TOWARDS A NEW GENERATION OF TERMINOLOGICAL RESOURCES: AN EXPERIMENT IN BUILDING A TERMINOLOGICAL KNOWLEDGE BASE

INGRID MEYER, DOUGLAS SKUCE, LYNNE BOWKER AND KAREN ECK
Artificial Intelligence Laboratory, University of Ottawa, Ottawa, Canada
ixmal@acadvm1.uottawa.ca

ABSTRACT

This paper describes a project to construct a terminological knowledge base, called COGNITERM. First, we position our research framework in relationship to recent developments in computational lexicology and knowledge engineering. Second, we describe the COGNITERM prototype and discuss its advantages over conventional term banks. Finally, we outline some of the methodological issues that have emerged from our work.

0 INTRODUCTION

The discipline of terminology[1] has received surprisingly little *focussed* attention in the literature of computational linguistics - an unfortunate situation given that NLP systems seem to be most successful when applied to specialized domains. We say *focussed* attention because when specialized lexical items are discussed in the literature, the research problems are often not clearly differentiated from the problems of non-specialized lexical items. A fundamental assumption of our research is that, while terminology can certainly benefit from advances in computational lexicology, it nonetheless has its own non-trivial research problems, which are ultimately related to the quantity and types of *specialized world knowledge* that terminological repositories must contain.

At the Artificial Intelligence Laboratory of the University of Ottawa, we are constructing a new type of terminological repository, COGNITERM, which is essentially a hybrid between a term bank and a knowledge base, or a *terminological knowledge base (TKB)*. COGNITERM is a bilingual (French/English) TKB constructed using a generic knowledge engineering tool (CODE) that has been used in terminology, software engineering and database design applications. The COGNITERM Project (1991-94) is focussing on the domain of optical storage technologies (e.g. optical discs, drives, processes, etc.).

In Section 1 of the paper, we position our research in relation to recent developments in computational lexicology and knowledge engineering; in Section 2, we describe the structure of COGNITERM as well as some of its advantages over conventional term banks; in Section 3, we outline some methodological issues that have emerged from our work.

[1] Space constraints preclude even a brief description of the discipline of terminology. Cf. Sager 1990.

1 RESEARCH ISSUES IN COMPUTATIONAL TERMINOLOGY

1.1 Terminological *vs.* Lexical Knowledge Bases

Much of the world's terminological data is stored in large terminological databases (TDBs) such as Canada's TERMIUM III, which contains over one million bilingual records. These TDBs are useful only to humans, and even then to only a small subset of potential users: translators remain the principal user category, even though TDBs have obvious applications in technical writing, management information and domain learning, not to mention a wide variety of machine uses such as information retrieval, machine translation and expert systems. A major weakness of TDBs is that they provide mainly *linguistic* information about terms (e.g. equivalents in other languages, morphological information, style labels); *conceptual* information is sparse (limited to definitions and sometimes contexts), unstructured, inconsistent and implicit.

Given these problems, a growing number of terminology researchers are calling for the evolution of TDBs into a new generation of terminological repositories that are knowledge-based. Since this vision of a TKB has been recently paralleled in computational lexicology by the vision of a *lexical knowledge base* or LKB (e.g. Atkins 1991, Boguraev and Levin 1990, Pustejovsky and Bergler 1991), we would like to briefly position our research framework in relation to these developments.

The LKB projected by Boguraev and Levin 1990 differs from an LDB in two ways: 1) the LDB states lexical characteristics on a *word-by-word* basis, while the LKB permits *generalizations*; and 2) the LKB permits *inferencing*, and thus the possibility of *dynamically extending* the

lexicon to accommodate new senses. Both characteristics are extremely important for the TKB as well: 1) a capacity for supporting generalisations is particularly relevant to terminology since terminological repositories have an important teaching function[2]; and 2) the accommodation of new senses is even more crucial to terminology than to the general lexicon since specialized languages grow so rapidly. While the TKB must share these characteristics, it differs from the LKB in one important way, which derives from the fundamental difference between general and specialized lexical items. This difference can be summarized in the following two principles:

- *an LKB must make explicit what a native speaker knows about concepts denoted by general lexical items*
- *a TKB must make explicit what a native speaker who is also a domain expert knows about concepts denoted by specialized lexical items*

While the lexicographer's ultimate source of lexical knowledge is his/her own intuition, the terminologist's challenge is to model experts' terminological intuitions, which stem in large part from their domain knowledge. The acquisition of domain knowledge, therefore, has traditionally been the starting point for any practical terminology project; only when the knowledge structures of a domain are systematized to some degree can terminologists proceed with term extraction, definition construction, analysis of synonymy and polysemy, identification of equivalents in other languages, etc. The crucial importance of modelling domain knowledge in a TKB necessitates a conceptual framework and technology which, in our view, should derive partly from recent insights in knowledge engineering.

1.2 Terminology and Knowledge Engineering

At the heart of the relationship between terminology and knowledge engineering is the fact that practitioners of both disciplines function as *intermediaries* in a knowledge communication context involving experts on the one hand and a knowledge processing technology on the other. This type of knowledge communication context entails three principal activities:

Knowledge acquisition. Acquisition of knowledge, whether by elicitation from a human expert or extraction from texts, is complicated by the fact that domain expertise consists of three elements - performance, understanding and communication - that require the expert to play the roles of practitioner, scientist and teacher, respectively (Gaines 1990). Unfortunately, experts vary widely in their teaching skills: they may not have the linguistic ability to express knowledge clearly; they may not provide exactly the knowledge that is required; etc. As well, they may vary in their understanding of the field, presenting the knowledge engineer/terminologist with problems of inconsistency and contradiction.

Knowledge formalization. Knowledge does not come "off the shelf, prepackaged, ready for use" (Hayes-Roth 1987:293). As already mentioned, it can be inconsistent and contradictory. It can be multidimensional, since experts' understanding of a conceptual system can depend on their point of view. It may be hard to "capture", since it is constantly changing, and since emergent knowledge can be incomplete and unclear. Finally, from the knowledge engineer/terminologist's point of view, it will exist in various degrees of "clarity" and "depth": since knowledge acquisition is incremental, certain concepts will be more clearly or deeply understood than others at any given time.

Knowledge refinement. Once formalized, knowledge may be refined in two ways: 1) it may be validated by testing the knowledge-based system on the intended application, and/or 2) it may be periodically updated, for example, as the knowledge engineer/terminologist's understanding of the field deepens or expands, when the field itself changes, or when the system needs more knowledge due to changes in the application. Knowledge refinement may again entail knowledge acquisition and formalization, making the knowledge engineering cycle a continuous process.

Over the last three years, we have developed and tested a knowledge engineering tool called *CODE* (Conceptually Oriented Description Environment), which is designed to assist a user who may or may not be a domain expert in acquiring, formalizing and refining specialized knowledge. Although generic by design, CODE emphasizes linguistic and particularly terminological support, which we feel is crucial to all knowledge engineering applications. From 1987 to 1990, a working prototype was developed and tested in three terminology-intensive applications: term bank construction, software engineering and database

[2] Most TDB users are not domain experts, and thus hope to acquire some domain knowledge when they look up a term.

design[3]. Our research has now entered a second three-year phase, with the goal of using CODE to help us develop a clearer concept of a TKB and of an associated methodology.

2 COGNITERM: A TERMINOLOGICAL KNOWLEDGE BASE

2.1 General Description

COGNITERM is essentially designed as a hybrid between a conventional TDB and a knowledge base. Each concept is represented in a frame-like structure called a *concept descriptor (CD)*, which has two main information categories. The Conceptual Information category is the knowledge base component, listing conceptual characteristics and their values. CDs are normally, though not necessarily, arranged in inheritance hierarchies. The Linguistic Information category is the TDB component, providing all the strictly linguistic information normally found in conventional TDBs.

The TKB can be visualized graphically in a variety of semantic net displays. Both hierarchical (e.g. generic-specific, part-whole) and non-hierarchical relations can be graphed. Since knowledge acquisition typically proceeds one subdomain at a time, subwindows may show only a restricted part of the knowledge structure (i.e. a subtree). There is also a masking capability which, for example, can show only concepts that fall within a given "dimension" of reality.

As an aid to definition construction, and specifically to assist in determining the differentiating characteristics, CODE offers a Characteristic Comparison Matrix that presents the union of all characteristics of coordinate concepts[4], with the exclusion of those that are identical in all coordinates.

Finally, navigation through COGNITERM is facilitated by CODE's Browser, which allows the knowledge to be accessed either by names of concepts or names of their characteristics, both of which can be presented in a conceptual (i.e. hierarchical) or alphabetical order. A variety of masks can be applied to restrict the knowledge.

2.2 Advantages of a TKB over a TDB

The differences between a conventional TDB and a TKB can be examined from three points of view: 1) the information itself, 2) support for acquiring and systematizing the information and 3) facilities for retrieving the information. A brief description[5] of each is found below.

The information. In a TDB, conceptual information is encoded implicitly in the form of definitions, contexts, indication of domain(s), etc. In a TKB, it is encoded explicitly. The resultant *degree of structure* imposed on the information has three important by-products. First, it allows for an explicit representation of conceptual relations (as opposed to implicit representations in TDB definitions or contexts). Second, it facilitates consistency: since generic concepts are explicitly indicated, for example, definitions of all coordinate concepts must have the same genus term; since characteristics inherit to subconcepts, they will correspond from one coordinate concept to another. Third, an explicit representation of conceptual relations facilitates graphical representations of knowledge structures; this aspect is particularly emphasized in the COGNITERM Project since graphical representations aid learning, providing the kind of conceptual "map" advocated by numerous educational psychologists[6].

Acquisition and systematization of information. Unlike conventional TDBs, a TKB such as COGNITERM provides not only a medium for storing information, but also mechanisms to assist in acquiring and systematizing the information in the first place. Inheritance mechanisms play an important role in this regard: on the simplest level, they free the terminologist from repeating information from one hierarchical level to another, and allow the possibility of "what-if" experiments; on a more interesting level, inheritance can be associated (as it is in CODE) with mechanisms for signalling conflicts when changes to one hierarchical level "percolate" through the knowledge structures. A browsing mechanism such as we have implemented provides additional support for acquisition, as it allows the kind of hypertext-like "navigation" through the knowledge structures that is needed to ferret out compatible knowledge "spaces" for a new concept. Other implemented

---

[3] This first phase of our research has already been documented elsewhere: a general technical description of CODE can be found in Skuce (in press b); an analysis of the relationship between terminology and knowledge engineering can be found in Meyer 1991 and (in press); the three terminology-intensive applications are described in Skuce and Meyer 1990a/b (term bank construction), Skuce (in preparation) (software engineering), and Downs *et al.* 1991 (database design).

[4] By *coordinate concepts* we mean concepts that share the same parent in a hierarchy.

[5] A much more detailed description, illustrated with examples of COGNITERM output, can be found in Meyer *et al.* 1992 (in press).

[6] Cf. Sowa 1991 (in press).

user interface features, such as masks, the Characteristic Comparison Matrix, and a highly developed graphical display, are just some examples of the potential facilities of a TKB environment designed to help terminologists "get the knowledge straight" throughout the acquisition process.

Retrieval of information. Conventional TDBs are severely handicapped by their fundamental term-to-concept orientation: knowing a term, one can expect the TDB to indicate (to some degree, at least) what it means, what its synonyms are, etc. Terminological research, however, is very often concept-to-term oriented: for example, "real-life" terminology is typified by questions like "What do you call the machine with function W?", "What do you call the material that has physical characteristics X, Y, and Z?" The inability of conventional TDBs to answer these kinds of questions leads to the proliferation of synonyms and quasi-synonyms, one of the greatest impediments to communication in specialized domains. Users of COGNITERM can access its data through any conceptual characteristic to determine whether the concept they have in mind already has a name.

3 METHODOLOGICAL ISSUES

In deciding on a preliminary methodology for our work, we naturally turned to the literature of both computational lexicology and knowledge engineering for inspiration, with little success. Even the world's largest knowledge acquisition project, CYC (Lenat and Guha 1990), provides only sparse methodological guidance (Skuce in press a). To date, our methodology has remained essentially grounded in that traditionally used by terminologists (Sager 1990), a reasonable starting point when one considers that, although terminologists have traditionally not built TKBs, conceptual analysis has always been a central part of their work[7] nonetheless. Terminologists are keenly aware of the importance of a certain depth of domain knowledge, and many of the conceptual analysis techniques that are advocated in the knowledge engineering literature - e.g. describing conceptual characteristics through attribute-value pairs, sketching concept networks - have been part of the terminological methodology for years.

The methodology we have developed can be very superficially described as follows[8]: 1) After introductory reading on the domain, the principal conceptual relations are sketched out, with the goals of establishing the boundaries of the domain and identifying the subdomains, from which the most fundamental is selected for further analysis. 2) A template of conceptual characteristics is established for the selected subdomain; it is used as a guide to the knowledge acquisition process, and inherits to lower levels of the conceptual hierarchy, where it can also be specialized. 3) Conceptual and linguistic information are entered into the system as they are acquired (mainly from the corpus). A concept is integrated into a hierarchy whenever its superconcept is known; when it is not, or when there is some doubt, the concept is labelled "unclassified" (unclassified concepts can occur at any level in a hierarchy, i.e. there can be different "degrees" of classification). 4) Intensional definitions are constructed with the help of the Characteristic Comparison Matrix. Steps 2-4 are then repeated for the next subdomain, until all subdomains have been completed.

A number of the more troublesome methodological issues with which we are currently grappling are briefly outlined below.

Knowledge acquisition "paths". Knowledge acquisition is not a journey down a straight path: there is no visible "goal". Although we have followed traditional terminology methodology in adopting a subdomain-oriented, top-down approach to acquisition, it often seems desirable to deviate from the principal subdomain when one encounters related terms in a neighbouring subdomain or field, and to work bottom-up as well as top-down within the principal subdomain. While the subdomains we have investigated so far (the majority of the concepts belonging to the semantic class of artefacts) are dominated by generic-specific and part-whole relations, subdomains related to other semantic classes may be more amenable to analysis based on different relations, as has been pointed out, for example, in the literature on the WordNet project (Miller 1991, Fellbaum 1991).

Multidimensionality. While terminologists are well aware that a given domain can be subdivided in different ways, depending on the expert's point of view, they have not traditionally attempted to account for it in any serious way, since this is difficult to do with pencil-and-paper techniques. Some problems that arise are how such "multidimensionality" affects knowledge acquisition "paths", how the technology can better support the maintenance of conceptual clarity as the number of dimensions grows (for example, through masking facilities of the kind we have implemented), how multidimensionality can be reflected in definition construction, etc.

[7] A detailed analysis of the role of conceptual analysis in terminology can be found in Meyer (in press).

[8] A detailed description of the methodology can be found in Meyer *et al.* 1992 (in press).

Validation. Validation by experts and other terminologists, which has always been an important part of terminology work, is complicated in our approach by the fact that our TKB is very hypertext-like, and thus requires revision techniques that go beyond those normally applied to "flat" texts such as conventional terminology records. We need to investigate further at which points validators should be consulted, what elicitation techniques should be used at each point, how to handle inconsistencies in opinion, etc.

Increased automation. To date, our research efforts are oriented towards *facilitating* (and not *automating*) the knowledge acquisition process for developing and implementing our concept of a TKB. This is consistent with the majority of knowledge acquisition projects in the world, including CYC. As the concept of a TKB becomes clearer, however, we hope that TKB and LKB researchers will collaborate in exploring possibilities for a more automated approach to acquisition.

ACKNOWLEDGEMENTS

The COGNITERM Project is supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and Research Services of the University of Ottawa. Development of CODE is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University Research Incentives Fund of the Government of Ontario, Bell Northern Research and Research Services of the University of Ottawa.

REFERENCES

ATKINS, B.T.S. 1991. "Building a Lexicon: The Contribution of Lexicography". *International Journal of Lexicography*, Vol. 4, No. 3.

BOGURAEV, Branimir and LEVIN, Beth. 1990. "Models for Lexical Knowledge Bases". *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research.* Waterloo: University of Waterloo.

DOWNS, Mary, GREENE, Reid and RISHEL, Diane. 1991. "Conceptual Data Modeling in a Materials R&D Organization". *29th Annual Symposium of the National Institute of Standards and Technology.* Washington, D.C.

FELLBAUM, Christiane. 1991. "English Verbs as a Semantic Net". *International Journal of Lexicography*, Vol. 3, No. 4.

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

HAYES-ROTH, F. 1987. "Expert Systems". *Encyclopedia of Artificial Intelligence.* Ed. Stuart Shapiro. New York: John Wiley and Sons.

LENAT, D. and GUHA, R. 1990. *Building Large Knowledge-Based Systems.* Reading, MA: Addison Wesley.

LEVIN, Beth. 1991. "Building a Lexicon: The Contribution of Linguistics". *International Journal of Lexicography*, Vol. 4, No. 3.

MEYER, Ingrid. (in press). "Concept Management for Terminology: A Knowledge Engineering Approach". *Proceedings of the Symposium on Standardizing Terminology for Better Communication: Practice, Applied Theory and Results* (Cleveland, Ohio, June 1991). Special Technical Publication of the American Society for Testing Materials (ASTM).

MEYER, Ingrid. 1991. "Knowledge Management for Terminology-Intensive Applications: Needs and Tools". *Proceedings of the ACL SIG Workshop on Lexical Semantics and Knowledge Representation.* (To appear as a book edited by J. Pustejovsky and S. Bergler and published by Springer Verlag.)

MEYER, Ingrid, BOWKER, Lynne and ECK, Karen. 1992 (in press). "COGNITERM: An Experiment in Building a Terminological Knowledge Base". *Proceedings of the Fifth Euralex International Congress.*

MILLER, George. 1991. "Nouns in WordNet: A Lexical Inheritance System". *International Journal of Lexicography*, Vol. 3, No. 4.

PUSTEJOVSKY, James and BERGLER, Sabine (Eds.). 1991. *Proceedings of the ACL SIG Workshop on Lexical Semantics and Knowledge Representation.* (To appear as a book published by Springer Verlag.)

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

SKUCE, Douglas. (in press a). "A Review of: Building Large Knowledge-Based Systems (Lenat and Guha)". *The Journal of Artificial Intelligence.*

SKUCE, Douglas. (in press b). "A Wide Spectrum Knowledge Management System". *Knowledge Acquisition.*

SKUCE, Douglas. (in preparation). "Managing Software Design Knowledge: A Tool and an Experiment".

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.* Frankfurt: Indeks Verlag.

SOWA, John. 1991 (in press). "Conceptual Analysis as a Basis for Knowledge Acquisition". In *The Cognition of Experts: Psychological Research and Empirical AI*, Ed. R. R. Hoffman. Berlin: Springer Verlag.