional term bank (17) and a knowledge base.

Since a general technical description of the current and forthcoming versions of CODE are found elsewhere (Skuce *et al.* 1989, Skuce and Meyer 1991), we shall just outline below some of the features that are receiving particular attention in light of the fact that we intend COGNITERM to facilitate the management of both lexical-semantic and encyclopedic information, and to be usable across the spectrum of terminology environments.

User interface. Given the many different types of users that can be engaged in terminology-intensive activities, and the fact that we see a knowledge-based term bank as both a communication tool (e.g. between terminologists, between terminologists and experts, between the various "links" in a document-production "chain") and a teaching tool, the user interface has been a top priority in system development from the start. Hence, the current version of CODE is already user-tailorable, i.e., the same knowledge base is accessible in different manners for different purposes. For example, a domain expert or a terminologist who is highly experienced in a domain will have a different set of options than a learner. In the current version of CODE, we have also placed a strong emphasis on graphical representation. The system can easily produce various types of semantic net diagrams, for both hierarchical and non-hierarchical relations. The graphical display updates automatically when changes are made to the knowledge base, and offers mechanisms for focussing on certain parts of the knowledge base, highlighting special concepts (e.g. concepts that are uncertain, unconfirmed, etc.), and comparing and contrasting knowledge substructures. In CODE Version 4, Hypercard-like bit map images will be available, so that one can ask of a term "show me one", or of an image "what is this called?"

Access to, and navigation through, the knowledge base. Since a knowledge base incorporates large amounts of encyclopedic information, and since different users will

require different information, it is important that the knowledge be easily accessible and navigable. A CODE knowledge base is essentially a hierarchically organized hypertext-like system, incorporating the notion of property inheritance. One may navigate in whatever manner is appropriate, with typical retracing abilities of hypertext systems. Unlike traditional term banks, in which access is strictly terminological (i.e. one must know a term in order to get conceptual information about it), CODE allows conceptual characteristics to be entry-points into the knowledge, so that one can ask questions like "what is the term for the machine with function X", "what is the term for the material with physical properties X, Y, Z?" Access to, and navigation through, the knowledge is facilitated by the graphical component described above, and also through a browsing capability. In Version 4, the browser will use a basic window whose behaviour is modelled after an outline processor, with the ability to dynamically expand and contract tree-structures. The user can easily tailor-make the browser to suit a given need. To facilitate the use of terms as entry-points into the knowledge, the current version of CODE has a search/rename browser that permits scanning of the entire knowledge base for every occurrence of a term, and can be restricted to certain contexts (e.g. concept names, names of conceptual characteristics, descriptions of conceptual characteristics) to speed up the search. Version 4 will include a clearly defined set of terminological "status levels", by which we mean attributes of a term such as how it is used (e.g. as a concept name or the name of a conceptual characteristic), whether it is defined or not, whether it is used in definitions but is not a knowledge base concept or property, etc.

Informal, trial-and-error knowledge experimentation. The system contains features, which we are still developing, for managing knowledge that is in different states of "clarity" (for want of a better term). Lack of clarity may be due to several causes: for example, a terminologist may be unclear about a concept because he/she does not have the domain expertise to understand it properly; a technical writer, translator, etc. in the document-production chain may be unclear about a concept because people at various preceding links in the chain have used a term inconsistently; a concept may be very new (e.g. in the case of neologisms) and thus intrinsically unclear; and so on. In all these situations, we find problems such as what to call a concept, what the superconcept is, what subconcepts it has, what characteristics it has, what the similar concepts are. CODE permits rapid entry of hunches, guesses, trials, etc., followed by experimentation with the consequences of entering new knowledge. For example, superconcept links may be changed on the graph just by dragging, and the consequences can be seen immediately in textual or graphical form. One may ask for "similar" concepts, or potential terminological conflicts. Previously made changes (up to three) can be discarded in one click.

Multidimensionality. It is well known that concepts and entire knowledge structures can be "seen" from various "viewpoints" (18), which correspond roughly to the needs or interests of the knowledge base user. CODE offers a "masking" facility that allows one to restrict what is visible in the knowledge base by Boolean conditions on concepts and characteristics. For example, different users might require different types of knowledge about a certain laboratory procedure. CODE allows one user to say "show me only things about this laboratory procedure related to the *tools* that are required", and another to say "show me only things related to the *types of organisms* that the procedure can identify". The masking facility also allows the notion of viewpoint to be extended to include a notion of depth of domain expertise. For example, the user may request information about the laboratory procedure that would be of interest (and understandable) to a beginning biology student, or to a seasoned researcher.

Ranking of conceptual characteristics. We are currently investigating the usefulness of ranking characteristics according to where they fall in the lexical-semantic/encyclopedic continuum. For certain purposes (e.g. users with different levels of domain expertise), it may be useful to at least distinguish between characteristics that are necessary and sufficient, those that are encyclopedic but useful to establishing interlingual

equivalence, and finally, all other encyclopedic characteristics. We are also investigating an algorithm proposed by Maybury (1990) for ranking characteristics according to concept similarity on the one hand (e.g. similarity of characteristics of co-ordinate concepts), and prototypicality on the other (e.g. the degree to which a concept's characteristics are reflected in its subordinate concepts), which offers the possibility of generating definitions of the genus-differentia type automatically.

Multiple knowledge bases. Facilities for managing multiple knowledge bases (under development for Version 4) are required in order to work in multidisciplinary fields, and in order to work multilingually (since knowledge structures rarely correspond perfectly from one language to another). Both situations require support for isolating areas of correspondence and non-correspondence, and for comparing and contrasting. Multilingual work will require support for automatically generating some knowledge substructures (i.e. those that do correspond for the most part); eventually, this would involve a machine translation component. CODE already includes a general high-level ontology, which is being regularly refined; it will eventually serve as a basis for integrating knowledge bases.

Quality control. The envisaged use of the system by various persons in a terminology environment necessitates a sophisticated capacity for quality control. CODE offers a capacity for detecting conceptual inconsistencies of various types, carrying out type checking, flagging entries as to source, entry person, date, state of correctness, etc. Database-like retrieval facilities permit queries such as "show me all entries about laser printers made by X since last month and not yet approved". In order to ensure terminological consistency, CODE offers a number of features for assisting users in naming a conceptual characteristic (a common terminological problem in knowledge base building). The system can display all currently used names of similar properties (e.g. all properties belonging to the same category of property), and will prompt if this property name has already been used elsewhere.

SUMMARY

We have described two issues that must be considered in the development of technology for managing knowledge in terminology-intensive environments: 1) the importance of encyclopedic as well as lexical-semantic knowledge, and 2) the wide spectrum of working environments in which terminological activities can be carried out. We have also described a number of features currently available and under development in a generic knowledge management tool called *CODE*. Many of the ideas described represent work in progress related to the design of a prototype knowledge-based term bank using the CODE system. The problems are difficult, and our ideas constantly evolving: any comments on this work would therefore be warmly welcomed.

ACKNOWLEDGEMENTS

We gratefully acknowledge the many useful suggestions of Douglas Skuce about this paper specifically, and also about the general research framework (both technical and methodological) for the COGNITERM project. Roda Roberts provided some useful ideas for Section 1 of this paper. Karen Eck and Lynne Bowker helped with proofreading and text preparation.

The COGNITERM project (1991-1994) is supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and Research Services of the University of Ottawa. Testing of the CODE system in terminology-intensive environments has been supported by the Terminology and Linguistic Services Directorate of the

Department of the Secretary of State of Canada, and by Bell Northern Research. Development of the CODE system and associated methodology is being supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University Research Incentives Fund (URIF) of the Government of Ontario, Bell Northern Research, and Research Services of the University of Ottawa.

NOTES

1. Discussions of this debate as it has been articulated within a lexicographic framework can be found, for example, in Bierwisch and Kiefer 1970, Haiman 1980, Frawley 1981. More references can be found in Haiman 1980.

2. In Canada, most terminologists have university-level training in Translation, since most of them do bilingual (French-English) terminology. Some universities even offer an M.A. in Terminology. Canada's largest professional organization of translators, the *Société des traducteurs du Québec*, offers a standardized exam for certification in terminology.

3. For example, at the Terminology and Linguistics Services Directorate of the Department of the Secretary of State of Canada, an 85-page handbook prescribes the methodology. Questions of methodology are also outlined in some well-known textbooks on terminology, e.g. Rondeau 1984, Wüster 1974, Sager 1990.

4. A general overview of the Terminology and Linguistic Services Directorate can be found in Gawn 1990.

5. We use the term *linguistic* (in opposition to *conceptual*) in the very general sense of 'related to the terms, as opposed to the concepts designated by the terms'. Of course, any attempt to separate terms from concepts, while useful for analysis, is fundamentally artificial: terms and concepts are intimately linked.

6. *Thematic* is opposed to *term-oriented*. In the latter type of terminology, work is done on isolated terms, normally in response to specific requests from clients.

7. A more detailed analysis can be found in Meyer 1991.

8. In effect, terminologists experience most of the problems of expertise elicitation summarized, for example, in Gaines 1990.

9. Semi-automation of the scanning process in terminology is the object of considerable research interest at the moment. For an overview, cf. Auger 1989. Eventually, the knowledge management tool we envision would be integrated with technology for semi-automated scanning.

10. Noting a variety of contexts for the terms is extremely important in terminology, since contextual analysis is the principal way in which terminologists determine the meanings of terms, and also how they are used in context. Most term banks include some contexts in the terminology records.

11. We say *begin to* analyze because analysis is still not very deep at this stage, given that scanning involves rather fast reading. The cognitive processes that occur at this stage are

not yet very well understood, but we are assuming that some level of analysis, even if it is partly unconscious, does happen.

12. According to the ISO International Standard 1087, a *co-ordinate concept* is a "concept in a hierarchical system which ranks at the same level as one or more other concepts."

13. A neologism is a term that is used in a new sense. Sometimes existing terms can be used (and just given an additional sense), and sometimes a totally new term is coined.

14. For terminological simplicity, we use the term *product* to include things like services, regulations, committee decisions, etc. (and not just *products* in the strictest sense of the term).

15. A more detailed overview of the various types of writing activities in a document-production chain can be found in *Language Technology* (April 1989, Special Issue on Documentation).

16. Terminological problems are such an impediment to translation, and particularly to machine translation, that many organizations are starting to impose an in-house "controlled language" on their technical writers. An interesting overview of this phenomenon can be found in Pogson 1988.

17. The traditional term bank we are using as a model is TERMIUM III, the Secretary of State's bilingual (French-English) term bank. COGNITERM will include all the information categories currently available in TERMIUM.

18. This idea is very similar to "view" in Murray and Porter 1989.

REFERENCES

Ahmad *et al.* 1989. "Terminology and Knowledge Engineering: A Symbiotic Relationship Explained". Technical Report, KITES Project, University of Surrey, Guildford, Surrey.

Auger, Pierre. 1989. "Informatique et terminologie: revue des technologies nouvelles". *META* 34, 3.

Bierwish, M. and Kiefer, F. 1970. "Remarks on Definitions in Natural Language". In: F. Kiefer (ed.), *Studies in Syntax and Semantics*. Dordrecht, Reidel, 55-79.

Czap, Hans and Nedobity, Wolfgang (eds). 1990. *TKE '90: Terminology and Knowledge Engineering* (2 vols). Frankfurt: INDEKS Verlag. (Proceedings of the 2nd International Congress on Terminology and Knowledge Engineering, Trier, Oct. 1990).

Frawley, W. 1981. "In Defense of the Dictionary: A response to Haiman." *Lingua 55:1*, pp. 375-383.

Freibott, Gerhard, and Heid, Ulrich. 1990. "Terminological and Lexical Knowledge for Computer-Aided Translation and Technical Writing". *Proceedings of the Second International Conference on Terminology and Knowledge Engineering*. Frankfurt: INDEKS Verlag.

Gaines, Brian. 1990. "Knowledge Acquisition Systems," *Knowledge Engineering (Vol. 1: Fundamentals)*, Hojjat Adeli, Ed. New York: McGraw-Hill.

Gawn, Peter. 1990. "Tools for Terminology". *META* 33, 2, pp. 452-455.

Haiman, J. 1980. "Dictionaries and Encyclopedias". *Lingua 50:4*, 329-357.

Knowles, Francis. 1988. "Lexicography and Terminography: A rapprochement?" *Zurilex '86 Proceedings*. Ed. Mary Snell-Hornby. Tubingen: Francke Publishers.

Maybury, Mark T. 1990. "Generating Natural Language Definitions from Classification Hierarchies", *Proceedings of the 1st ASIS SIG/CR Classification Research Workshop*, Toronto, pp. 101-108.

McArthur, Tom. 1986. *Worlds of Reference: Lexicography, Learning and Language from the Clay Tablet to the Computer*. Cambridge University Press.

McNaught, John. 1990. "Reusability of Lexical and Terminological Resources: Steps Towards Independence". *Proceedings of the International Workshop on Electronic Dictionaries* (Oiso, Kanagawa, Japan, Nov. 1990). Japan Electronic Dictionary Research Institute Technical Report TR-031.

Mel'cuk, Igor. 1988a. "Principes et critères de description sémantique dans le DEC". In: *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques II*. Mel'cuk *et al.* Montreal: Presses de l'Université de Montréal.

Mel'cuk, Igor. 1988b. "Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria". *International Journal of Lexicography*, Vol. 1, 3.

Meyer, Ingrid. 1991. "Concept Management for Terminology: A Knowledge Engineering Approach". Paper presented at the ASTM Symposium on Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results (Cleveland, Ohio, June 1991).

Meyer, Ingrid and Paradis, Line. 1991 (in press). "Applying Knowledge-Engineering Technology to Terminology: A Pilot Project". *Terminology Update*. Ottawa: Department of the Secretary of State of Canada.

Meyer, Ingrid and Skuce, Douglas. 1990. "Computer-Assisted Concept Analysis for Terminology: A Framework for Technological and Methodological Research". Paper presented at the Fourth International Congress of the European Association for Lexicography (EURALEX 90). Malaga, Spain, Aug. 28 - Sept. 1, 1990.

Meyer, Ingrid, Miller, David and Michaud, Diane. 1991 (in press). "Terminologie et analyse notionnelle assistée par ordinateur". *Actes du colloque international sur les industries de la langue* (Montreal, Nov. 1990).

Murray, Kenneth and Porter, Bruce. 1989. "Controlling Search for the Consequences of New Information During Knowledge Integration". *Proceedings of the 6th International Workshop on Machine Learning*.

Parent, Richard. 1989. "Recherche d'une synergie entre développement linguistique informatisé et systèmes experts : importance de la terminologie". *META*, 34, 3.

Pogson, Geoff. 1988. "Controlled English: Enlightenment Through Constraint". *Language Technology*, Vol. 6, pp. 22-25.

Rondeau, Guy. 1984. *Introduction à la terminologie*. Chicoutimi: G. Morin.

Sager, Juan. 1990. *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Skuce, Douglas. 1991 (in preparation). "Experiences in Acquiring Knowledge About Industrial Software" (working title).

Skuce, Douglas. 1989. "A Generic Knowledge Acquisition Environment Integrating Natural Language and Logic". *Proceedings IJCAI Workshop on Knowledge Acquisition* (Detroit, Aug. 1989).

Skuce, Douglas and Meyer, Ingrid. 1990a. "Concept Analysis and Terminology: A Knowledge-Based Approach to Documentation". *Proceedings of the Thirteenth International Conference on Computational Linguistics* (COLING 90).

Skuce, Douglas and Meyer, Ingrid. 1990b. "Computer-Assisted Concept Analysis: An Essential Component of a Terminologist's Workstation". *Proceedings of the Second International Congress on Terminology and Knowledge Engineering Applications* (TKE 90), Frankfurt: INDEKS Verlag.

Skuce, Douglas and Meyer, Ingrid. 1991. "Terminology and Knowledge Acquisition: Exploring a Symbiotic Relationship". Paper submitted to the 6th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, to be held in Banff, Canada, October 1991.

Skuce, Douglas and Monarch, Ira. 1990. "Ontological Issues in Knowledge Base Design: Some Problems and Suggestions". *Proceedings of the 5th Workshop on Knowledge Acquisition for Knowledge-Based Systems* (Banff, Canada, Oct. 1990). Also available as Technical Report CMU-CMT-90-119, Centre for Machine Translation, Carnegie Mellon University, Pittsburgh.

Skuce, Douglas, Wang, S., and Beauvillé, Y. 1989. "A Generic Knowledge Acquisition Environment for Conceptual and Ontological Analysis". *Proceedings Knowledge Acquisition for Knowledge-Based Systems Workshop* (Banff, Canada, Oct. 1989).

Vinay, J.-P. and Darbelnet, J. 1976. *Stylistique comparée du français et de l'anglais: méthode de traduction.* Paris: Didier.

Wijnands, Paul. 1989. "Systèmes experts et terminologie". *META*, 34, 3.

Wüster, Eugen. 1974. "Die allgemeine Terminologielehre - ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften". *Linguistics 119* (1974), pp. 61-106.