Computer-assisted concept analysis for terminology:
a framework for technological and methodological research
Ingrid Meyer - Douglas Skuce

Abstract

Concept analysis, a crucial component of the terminologist’s work, is clearly the most neglected area of computational terminology research. To begin filling this gap is one of the objectives of the Artificial Intelligence Laboratory at the University of Ottawa, Canada, where we have developed a generic knowledge engineering tool that has already undergone preliminary testing in two terminology applications. Drawing on this work, we propose a general framework for research on computer-assisted, terminology-oriented concept analysis. Our framework describes a technological as well as a methodological avenue of investigation: we set forth a list of desiderata for a tool that will facilitate concept analysis, as well as for a methodology to support the tool.

Introduction

The goal of concept analysis for terminology is to represent the knowledge structures of specialized subject fields. This is done by describing conceptual properties, i.e. the attributes of individual concepts, and the relations that hold within systems of concepts. Since terminology has both a conceptual and a linguistic dimension, concept analysis plays a crucial role in numerous terminological activities, of which the following are just some of the most important:

- **Orientation of documentary research.** An understanding of a subject field’s knowledge structure is crucial to determining the limits of the field and the principal subfields.
- **Construction of definitions.** The classic «genus-differentia» definition requires a knowledge of generic-specific relations (for the genus term), and of conceptual characteristics of co-hyponyms (for the differentia). An extensional definition also requires a knowledge of generic-specific relations, since it lists hyponyms, i.e. subordinate terms. A contextual definition (which uses the term in an explicative context) requires an evaluation of potential contexts, which in turn requires an understanding of conceptual characteristics.

1. Various terms have been used in the linguistic, philosophical and Artificial Intelligence literature to designate what we term property, attribute and relation. We shall use property as a hyponym (i.e. generic term) for both attribute and relation. By attribute we mean a conceptual characteristic (e.g. age, sex, colour) that refers to a concept in itself, without involving a relation to another. By relation we mean a conceptual characteristic (e.g. spatial position, family relationship, cause) that belongs to a concept only in its relation to another.
2. Throughout this paper, we use terminology in a very broad sense, described in 1.1.
3. The practical applications of concept analysis that we propose derive partly from Picht and Draskau 1985, p. 92. For a comparison of the usefulness of concept analysis in term-oriented as opposed to subject-oriented terminology, See Meyer et al. 1991 (in press).
• **Description of neology.** Emerging concepts typically use as their «building blocks» conceptual characteristics of existing knowledge structures, i.e. those of the subject field into which a neologism is emerging, and/or those of a neighbouring subject fields.

• **Communication with subject field experts.** When the terminologist is not a subject field expert, a sound grasp of the knowledge structure of the field is essential for effective communication with experts, and for maintaining the terminologist’s credibility with these experts (this is important since experts are often extremely busy and, understandably, reluctant to explain basic concepts).

• **Multilingual work.** Univocal interlinguistic correspondence between terms is often not possible since knowledge structures frequently differ from one language to another. To account for such interlinguistic «mismatches», the terminologist must clearly understand the knowledge structures of the subfield for the languages concerned and isolate the areas of non-correspondence.

While the terminology literature has long stressed the importance of concept analysis, it clearly remains the most neglected area of computational terminology research. In the words of Sager (1990: 9):

> A great deal of attention has... been devoted to the structure of conceptual systems and the best way of representing them on paper. In this area in particular, conventional thinking is still largely dominated by pen and paper processing techniques and by the relatively simple relationships that suffice for structuring documentation thesauri and respond well to human attribution and processing.

Indeed, while terminologists recognize the importance of concept analysis, their methodology and the resultant quality of their work have been limited by «paper-and-pencil» and «do-it-in-my-head» techniques. While these techniques may be workable (though by no means ideal) approaches to concept management for subject field experts with years of training and experience in their fields, they are particularly problematic when the terminologist is not a subject field expert, and when the field is large, multidisciplinary, and rapidly evolving. To compound the problem, «non-expert» terminologists may even be required to work in several fields simultaneously or to take on new fields without any formalized introduction to their subject matter.

Although professional backgrounds and tasks may vary from one working environment to another, all terminologists share to some degree the fundamental pro-

4. Throughout this paper, for the sake of simplicity, we use the term **terminologist** in the very broad sense of any person engaged in terminology work, as described in 1.1.

5. This is the case, for example, in the Terminology and Linguistic Services Directorate of the Department of the Secretary of State of Canada, where terminologists are typically trained in translation or linguistics, and are therefore greatly dependent on collaboration with subject field experts.


7. For an explanation of the various problems entailed by non-computerized concept analysis, see Skuce and Meyer 1990b.
blems of concept analysis: how to acquire, systematize, and retrieve expert knowledge. Fortunately, these «knowledge processing» problems are not unique to terminology; rather, as is being stressed more and more in the terminology literature, they are general concerns of knowledge engineering that are now receiving extensive attention in Artificial Intelligence research.9

Consistent with this knowledge engineering view of terminology, the Artificial Intelligence Laboratory at the University of Ottawa, Canada, under the direction of Douglas Skuce, has over the past few years developed a generic knowledge engineering system called CODE (Conceptually Oriented Design Environment), which is particularly suited to terminology work since it calls for a linguistically-oriented methodology (Skuce and Monarch, 1990), and stresses conceptual clarity and terminological consistency. CODE, which is written primarily in Smalltalk and runs on a Macintosh, 386, or UNIX platform, is intended as a «knowledge processor» for any person (including the non-expert) faced with the task of concept analysis. The system has purposely been designed generically, i.e. with the idea of allowing eventual specialization for a number of possible applications.10 To date, CODE has received preliminary testing in two terminology intensive applications: a bilingual vocabulary project at the Terminology and Linguistic Services Directorate of the Department of the Secretary of State of Canada, and a software documentation project at Bell Northern Research (the Canadian equivalent of Bell Labs in the United States).

Since the technical details of CODE, the system’s particular suitability to terminology, and the two terminology-oriented test experiences have already been well documented,11 we shall not discuss them here. Rather, the purpose of this paper is to propose a very general framework for research in computer-assisted concept analysis based on our initial experience in this field. This framework, whose underlying premises are outlined in Section 1, comprises a technological as well as a methodological avenue of investigation: on the one hand (Section 2), we outline a number of desiderata for a tool that will facilitate terminology-oriented concept analysis (many of these desiderata have already been realized in the current version of the CODE system, cf. Skuce and Meyer, 1990b); on the other hand (Section 3), we propose a number of key issues for the development of a methodology to support the tool.


9. See, for example, Chorafas 1990, Lenat and Guha 1990, as well as the 1989 and 1990 Proceedings of the Annual Workshops on Knowledge Acquisition for Knowledge-Based Systems, held in Banff, Canada.

10. Some of the applications that we have in mind are software engineering, database design, teaching technical subjects, and standardizing military rules and regulations.

1. Basic Premises

The technological and methodological research framework outlined below is based on the following general premises about the types of terminology environments in which the tool and methodology could be used, and about the goal of computer-assisted concept analysis.

1.1. Terminology Environments

Terminology work can be performed in two types of working environment. In its most narrow sense, terminology can be described as a distinct specialization, i.e. as an activity carried out in environments where persons officially designated as terminologists (often with professional training and/or certification in terminology) are responsible for compiling information on specialized terms, which is typically stored in a database.

In its broadest sense, terminology can be described as part of the documentation chain, i.e. as just one component of a wide spectrum of document production and dissemination activities, including, for example, product design specification, technical writing, revision, proofreading, translation, abstracting, management information, and marketing. Conceptual and terminological consistency are crucial throughout all phases of the documentation chain in order to avoid the «pass-my-confusion-on-to-the-next-person» phenomenon that is particularly felt by those persons at the end of the chain, such as editors and translators, not to mention the poor users, i.e. readers, of the final product. The concept analysis tool we foresee would provide an interface between the various links in the documentation chain, ensuring conceptual and terminological consistency, and consequently, enhancing both internal and external communication.

Throughout this paper, we shall use the terms terminology and terminologist in their most general senses, to refer, respectively, to both types of terminology environment, and to all persons in either environment concerned with establishing, using, or verifying specialized terms.

1.2. Goals of Computer-assisted Concept Analysis

We foresee both a short-term and a long-term goal for computer-assisted concept analysis. In the short term, the goal is to improve both qualitative and quantitative aspects of conventional repositories (typically, term banks) of terminological data, which we shall describe very generally as repositories that are highly term- (as opposed to knowledge-) oriented, and that are not designed to be multifunctional and shareable. As regards quality, we assume that computer-assisted concept analysis will improve the various terminological activities mentioned in Introduction above, and consequently lead to more accurate and consistent linguistic output. As regards quantity, we assume that computer-assisted concept analysis will reduce the amount of time that is often wasted in terminology work, for example in correcting errors and inconsistencies, in dealing with communications problems due to conceptual and terminological confusion within a documentation chain, and in training new terminolo-
gists (who typically do not have formalized knowledge structures of their predecessors to draw on).

In the longer term, and consistent with an increasingly prevalent view (e.g. Auger, 1989; Czap and Galinski, 1988; Galinski, 1988a/b; Knowles, 1988; Parent, 1989; Sager, 1990; Wijnands, 1989), we assume that terminology banks are «on the threshold of significant transformation as the accent gets placed more fairly and squarely on knowledge engineering and transfer» (Knowles, 1988: 335). Hence, computer-assisted concept analysis should aim at contributing to the development of terminological knowledge bases, i.e. terminological repositories whose linguistic data is enhanced with a rich and formalized knowledge component that facilitates multifunctionality and shareability, allowing various possible applications such as documentation and management information (Galinski, 1988a, Meyer et al., 1991), training (Knowles, 1988: 334), technical writing (Freibott-Heid, 1990), expert systems (Wijnands, 1989), and machine translation (Isabelle, 1988).

2. A framework for technological research

The concept analysis tool we envisage could take two different forms. In the short term, we see it as a component of a terminology workstation —a «knowledge processor» that helps the terminologist acquire, systematize, and retrieve information about the conceptual structures of specialized fields. This knowledge processor could also serve as a training tool for a terminologist taking on a new field, as a basis for communication between a non-expert terminologist and a subject field expert, and as an interface between different links in a documentation chain. In the longer term, we envisage the tool as just one component of a terminological knowledge base, which would also have to provide linguistic, pragmatic, bibliographic and administrative data.

The following is an outline of the principal desiderata that we foresee for this tool. We shall restrict our discussion to design aspects of the tool itself. Space limitations do not allow us to discuss the important question of interface between the tool (as we envisage it for the short term) and other components of the terminology environment, particularly the terminology bank: nor will we address the matter of how the tool (as we envisage it for the long term) will be integrated into a terminological knowledge base.

**User interface.** As well as being user-friendly in the most general sense of this term, the technology must be flexible enough to suit (or at least to be specializable for) a wide range of contexts: terminology as a distinct specialization as well as terminology as part of the documentation chain; use by persons who know a subject field extremely well (e.g. true subject field experts), by persons who know it reasonably well (e.g. terminologists who have been working in a subject field for a long time), and by persons who do not know the field at all (e.g. a terminologist starting in the field); finally, it should be useful as a communication tool, for example as an interface between links in the documentation chain, as a basis for discussions between terminologists and experts, and as the foundation for work by standardizing committees.

**Textual representation.** The tool we envisage would create a knowledge base that stores textual information in units like frames (or frame-like structures). Conceptual properties (both attributes and relations) would thus correspond to slots. Inheritance
mechanisms would prevent manual repetition of information from one level to another and assure quality control (see below). Eventually, it would be desirable to establish a standard set of high-level concepts and properties for different subject fields (see Section 3). But terminologists also need the flexibility of being able to create their own properties for more specialized concepts. When changes are made to the textual form of the knowledge base, updating should occur automatically in the graphical display (see below), and vice versa.

**Graphical display.** In our experience (Skuce and Meyer, 1990b), terminologists place an extremely high value on graphical representations of knowledge structures: a «picture» of a field or subfield can give a quick overview, and help in understanding where a concept fits in. A sophisticated graphical display should have mechanisms for: a) focussing on certain parts of the graph (e.g. hiding subfields that are not currently being worked on, comparing groups of related concepts); b) isolating «special» (i.e. fuzzy, uncertain, unconfirmed) concepts or groups of concepts; c) showing, comparing, and contrasting multiple graphs; d) representing multidimensionality and fluidity (see below); and e) showing both hierarchical and non-hierarchical relations.

**Hypertext browsing.** When a subject field is very large, multidisciplinary or rapidly changing, when multiple knowledge bases must be consulted simultaneously (see below), or when the knowledge base is used as a learning tool, quick and easy navigation through the data is crucial. A «browser», such as the one already highly developed in the CODE system, would allow the terminologist free movement from one concept to another, from one property to another, and between concepts and properties—all of which can be visualized textually and/or graphically. Browsing should also be supported by a global lexicon facility that could search on any word in the knowledge base, and by a capacity for fuzzy searching.

**Multiple knowledge bases.** Two particular situations would require a multiple knowledge base capability. The first is the case of interdisciplinary or overlapping fields. The second is that of multilingual terminology work (since knowledge structures rarely correspond from one language to another). In both situations, one requires support for isolating areas of correspondence and non-correspondence, for comparing and contrasting, and for generating parts of knowledge structures automatically (i.e. when duplication exists). This would involve a machine translation component for interlingual work.

**Multidimensionality.** It is very common in terminology work to find that a field or subfield can be hierarchically partitioned in many different ways, depending on which properties of concepts are stressed. The complexity of this phenomenon is increased when concepts occur in numerous non-hierarchical systems. Multidimensionality can also occur when a concept is placed in more than one hierarchical relation (i.e. when it has more than one superordinate concept), a case which will call for the allowance of multiple inheritance of properties.

**Fluidity.** In the words of Sager (1990: 13), «knowledge structures are not absolute... conceptual systems are relatively fluid entities constantly undergoing change». The tool we envisage would allow the terminologist to modify knowledge structures as they change, or as his understanding of these structures deepens.

**Quality control.** Quality control mechanisms should focus on ensuring consistency, since any changes to one part of the knowledge base can have repercussions.

12. See Sager 1990, section 2.2.4, for a very clear introduction to this phenomenon.
throughout, and affect definitions as well. Quality control is especially important when a number of persons are contributing to the construction of the knowledge base. We envisage a sophisticated facility for consistency checking based on inheritance mechanisms, and involving queries whenever properties are changed or concepts moved. Inheritance mechanisms can also allow the terminologist to do «what-if» experiments when a concept’s position within a knowledge structure is unsure.

Support of definition construction. Since a definition is, in the words of Martin (1988), a «knowledge representation», i.e. a description of a concept based on an enumeration of a number of its properties or subconcepts, one would expect that some definitions could be constructed automatically once the knowledge base was constructed. This would be the case for the classic genus-differentia definition, which would require the concept’s superconcept and those properties that are not shared by its co-hyponyms; for the extensional definition, which would require the concept’s subconcepts; and for the encyclopedic definition, which would require the superconcept and a variety of properties (not just the distinguishing ones as in the genus-differentia definition). Mechanisms for supporting definition construction should also include quality control devices aimed at maintaining consistency and preventing circularity.

3. A framework for methodological research

To date there exists no standardized methodology for terminological concept analysis in general, let alone for computer-assisted concept analysis. With the latter, the need for methodology becomes essential, since computerization allows the possibility of distributed information collection, for example between non-expert terminologists and subject field experts, and between nations (e.g. for multilingual terminology, to get native speakers and high-quality documentation). Methodology will also become increasingly important in the construction of the shareable and multifunctional terminological knowledge bases of the future. The following is a brief description of some of the more important methodological issues that must be addressed.

Selection and evaluation of knowledge sources. Both texts and human experts will have to be considered as knowledge sources. Elicitation techniques will have to be investigated in the case of human experts, and eventually, the methodology will have to be compatible with semi-automatic acquisition from text.

Quantity and type of conceptual information. These will vary according to the type of human user, and particularly when machine uses (e.g. expert systems, machine translation systems) are envisaged. A methodology for deciding what conceptual information is required will first require an analysis of the needs of the various potential user categories (some of which will not become clear until the new generation of terminological knowledge bases is in place).

Bottom-up vs. top-down approaches. In our experience (Skuce and Monarch, 1990) knowledge entry tends to involve a mixture of top-down and bottom-up strategies, and very often, can proceed in a «middle-out» direction as well, supporting the arguments of Rosch (1978). A complete methodology should include a classification

13. See Sager 1990, Section 7.4, for an excellent overview of the different possible user types for both conventional and future term banks.
of typical knowledge acquisition problems, and specify one approach (or mixture of approaches) for each problem.

**Consensus on high-level concepts and properties.** Since these are the building-blocks of the knowledge base, it is crucial to develop strategies for achieving agreement on them, so that terminologists will have some guidance when they start the acquisition process.

**Lexical and syntactic conventions.** Property names, as well as the values of the properties, should be as clear and meaningful as possible, following an established syntax. Particular attention should be paid to avoiding polysemy and circularity, in keeping with lexicographic principles such as, notably, those of Meaning-Text lexicography (Mel'čuk, 1988). It may be useful to implement a restricted sublanguage to facilitate automatic logical checking and machine translation.

**Non-hierarchical relations.** While the importance and use of hierarchical relations (e.g. generic-specific, whole-part) has received extensive attention in the terminology literature, this has not been the case for non-hierarchical relations: a methodology for concept analysis will need strategies for integrating non-hierarchical relations into the knowledge base. This will require a classification of such relations (which in turn will require the analysis of many different subject fields since these relations tend to be largely field specific), and an analysis of their importance in definitions.

**Definition construction.** Once an enumeration of properties has been established for a concept, the terminologist needs to decide which properties to incorporate into the definition (this implies strategies for ranking properties). A methodology should provide guidelines for deciding which definition type is most appropriate to the concept and to the envisaged user, and subsequently, for deciding which properties are required for the selected definition type. It will be useful to develop definition templates for various terminological purposes; this will require a much more extensive typology of terminological definitions than currently exists.

**Validation techniques.** Analogous to software engineering, an unverified knowledge base is as undesirable as an untested program. A methodology, therefore, must present guidelines for deciding how to do controlled testing, against what, at what time (during and/or after construction), and by whom. Validation techniques will vary according to the knowledge sources used (e.g. texts only, combination of texts and human experts, or human experts only).14

Acknowledgements

Skuce's research and development of the **CODE** system have been supported by COG-NOS, Inc., the National Sciences and Engineering Research Council of Canada (NSERC), the University Research Incentives Fund (URIF) of the Government of Ontario, Canada, and Research Services of the University of Ottawa. Meyer's research has been supported by Research Services of the University of Ottawa and by the Centre for Machine Translation, Carnegie Mellon University. Meyer and Skuce's joint research has been supported by the Department of the Secretary of State of Canada.

14. For more details on the problems of validation techniques, see Skuce and Monarch 1990.
References


ISABELLE, Pierre, 1988. «Terminologie et traduction automatique» (report of an interview with Isabelle). Terminogramme No. 46. (Publication of the Office de la langue française, Gouvernement du Québec, Canada.)


SKUCE, Douglas, 1989. «A Generic Knowledge Acquisition Environment Integrating Natural
Language and Logic. *Proceedings IJCAI Workshop on Knowledge Acquisition* (Detroit, August 1989).


**WINANDS, Paul**, 1989. «Systèmes experts et terminologie». *META, 34, 3*.