Ingrid Meyer and Kristen Mackintosh
University of Ottawa, Canada

Phraseme Analysis and Concept Analysis: Exploring a Symbiotic Relationship in the Specialized Lexicon

Abstract

This paper analyzes ways in which phraseme analysis can facilitate concept analysis, and vice versa, in terminography work. We compare the phraseology of a number of conceptually related terms with conceptual information in our terminological knowledge base on optical storage technologies. We propose that a better understanding of phraseme–concept relations is important for both knowledge– and corpus–based approaches to terminography, approaches which we believe will merge in the next generation of term banks.

1. Introduction

This paper is concerned with phraseology as it pertains to terminography, by which we mean the identification, analysis and description of terms (=specialized lexical items). With the help of specialized texts and interviews with domain (=subject–field) experts, the terminographer carries out three basic tasks, which are not strictly sequential.

1) Identification of object of study. The terminographer circumscribes the domain and finds related domains; he identifies the principal concepts and terms, and eliminates lexical items belonging to general language.

2) Analysis. The terminographer analyzes the specialized corpus from both a conceptual and linguistic point of view. On the conceptual side, the concepts' principal attributes and relations (collectively, *characteristics*) are determined, a process which goes hand–in–hand with building up the conceptual structure of the domain, and mapping out links between these systems and those of related domains. On the linguistic side, various aspects of the terms are identified, such as collocational behaviour, grammatical features, and usage restrictions.

3) Synthesis. The terminographer's findings are typically presented in the form of a paper–based specialized dictionary or a term bank (=terminological database), which may be unilingual or multilingual. Definitions may be hand–crafted by the terminographer or taken from specialized texts. The terminographer may occasionally be required to propose a neologism when a concept has not yet been lexicalized. In working environments where standardization is essential, the terminographer may be required to propose a preferred term when competing candidates exist.

In this paper, we restrict ourselves to the second of these three tasks. Our goal is to identify some of the ways in which phrasemes are linked to the analysis component of terminography, and more specifically, to the *concept* analysis component. By concept analysis, we mean the process of discovering and representing the conceptual structures underlying the terms of a domain. Concept analysis is the foundation of terminography: without some understanding of the conceptual structure of a domain, it is impossible to carry out important linguistic tasks such as constructing definitions, dealing with quasi–synonyms, creating neologisms, etc. Explained simply, [1] concept analysis has three goals: 1) to establish concept systems within the domain, and links between these systems and those of related domains; 2) to develop conceptual frames (explained in 2.2 below) for the terms of the domain, by analyzing the attributes and relations of the concepts (this task goes hand–in–hand with the first); 3) to discriminate between closely related concepts. Multilingual work, which we do not deal with here, entails a fourth task of matching concepts between languages.

For our purposes, we will take phrasemes to include *noun compounds* (in the sense of Sager 1990:55–79) and *collocations* (in the sense of Benson *et al* 1986). We realize that these are different, in that a compound normally designates a single concept while a collocation does not. However, they share important relations to conceptual structure, and hence are grouped together here. We will examine phraseology from a practical angle, outlining ways in which phrasemes can help the terminographer with the conceptual side of analysis work, and conversely, ways in which the results of concept analysis can help the terminographer deal with phrasemes.

1.1 Motivation for this research

A better understanding of the relationship between concept and phraseme analysis has implications for the following two research areas:

1) A knowledge–based approach to terminography. Despite its importance, concept analysis remains largely unformalized, as is evident when one consults general textbooks on terminology. Here, it is common to find more space devoted to the *importance* of concept systems than to *methods for constructing them.* Concept analysis will likely never become an exact science; however, we believe that by exploiting the many regularities in 1) the way that phrasemes encode conceptual information, and 2) the way that conceptual structures generate phrasemes, we can at least develop better guidelines to incorporate into textbooks.

Increased formalization of concept analysis has implications beyond the didactic, however. The computerized terminological lexicon of the future needs to be rich not only in linguistic data, but also in domain knowledge. We have termed this model of the specialized lexicon a *terminological knowledge base* (TKB) (Meyer *et al* 1992a/b). Very simply, a TKB can be described as

a hybrid between a conventional term bank (containing all the strictly linguistic information one finds therein) and a *knowledge base*, as this concept is known in Artificial Intelligence. A TKB would function not only as a dictionary, but as a *general knowledge resource*, an invaluable asset to language professionals (writers, translators) and others dealing with specialized texts (students, information retrieval specialists, software engineers), as well as computer systems (machine translation, natural language processing).

2) A corpus–based approach to terminography. The increasing availability of specialized on–line corpora, and the parallel development of corpus analysis tools, offer exciting potential for facilitating one of the terminologist's most labour–intensive tasks: identifying the specialized lexical items — both single– and multi–word — for a given domain. This job used to be (and often still is) done by manually scanning texts. The new corpus analysis tools, in contrast, offer the possibility of extracting phrasemes (and their immediate contexts) automatically. As phrasemes are becoming easier to acquire, terminographers need to get clearer on *what to do with* the phrasemes thus extracted. A certain amount of research has recently targeted the problem of *how to classify* phrasemes;[2] however, despite a few notable exceptions,[3] very little work has addressed the question of *their relationship to the process of terminography.* This paper aims at helping to fill this gap.

1.2 Methodology

This paper is related to a broader terminography project for the domain of optical storage technologies. This particular investigation, however, is limited to two interrelated concept systems, *optical storage media* and *optical disc production processes*, the core concepts of which are illustrated in Figure 1. Consistent with the first step in a terminographer's methodology — identification of object of study — we first identified three major subcategories of optical disks (CD–ROMs, WORMs and erasable disks), and established that there was some kind of relationship between *mastering* and *CD–ROMs.* We also knew that the domain of optical storage technologies had close links to audio recording and paper–based publishing, which we term *ancestral domains.* This rough conceptual profile, however, needed to be filled out in several ways: the links to ancestral domains had to be clarified, our understanding of most concepts had to be sharpened, and in particular, the concept *mastering* — which we later learned was actually three separate concepts — had to be developed.

The goal of our investigation was to discover what roles phrasemes could play in fleshing out this conceptual skeleton, and conversely, what roles even a rough conceptual structure could play in facilitating phraseme analysis. To start with, we identified the phrasemes associated with five terms from this

concept system, using a one–million word corpus [4] and the concordancing tool *TACT*, developed at the University of Toronto. Our search terms were as follows (we included all spelling and morphological variants and senses): *optical disc, CD–ROM, erasable* (and its synonym *rewritable*), *WORM*, and *master*. The phrasemes found for these terms were compared with partially completed knowledge base entries for the associated concepts in our TKB. Essentially, for each phraseme identified, we asked ourselves the question: "Can this phraseme augment the existing TKB?", and conversely, "Can the conceptual information in the TKB facilitate our analysis of the phraseme?" Our observations for these two questions, respectively, are found in Sections 2 and 3 below.

2. Usefulness of phraseme analysis for concept analysis

As mentioned above, concept analysis has three goals: 1) to establish concept systems within the domain and links with related domains; 2) to develop conceptual frames for the lexical items of the domain; 3) to discriminate between closely related concepts.The usefulness of phraseology for each of these is discussed in turn below.

2.1 Establishing concept systems and links with related domains

Compounds constitute the most obvious link between phraseology and concept systems. In the words of Sager (1990:73), compounding serves the purpose of a "closer determination of a concept...while at the same time showing the relationship that exists between the new concept and its origin." For example, an *erasable* disc is a specialization of a *disc*, distinguished from other specializations by its erasability. The importance of compounds for indicating the general structure of a concept system is well known in the terminology literature and hence will not be discussed further here. Rather, we focus on a less studied aspect of conceptual structure which we term *multidimensionality* (Bowker and Meyer 1993).

Multidimensionality. We use this term to designate a phenomenon that occurs when a concept can be classified according to more than one characteristic, e.g. vehicles can be land/air/water or motorized/non–motorized, according to the characteristic 'place of transportation' and 'means of propulsion', respectively. Multidimensionality can be *top–down* as in the subclassification of vehicles just mentioned, or *bottom–up*, i.e., a concept may have several generic concepts, each belonging to a different dimension (e.g. a plane can be seen from the perspective of air vehicles or motorized vehicles).

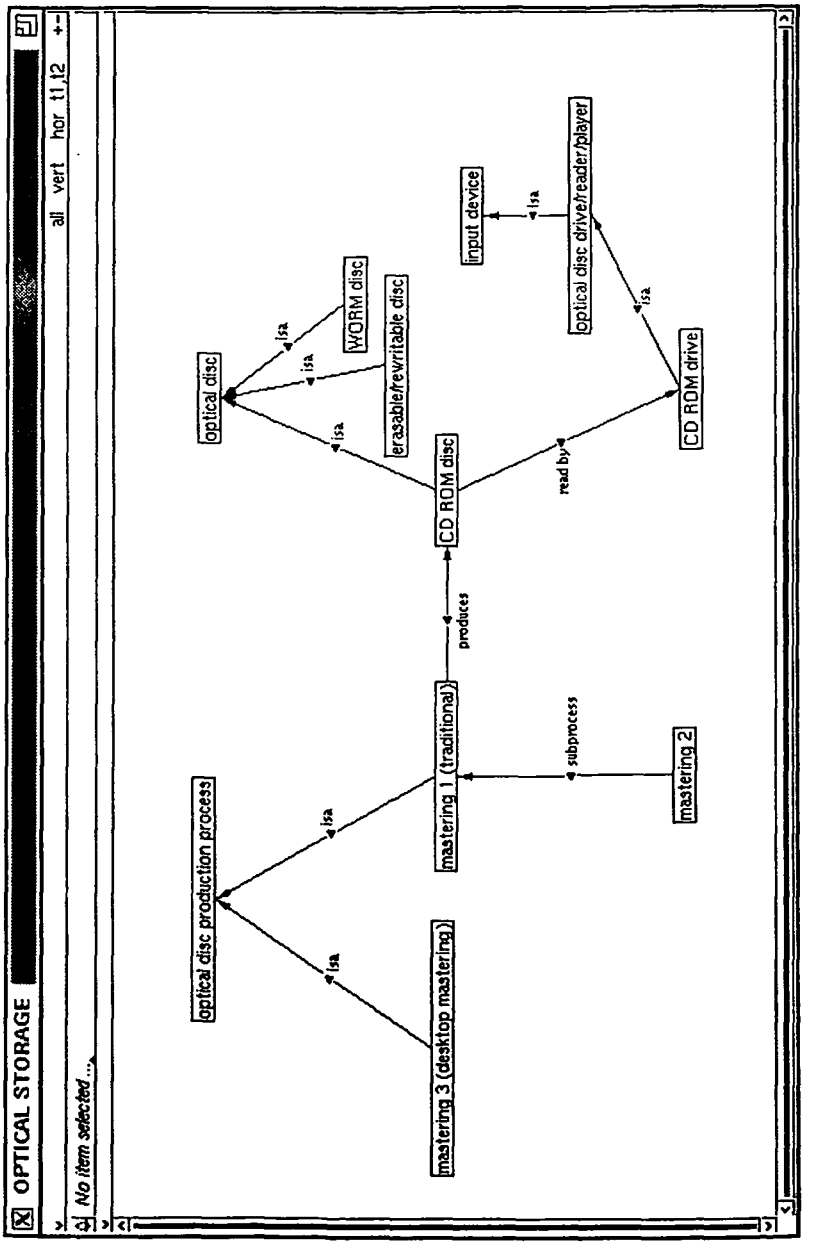

Figure 1
Two small concept systems in the *optical storage* domain

Top–down multidimensionality can frequently be extrapolated from noun phrases: for example, a KWIC (key–word–in–context) search on the term *disc* produced the phrasemes *CD–ROM/erasable/WORM disc* (characteristic = 'degree of writability'), *high–capacity/low–capacity disc* (characteristic = 'storage capacity'), etc. Evidence of bottom–up multidimensionality was found in both noun phrases and collocations. For example, *author a CD–ROM, subscribe to CD–ROM, CD–ROM publishing, CD–ROM library* illustrate a relationship with the ancestral domain of paper–based publishing, while *cut a CD–ROM, record a CD–ROM, CD–ROM player, CD–ROM jukebox* point to audio recording.

2.2 Establishing conceptual frames

By *conceptual frame*, analogous to *semantic frame* in Fillmore 1985, we mean the set of principal characteristics of a concept. Following a long tradition in epistemology, we include as characteristics both *attributes* (characteristics instrinsic to the concept itself, e.g. COLOUR, HEIGHT) and *relations* (characteristics involving a relation to another concept, e.g. GENERIC– SPECIFIC, PART–WHOLE, ACTOR–ACTION). Consistent with Martin 1992, we have found that both noun phrases and collocations provide valuable indications of conceptal frame elements. As Sager (1990:64) explains in some detail, noun compounds often (though not always) indicate the generic concept, as in *WORM optical disc*, a kind of optical disc. However, both compounds and collocations signal other important relations as well: *built–in CD–ROM drive*, for example, indicates that a CD–ROM drive can be seen as an optional part of something (the computer); *glass optical disc* indicates a relation between the disc and a CONSTITUENT SUBSTANCE. Attributes are also signalled by phrasemes, particularly by noun phrases: *built–in /internal/ integral disc* indicates the attribute LOCATION for disc; *portable/non– portable CD–ROM reader* indicates the attribute PORTABILITY for CD–ROM readers (and most likely for *readers* in general). An interesting research question is to what degree conceptual frames might be derived automatically. Obviously, there are many problematic cases. For example, the generic–specific relation can be obscured completely in single–word items like *jukebox* GENERIC = *optical disc drive*); some attributes cannot be analyzed without domain knowledge, as in *read–only drive*, where *read–only* modifies the disc *read* by the drive, not the drive itself (unlike, for example, *portable drive*)

2.3 Discriminating between related concepts

As in general language, different senses of a term tend to exhibit different phraseology. Since terminographers typically describe the lexicon of *only one domain at a time*, they are primarily concerned with *intra*–domain polysemy, rather than *inter*–domain (e.g. *docking station* in computing vs

space) While terminological polysemy is much rarer than in the general lexicon, it can be extremely problematic when it does occur. For example, terms often exhibit the insidious type of polysemy in which one of the polysemes stands in a generic–specific relation to another. For example, it is accurate to say that *a PC* (in the sense of 'the original IBM PC') *is a kind of PC* (in the sense of 'all IBMs and compatibles') *is a kind* of PC (in the sense of 'personal computer'). In our domain, we discovered that *master* (the verb) actually has three different senses, indicated in Figure 1: *master 1* designates the entire process of converting digital data into mass-produced CD-ROMS in the traditional way (i.e. in a *mastering facility*); *master 2* designates one subprocess of *master 1*; *master 3* designates an emergent coordinate concept for *master 1* (sharing the genus *optical disc production process*). Fortunately, these three senses generate somewhat different phrasemes: for example, *master 1* generates *mastering facility, metal master, trial master, mastering machine*, while *master 2* generates *desktop mastering, do-it-yourself mastering, in-house mastering*, etc.

3. Usefulness of concept analysis for phraseme analysis

An understanding of the conceptual structure of a domain helps in *understanding* phrasemes and *predicting* them, as discussed below.

3.1 Understanding phraseology: ambiguity and synonymy

The crucial importance of some domain knowledge for disambiguating complex noun phrases is well known. In our study, for example, we encountered the example *multi-function mini optical disc jukebox*, where it is unclear what is *mini* and *multi-function*, the disc or the drive. Resolving the problem means knowing that 1) a jukebox *is a kind* of disc drive, 2) drives can be multi-function or single-function, and 3) size is a key attribute of discs, not of drives (though drives reading smaller disks are consequently smaller themselves). Understanding conceptual multidimensionality can also play an important role in meaning discrimination. For example, understanding that optical storage has relations both to computer hardware/software and to audio recording helps one determine that *CD-ROM drive/reader* is synonymous with *CD-ROM player* (the first inheriting from computing, the second from audio). The same goes for *to write (data) to optical disc* and *to record (data) on optical disc.*

3.2 Predicting phraseology: inheritance

Since terminography must keep pace with the forefront of developments in specialized domains, it comprises a significant "predictive" element: on the one hand, when the concept has not yet been lexicalized, the terminographer must predict the most logical term for an emergent concept

— in other words, create a neologism; on the other hand, when several competing terms for a new concept exist, the terminographer may have to predict which will be most naturally accepted by users (in cases where standardization is required).

Phraseology may be predicted through two aspects of conceptual structure: 1) the *conceptual frame*, and 2) *inheritance* within a generic–specific hierarchy. The first has been aptly described by Martin 1992, and taken up also by Heid 1992/1993. Martin's examples include *system*, whose conceptual characteristic FUNCTION generates the phraseme *nervous system*; *vowel*, whose characteristic ARTICULATION generates *tense vowel*; *dictionary*, whose characteristic TYPE generates *etymological dictionary*; etc.

The second aspect of conceptual structure which imposes regularities on phraseology is inheritance. Most domains in science and technology can, to some degree at least, be organized in generic–specific hierarchies where characteristics inherit from general to more specific concepts. Just as conceptual characteristics inherit from generic to specific concepts (e.g. *CD–ROM disc* would inherit the characteristics of *disc* and add a few more), so too may phrasemes. For example, the seemingly deviant syntax in the collocation *master (data)* ***to*** *CD–ROM* (instead of, for example, ***on*** CD–ROM) makes sense when one considers that *mastering* is a specialization of *writing (data) to sth* in computing.

As a more complex example, consider the phraseme *CD–ROM player reads discs*. Here, a form of multiple inheritance (simultaneous inheritance from several generic concepts) appears to apply: *read* can be explained by the fact that a *CD–ROM player* is a kind of *disc drive*, which in turn is a kind of *input device*, and that all input devices are said to *read* data; *player* can be explained by the fact that optical disc technologies inherit from audio recording. Multiple inheritance is also a factor in the collocation *publish on CD–ROM*. While *publish* has obviously inherited from the domain of publishing, it does not take the preposition ***in*** that one might expect (as in to *publish sth* ***in*** *a journal*). Rather, it appears the preposition has inherited from the domain of audio recording, where one speaks of *recording* ***on*** disc. All these examples of multiple inheritance illustrate that while inheritance is a powerful factor in phraseme formation, it can be extremely complex, interfering with easy predictability of phrasemes.

4. Discussion

We have tried to demonstrate that phraseme analysis facilitates the concept analysis tasks of 1) establishing general conceptual structures within a domain and mapping links to related domains, 2) establishing conceptual frame elements, and 3) discriminating between related concepts; on the other hand, we have tried to demonstrate that a certain understanding of

conceptual structure facilitates phraseme analysis by 1) clarifying ambiguity and synonymy, and 2) predicting phraseology.

We have also proposed that a better understanding of the symbiotic relationship between concept and phraseme analysis will promote the development of both knowledge–based and corpus–based approaches to terminography. If, as we believe, term banks become *richer in knowledge* and *closer to text*, an interesting question is how these two approaches can be integrated in a terminographer's workstation environment. For example, one could imagine a terminographer wanting to guide his concept analysis through corpus queries such as "show me all the phraseology for this term that suggests what the subconcepts might be". Conversely, the terminographer might want to guide his corpus search conceptually, with queries such as "show me all the verbal collocates that precede *disc* or any specialization of disc found in the knowledge base."

In the short term, we hope that a clearer "map" of phraseme–concept relations will at least be useful to anyone trying to teach the difficult skill of concept analysis in terminography.

Acknowledgements

This research has been made possible by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC) and a variety of support from the University of Ottawa. Mackintosh's graduate studies have been supported by the Ontario Graduate Scholarship (OGS) programme.

Notes

1 For a more detailed analysis, Cf. Meyer 1993.
2 For example, Béjoint and Thoirion 1992, Heid 1993, Martin 1992.
3 In particular, Blampain 1993, Heid 1992, Humbley 1993, and Martin 1992, which have inspired much of our thinking for this paper.
4 Our corpus was extracted through key–word searches from a commercially available CD–ROM called *Computer Select* (Davis Publishing, Computer Library, NY), which contains English–language articles from several hundred journals on all areas of computing.

References

Béjoint, Henri et Thoiron, Philippe. 1992. "Macrostructure et microstructure dans un dictionnaire de collocations en langue de spécialité". *In Terminologie et traduction 2/3 – 1992.* Commission des Communautés européennes, Luxembourg.

Benson, Morton, Benson, Evelyn, and Ilson, Robert. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Amsterdam/Philadelphia, John Benjamins.

Blampain, Daniel. 1993. "Notions et phraséologie : Une nouvelle alliance?" *Terminologies nouvelles*, 10, pp. 43–49

Bowker, Lynne and Meyer, Ingrid. 1993. "Beyond "Textbook" Concept Systems: Handling Multidimensionality in a New Generation of Term Banks". *Proceedings of the 3rd International Congress on Terminology and Knowledge Engineering* (Cologne, Germany, August 1993). Frankfurt, INDEKS Verlag.

Fillmore, Charles J. 1985. "Frames and the semantics of understanding". *Quaderni di Semantica*, Vol. 6, No. 2, pp. 222–254.

Heid, Ulrich 1992. "Décrire les collocations – deux approaches lexicographiques et leur applications dans un outil informatisé". In *Terminologie et traduction 2/3 – 1992.* Commission des Communautés européenes, Luxembourg.

Heid, Ulrich. 1993. "On the representation of collocational phenomena in sublanguage lexicons." *Proceedings of the 3rd International Congress on Terminology and Knowledge Engineering (TKE 93).*

Humbley, John. 1993 "Exploitation d'un vocabulaire combinatoire : syntaxe, phraséologie, analyse conceptuelle". *Terminologies nouvelles, 10*, pp. 95–102.

Martin, Willy. 1992. "Remarks on Collocations in Sublanguages". *Terminologie et traduction 2/3 – 1992.* Commission des Communautés européenes, Luxembourg.

Meyer, Ingrid. 1993 . "Concept Management for Terminology: a Knowledge Engineering Approach". *Standardizing Terminology for Better Communication: Practice, Applied Theory and Results.* (Special Technical Publication of the ASTM, No. 1166). Eds. R. A. Strehlow and S. E. Wright. Philadelphia, American Society for Testing and Materials, pp. 140–151.

Meyer, Ingrid, Bowker, Lynne and ECK, Karen. 1992a. "COGNITERM: An Experiment in Building a Terminological Knowledge Base". *Proceedings of the 5th Euralex International Congress*, pp. 159–172.

Meyer, Ingrid, Skuce, Douglas, Bowker, Lynne and ECK, Karen. 1992b. "Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base". *Proceedings of the 14th International Conference on Computational Linguistics* (COLING 92), pp 956–960.

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