Building a concept hierarchy from corpus analysis*

Caroline Barrière

Corpus analysis is today at the heart of building Terminological Knowledge Bases (TKBs). Important terms are usually first extracted from a corpus and then related to one another via semantic relations. This research brings the discovery of semantic relations to the forefront to allow the discovery of less stable lexical units or *unlabeled concepts,* which are important to include in a TKB to facilitate knowledge organization. We suggest a concept hierarchy made of concept nodes defined via a representational structure emphasizing both labeling and conceptual representation. The Conceptual Graph formalism chosen for conceptual representation allows a compositional view of concepts, which is relevant for their comparison and their organization in a concept lattice. Examples manually extracted from a scuba-diving corpus are presented to explore the possibilities of this approach. Subsequently, steps toward a semi-automatic construction of a concept hierarchy from corpus analysis are presented to evaluate their underlying hypothesis and feasibility.

Keywords: unlabeled concepts, concept hierarchy, hyperonymy, semantic relations, corpus analysis, knowledge extraction, domain modeling, computational terminology

1. Introduction

Terminological Knowledge Bases (TKBs) aim to capture the essence of a domain via its important concepts and relations. As much as Wordnet (Fellbaum 1998) provides a semantic network of concepts for the general language, TKBs must be built for specialized domains. Six possible uses of TKBs can be identified:

Terminology 10:2 (2004), **241–263.**
ISSN 0929–9971 / E-ISSN 1569–9994

1. defining the terms used in a subject field to build a specialized dictionary;
2. defining language curriculum (Pearson 1998: 19) for the L2 learner in a specific field;
3. providing a better understanding of a domain for training;
4. providing a schematic view of important information in text for domain experts;
5. developing standards for naming concepts in the domain;
6. developing a view of a domain to assist in translation.

All these goals rely on a better understanding of the organization of domain knowledge. As this knowledge is present in a domain corpus, a knowledge extraction tool should transform textual information into an explicit, storable, reusable and retrievable encoding of the information expressed explicitly or implicitly in text. This research contributes to the fields of knowledge extraction and TKB construction by presenting (1) a "relation-first" view on corpus analysis in complement to "term-first", (2) a uniform representation formalism allowing a hierarchical representation of both stable terms and unlabeled concepts found in text, and (3) a method for extracting such information. Although the implementation of a fully automatic extraction tool has not yet been achieved, our goal hereafter is to establish the foundations for a representation formalism, to explore the different processes required and their inherent difficulties, as well as to show the anticipated results through manually produced examples.

Our main proposal is to complement a "term first" view of TKB construction and explore the corpus as a source of concepts sometimes explicit (having a corresponding term), and sometimes implicit (what we call *unlabeled concept*). This exploration is primarily based on the discovery of semantic relations.

Section 2 examines the two main steps of constructing Terminological Knowledge Bases (TKBs): term extraction and semantic relation extraction. Emphasis is on the idea of semantic relation extraction as an independent process which will naturally lead to the discovery of unlabeled concept nodes.

Section 3, at the heart of this research, presents in more details the proposed formalism for representing concept nodes in the concept hierarchy. This representation structure focuses on four fields: labeling (or denomination), hyperonym (or superordinate), definitional characteristics, and examplification (or subclasses). This section further expands on the need for a compositional expression of the label of a concept node and consequently for the need of a representation formalism which allows compositionality. A flexible representation formalism, Conceptual Graphs (CGs), is presented and discussed for such a purpose.

More examples from the scuba-diving corpus are presented in Section 4

showing how the concept nodes can be naturally organized in a hierarchy. This section also introduces the possibility of combining information from lexical sources to information found in the corpus to better organized the concept nodes into a concept hierarchy.

Section 5 investigates the hypothesis and difficulties related to all steps of the process of semi-automatic construction of the concept hierarchy from text.

A final section summarizes the contributions made by this research and looks at future work.

2. Relation-first view of corpus analysis

Terms, stable and conceptually significant phrases of a domain expressed in text, can be found through corpus analysis. By themselves, they provide an overview of the domain described in the corpus. Many systems perform term extraction automatically. We refer the reader to Cabré Castellvi et al. (2001) for an excellent review of systems using different algorithms and heuristics (probabilistic, linguistic, pattern-based) with their strengths and weaknesses. Some recent approaches even go beyond looking inside a single corpus for terms, but use a comparison approach based on different domain corpora (Chung 2003). The purpose of this research is not to favor one type of approach over another, but simply to emphasize that term extraction is a very common process and often considered a natural first step to any TKB construction, and simply to illustrate the possible results of such a proces. To familiarize the reader with some common terms in the scuba-diving domain, which we will be using throughout this paper, we present in Table 1 and 2 some results of a term extraction process, using pattern matching alone, employing a stoplist of tokens (Barrière and Copeck 2001) to act as delimiters between terms.

Tables 1 and 2 show single word and multiword terms with their number of occurrences in the text. The corpus on scuba-diving is a one megabyte corpus composed mainly of informative texts (non-narrative and fact-based as defined in Barrière 2001) found on the Internet on the topic of scuba diving. The texts are quite varied, some being technical (scuba equipment), others more medical (risks and related health problems), and still others giving important advice about diving conditions.

Once the terms have been extracted by means of a semi-automatic system, they can be verified by a domain expert and stored in the TKB. To be kept, terms usually need a certain stability of form and meaning and a definite

Table 1. Single word terms extracted from the scuba-diving corpus.

Term	Frequency	Term	Frequency	Term	Frequency
dive	205	oxygen	105	cave	58
water	194	boat	84	day	54
divers	189	depth	80	training	53
diving	159	regulator	75	tank	50
time	149	dives	73	Click	49
air	145	minutes	69	symptoms	48
feet	137	gear	62	cavern	47
diver	128	equipment	61	buddy	47
surface	126	stage	59	people	46
mask	116	pressure	58	pain	46
Top	106	nitrogen	58	gas	45

Table 2. Multiword terms extracted from the scuba-diving corpus.

Term	Frequency	Term	Frequency
cave diving	59	cave divers equipment	29
nitrogen narcosis	45	breathing gas	21
Subscription retailers	36	oxygen toxicity	20
RDS Online	36	underwater photography	19
Rights Reserved Mail	36	1994–1999 Rodale Press	19
breathing loop	35	cave diver training	18
technical diving	31	world travel search rds online	17
decompression sickness	31	underwater u.s.a	17

conceptual role within a domain or even more precisely within an application. The organization of the TKB then resides in the interaction between the terms chosen, that is in the semantic relations which link them. These semantic relations can also be found in text. Many researchers have studied which types of patterns indicate semantic relations and the process of discovering them in real text (Bowden et al. 1996; Cartier 1997; Condamines and Rebeyrolle 1998; Davidson et al. 1998; Meyer et al. 1999). Early work by Hearst (1992) marked the first step in corpus analysis research in this direction and investigated the hyperonym relation. Although recent work has also focused on meronymy (part-of), function, and more recently, causality (Garcia 1997; Barrière 2001) the hyperonym relation keeps its special status of being at the core of the hierarchical organization of a domain allowing not only categorization but also property inheritance.

This research suggests not to limit the search for semantic relations to sentences containing one or more of terms (found in a previous term extraction

step), but instead, to look for semantic relations anywhere in the corpus to allow the discovery of concepts not expressed by terms but which can be very useful for organizing the information of a domain. To illustrate this idea, we use SeRT ('Semantic Relation in Text' — an interactive software for TKB construction, see Barrière and Copeck 2001) to identify a few hyperonym patterns and show, in Figure 1, some sentences of the scuba-diving corpus with these surface patterns emphasized in bold. Interestingly, as we attempt constructing a taxonomy using the information in text surrounding these patterns, we see that they naturally suggest the existence of unlabeled concepts as shown in Figure 2.

1. **Other types of** overhead environment diving **include** wreck penetration, ice diving and **some types of** commercial diving, **such as** pipeline inspection.
2. Consider cave diving **a form of** technical diving.
3. How do cavern and cave diving **differ from other types of** overhead environment diving?
4. Everywhere however ice diving, wreck penetration or cave dives **as well as** diving on something other than air **are considered** technical dives.
5. Cavern diving **is a form of** recreational diving.

Figure 1. Examples of hyperonymy surface patterns.

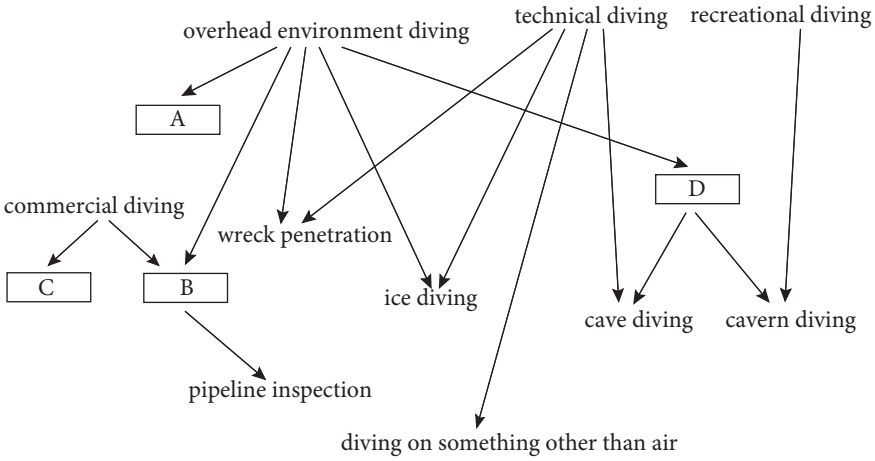

Figure 2. Partial tangled hierarchy from sentences of Figure 1.

We further look at each sentence from Figure 1 to illustrate the generation of the unlabeled concepts.

Sentence 1: **Other types of** (*overhead environment diving*) means that at least one other type exists (label A in Figure), perhaps previously mentioned in the text, apart from the ones in the enumeration announced by **include** (*wreck penetration, ice diving*). **Some types of** (*commercial diving*) has a dividing sense, indicating that some types (label B) will be *overhead environment* diving while others will not (label C). Within the ones that are, then **such as** introduces *pipeline inspection* as an example.
Sentence 2: The pattern **a form of** is an indicator of a subclass/superclass relation between *cave diving* and *technical diving*.
Sentence 3: The pattern **differ from other types of** indicates that the behavior of the first group (*cave and cavern diving*) is particular and therefore represents a different category (label D).
Sentence 4: The enumeration (*ice diving, wreck penetration, cave diving*) followed by **as well as** is an indicator of a group of elements sharing a common superclass. The surface pattern **are considered** actually names (*technical diving*) the common superclass.
Sentence 5: The pattern **a form of** is again an indicator of subclass/superclass, here between *cavern diving* and *recreational diving*.

Surface patterns allow structuring of terms into a tangled hierarchical organization, and also naturally allow the inclusion of unlabeled nodes. These nodes represent concepts in the sense of sharing properties, or representing abstract classes of objects. We are not alone in viewing text as rich in information about concepts, as we can see in Cruse (1986):

> In some cases the linguistic motivation for grouping a set of lexical items together as taxonyms of a non-existent superordinate may not be so immediately apparent. Consider the class of movable items one buys when moving into a new house: furniture (chairs, tables, beds, etc.), appliances (refrigerator, television, washing-machine, etc), carpets, curtains, etc. Once again there is no label for this overall category. Nor is there a simple diagnostic frame. But the use of everything in *Of course, we had to buy everything when we bought our first house …. beds, carpets, cooker …* is suggestive (it means "everything in a category we both know about but I can't name." (Cruse 1986: 148)

Cruse's 'linguistic motivation' can be discovered by surface patterns in text. Whether called covert categories or unlabeled concepts, these concepts will be a valuable addition to a TKB. We further look at developing an appropriate representation for them.

3. Formalism for concept representation and organization

As we identify nodes A to D in Figure 2, we position them within the hierarchy along with the terms, but so far there is no common formalism to express all the nodes. To define our formalism, we take inspiration from Wüster and Trimble, and work under the hypothesis that a TKB should be an organized collection of informative nodes, each one defining a concept of particular importance for the understanding of a domain.

Text as a source of information, and therefore a source of definitions, can be analyzed to find complete defining expositives that would satisfy Trimble's (1985) formula for formal definition of a term. A complete defining expositive requires all elements be present: the definiendum (the word defined or label), its superordinate (also referred to as hyperonym, genus or superclass), and at least one differentiating characteristic often introduced by a preposition or contained in a relative clause. Let us suggest the following condensed representation [D (definiendum), S (superordinate), C (characteristics)]. For example, the sentence: "*A squeeze, defined as ear pain on descent, is the most common medical complaint from divers.*" contains all elements leading to the following representation: [D = *squeeze*, S = *ear pain*, C = *on descent (time)*]. While a text rarely provides complete defining expositives, it often offers partial defining expositives corresponding to what Trimble calls *semi-formal definitions*. A semi-formal definition explains a term by listing differences between it and other members of its class, but no superordinate is given. Semi-formal expositives might use patterns like *is used to* or *has* to identify function or parts, thereby giving some information about the term. For example, the sentence: "*For deep diving, nitrox is used to accelerate decompression times.*" leads to a structure [D = *nitrox*, S = ??, C = *accelerate decompression times (function)*]. A third possibility, not mentioned by Trimble, is the case when a term is given a superordinate but no differentiating characteristics. For example, the sentence: "*Overhead environment diving includes wreck penetration.*" leads to this structure: [D = *wreck penetration*, S = *overhead environment diving*, C = ??]. If *wreck penetration* is partially defined by its superclass *overhead environment diving*, we can reciprocally say that *overhead environment diving* is partially defined by its subclass *wreck penetration*.

This corresponds to the notion of 'extensional definitions' from Wüster (2004): "But frequently extensional definitions are clearer and easier to understand….. An extensional definition consists of the enumeration of subordinated concepts." This suggests that a fourth component E (examples, subclasses)

should be added to our structure to make it complete. Now, our concept characterization [D, S, C, E] needs refinement. First, let us say that each element D, S, C, E can be a set, as there could be multiple labels, multiple superclasses, multiple defining characteristics and multiple examples or subclasses. Second, each field of the node can be assigned with more or less precision. Let us explore hereafter some ranges for their possible values. Although the ranges are discretized below into a few possibilities, these possibilities should only be seen as points on a continuum.

Definiendum/Label (D):
1. Naming expression (dictionary entry — label frequent and fixed enough, giving a conventional relation to meaning, not compositional);
2. Descriptive expression (given by a combination hyperonym + definition (S+C);
3. Unlabeled (either S or C, but not both are present in structure).

Superordinate (S):
1. Specific hyperonym;
2. Overgeneral hyperonym (general term found outside of the domain — generality is also a vague notion and can probably best be expressed by informativess such as defined in Resnik (1995) which is based on frequency of occurrence in text).

Defining Characteristics (C):
1. Defined (a few definitional elements provided);
2. Partially defined (minimally one attribute or function is used to restrict the extent of the concept — in the sense of semi-formal definitive, again from Trimble 1985);
3. Undefined.

Examples/Subclasses (E):
1. Exemplified (some subclasses are given in the text);
2. Not exemplified (text does not mention any subclasses).

Having established our node representation, we return to corpus analysis to discover these nodes from textual information. In Section 2, it was by looking at linguistic patterns of the hyperonym relation that we discovered places in the hierarchy for unlabeled nodes. The hyperonym relation is of importance for the placement of a node in the concept lattice as it dictates the S (superclass) and E (examples) field. In this section, we emphasized the search for more relations in text to expand the definition of each word. This is of importance for the C (defining characteristics) field. The remaining field L (label) will be either

occupied by a stable term (which could have been identified in the term extraction process), can be left "unlabeled", or can be automatically generated from the Conceptual Representation, as we will discuss later on.

Table 3 shows examples of sentences found via identification of semantic relation patterns in the corpus on scuba-diving. We organized Table 3 in a continuum in the range of labeling stability. We also highlight the hyperonym patterns in bold and the other defining relations patterns by underlining them. We also identify each semantic relation within the Defining Characteristics column.

Examples 1 and 2, *squeeze* and *nitrox*, are pure naming expression, opaque and only understood by convention. Example 3, *overhead environment diving*, is at the limit of naming and describing expression, since it is a fairly stable (it occurs 5 times in the small corpus) but contains some level of compositionality. For examples 4 and 5, we made up describing expressions, but with quite specific hyperonyms (*diluant gas* and *diving*). Example 6 is slightly less specific (*irritation*); so is example 7 (*effect*) and example 8 (*problem*).

Only examples 1 to 3 correspond to stable terms as found via a process of term extraction. But, in Table 3, we also find much interesting information about domain related concepts through the extraction of relation patterns. For the concepts not found in the list of terms, a label can always be generated, as is done manually and shown in italics. The automatic generation of such a label is more problematic. We believe this generation should come from the conceptual representation of the node, and therefore suggest to focus on such representation, investigating Conceptual Graphs (Sowa 1984) to do so.

3.1 Conceptual Graph representation

Conceptual Graphs (CGs) are well established in the Knowledge Representation (KR) community as a knowledge formalism equivalent to a first order logical form, including properties of inheritance and operators which could be linked to the four concept formation operators of Wüster (2003): disjunction, conjunction, determination and integration. All these operators are based on the principle of compositionality.

To allow a consistent organization of concepts in the hierarchy, as much should be known about them as possible, and where decomposition is a possibility, it should be taken. Single words are opaque but multiword expressions such as *cave diving, cavern diving, wreck penetration, pipeline inspection* and *ice diving* have a compositional meaning.

Table 3. Concept labeling, superordinate, defining characteristics and examples.

	Corpus sentence	Label	Super-ordinate	Defining Characteristics	Examples
1	A squeeze, **defined as** ear pain on descent, is the most common medical complaint from divers.	squeeze	ear pain	descent (time)	
2	For deep diving, nitrox is used to accelerate decompression times.	nitrox		accelerate decompression times (function) deep diving (where)	
3	overhead environment diving **includes** wreck penetration and ice diving	overhead environment diving			wreck penetration, ice diving
4	The diluent gas for the oxygen in the breathing mixture **is not limited to** nitrogen and may **include** helium and other gases.	*breathing mixture's oxygen diluent gas*	diluent gas		nitrogen helium
5	diving on something other than air **is considered** a technical dive	*non-air diving*	technical dive		
6	**any type of** irritation which inflames the mucus membrane and causes swelling and mucus discharge	*type of irritation*	irritation	inflames the mucus membrane (cause) swelling and mucus discharge (cause)	
7	The immediate effects of a sea urchin sting **are usually** a burning sensation, followed by swelling, redness and an aching pain.	*effects of a sea urchin stings*	effect		burning sensation, swelling, redness, aching pain
8	problem **like** a recent cold, allergies	*some type of problem*	problem		recent cold, allergies

A few researchers (Vanderwende 1984; Jacquemin and Royauté 1994; Barker and Szpakowicz 1998) have explored relations within multiword expressions. Let us illustrate this idea with four examples from the scuba corpus. *Cave diving* is diving that takes place in a cave, *technical diving* is a type of diving, *pipeline inspection* refers to a pipeline as the object of the inspection, and *breathing gas* refers to a gas that has the function or purpose to help breathing. This construction is important since compositionality is essential to resolving issues of placement of concepts in the hierarchy. However, in a case like *pipeline inspection* it is not sufficient since no component refers to diving. Conceptual graphs (CGs) (Sowa 1984) will allow for the expression of compositionality, and the noun compounds given above can be expressed as:

cave diving:	[diving] →	(location) →	[cave]
technical diving:	[diving] →	(type-of) →	[technical]
pipeline inspection:	[inspection] →	(object) →	[pipeline]
breathing gas:	[gas] →	(purpose) →	[breathing]

Any CG is minimally a concept (e.g. [diving]), and that concept can be modified by different relations to other concepts. CGs therefore allow for much complexity if a concept is in relation to many other concepts which are themselves in relation to other concepts and so on. The total set of relations used in CGs differs from one researcher to the other, but usually a basic set is agreed on which would contain case roles relations (agent, object, patient, beneficiary), paradigmatic relations (kind-of, synonym, antonym, meronym, is-a), locative and time relations (location, point-in-time, frequency), complement relations (manner, purpose) and attributive relations (attribute, color, size).

In Figure 3, conceptual graphs are used to show the different concepts included in the hierarchical relation of Figure 2 (presented in Section 2). The explicitness in the conceptual representation allows for finding commonality between nodes. CGs being graphs, graph operations are defined on them, such as Maximal Common Subgraph (MCS — finding the common part of two graphs, this would correspond to the disjunction from Wüster), or External Join (putting two graphs together, this would correspond to the integration and the conjunction from Wüster). With the CG representation of *commercial diving, technical diving, overhead environment diving* and *recreational diving,* using the MCS operator, we obtain a common superclass [*diving*] (shown in Figure 3, but which was not present in Figure 2). Such implicit groupings are much easier to find via the conceptual representation of terms than via their labeling part.

Not only are CGs well suited for representing the concepts in the lattice, they are also quite flexible in their manipulation and allow for generalization.

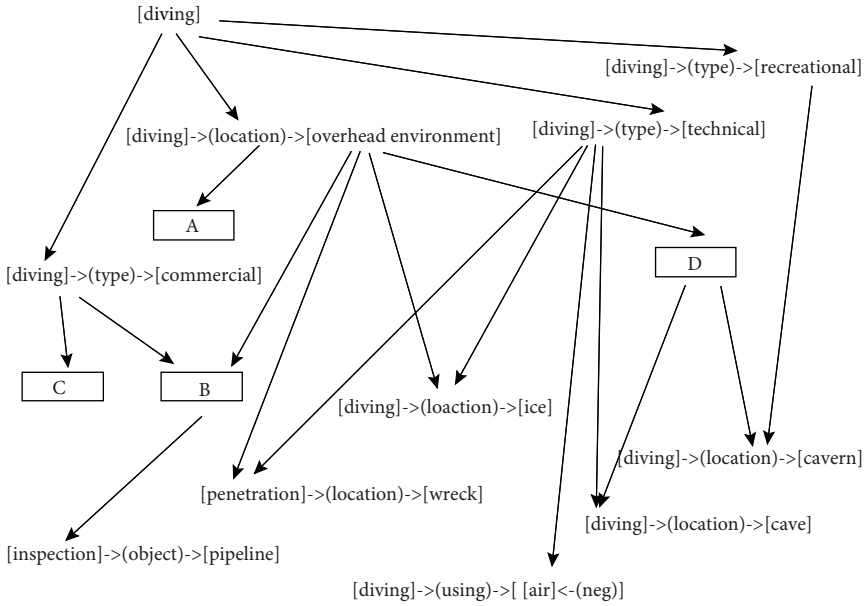

Figure 3. Concept network representation of the term network in Figure 2.

4. Examples from a Scuba-diving corpus

Having stated our formalism for node definition and node representation in the previous section, we wish to further show through examples found in the scuba-diving corpus, the value of such representation and the richness of its expression.

A concept node was defined as [D = d, S = s, C = c, E = e]. Focusing on the concept lattice, S and E are implicitly expressed by the parents and children links, and therefore, each node can be reduced to [D = d, C = c].

In each Figure (4 to 7), each concept node is given a label (D) and a Conceptual Graph representation (C). Within the CG representation, a unique identifier/reference is given for the concept expressed by the CG (*X1, *X2, *Y1, etc.). This is essential for storage and also to allow further reference and refinement of a CG in its children's nodes or other nodes. As for the label part,

either a term is used, a single or multi-word component of a term (extracted from decomposition), or a made-up label (as could be generated from a CG).

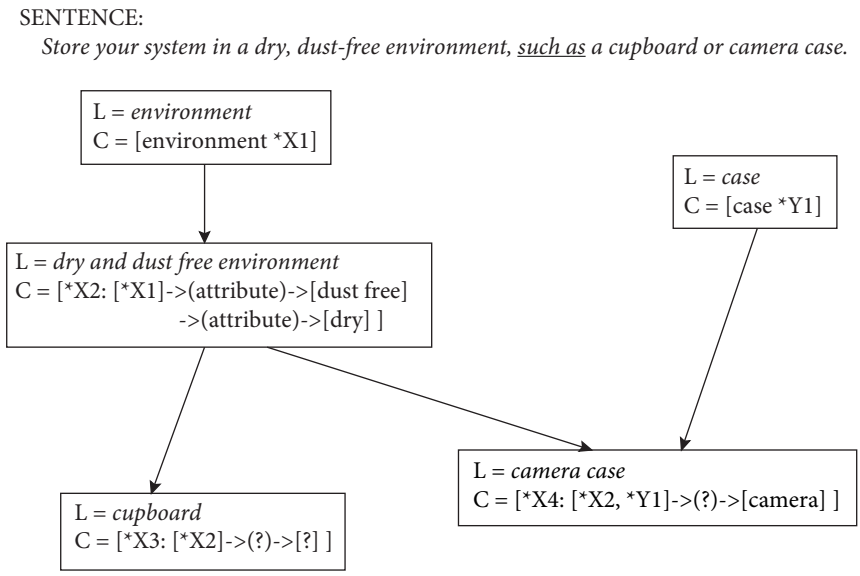

Figure 4. Issues of compositionality.

Looking first at Figure 4, we see at the top of the tree a label *environment* referring to the concept [environment] identified as *X1. Then some defining characteristics *dry and dust free environment* are added to *X1 to build the composite meaning *X2. *X2 is examplified by *cupboard* and *camera case* providing an unlabeled node to allow this group. The first example *cupboard*[1] has a label but no extra information. Therefore the CG representation of *X3 is based on *X2, but we indicate via interrogation marks that no more is known, although there must be more. The other example *camera case*, is providing some additional information by the fact that it can be decomposed, though the nature of the relation between *case* and *camera* is unknown as shown in the CG *X4. This is of course assuming that the corpus evidence is our only evidence. If we had access via a dictionary to its conceptual representation, we might find that *case* relates to a notion of [storage]. A dictionary definition expressing *case* as [case *Y1] ← (location) ← [store] → (object) → [??], that is a location to store something, would allow *camera case* to be compositionally understood as: [case *Y1] ← (location) ← [store] → (object) → [camera].

Figure 4 illustrates the "terminological orientation" of the ideas presented in this paper, as we see how a concept node takes on its importance only within the organization of concepts in a specific domain. The two phrases *cupboard* and *camera case* certainly do not share the same genus in a general dictionary. The Merriam Webster Online Dictionary gives *closet* as superclass of *cupboard,* and *camera case* has no entry, but *case* has *receptacle* as superclass. Their grouping makes sense here in the context of video equipment storage.

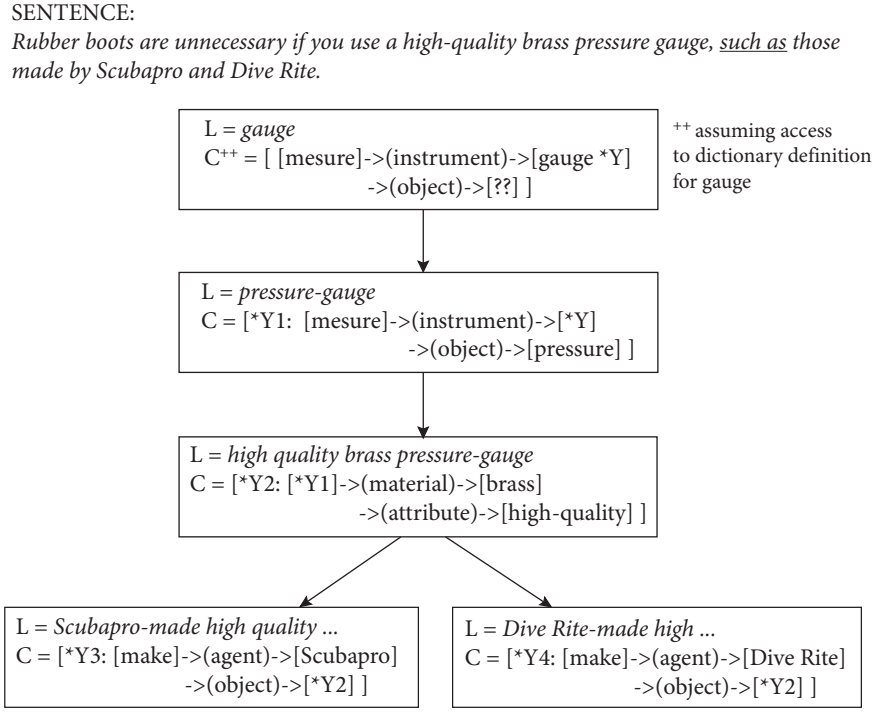

Figure 5. Corpus and dictionary interrelation.

Figure 5 shows a similar example with *pressure gauge*, and further illustrates the possible combination between information from dictionary and corpus, by taking one of the Merriam Webster Online Dictionary's definitions of *gauge* (an instrument for or a means of measuring or testing) to add a conceptual representation for *gauge* in the lattice. Figure 5 also illustrates that labelling could be generated from the CG representation as in *Scubapro-made high quality brass pressure gauge.*

Figure 6 shows three small trees rooted at *problem*, *barotrauma* and *dysfunction*

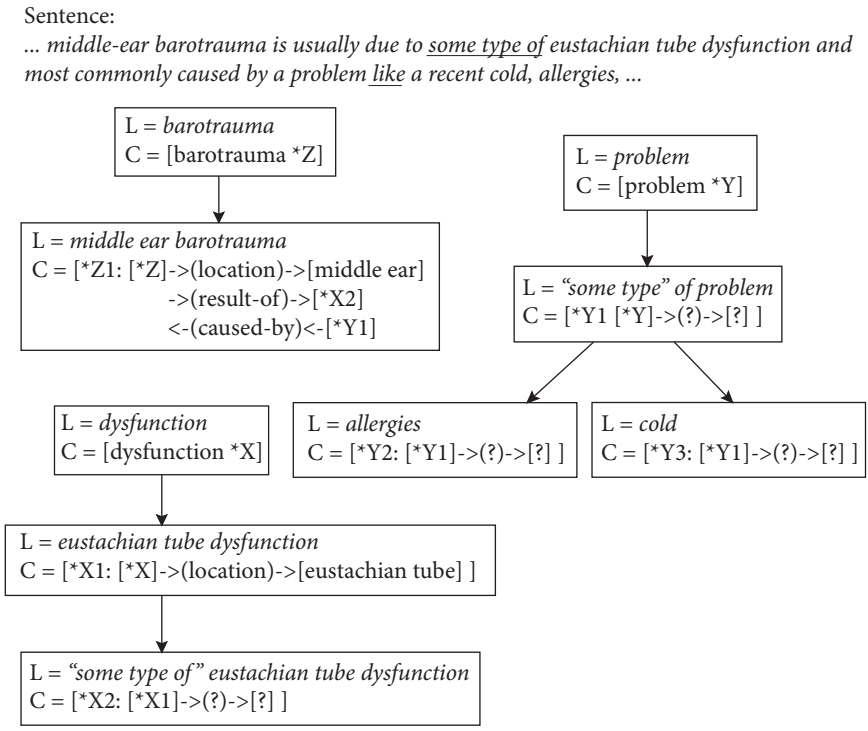

Figure 6. Unlabeled concepts.

which represent different parts of the concept lattice and how they are intertwined through their conceptual representations. The sentence introduces causality which is a very important relation in many domains (medicine, pharmaceutical, any instruction or repair manuals). While causality is rarely represented in TKBs, we believe it does have roots in terminology (Barrière and Hermet 2002) and it is essential toward building a domain model that will allow inferencing and reasoning. Nodes *Y1 and *X2 in Figure 6 require further explanation. These are two cases of structures in which both the label and the definitional part are missing, but the existence of the node can be extracted from sentence analysis. In such case, even if we can build a very vague descriptive expression, using *some type of*, we consider these nodes as unlabeled concepts.

Figure 7 shows two examples of groups of objects defined through their case relation to a verb. Previous studies (Barrière and Popowich 1996, 2000) have shown that many unlabeled concepts exist in relation to an action. The concept in question either facilitates the action (instrument), makes it happen (agent), or is the object (patient). Food is a good example. While the Merriam Webster

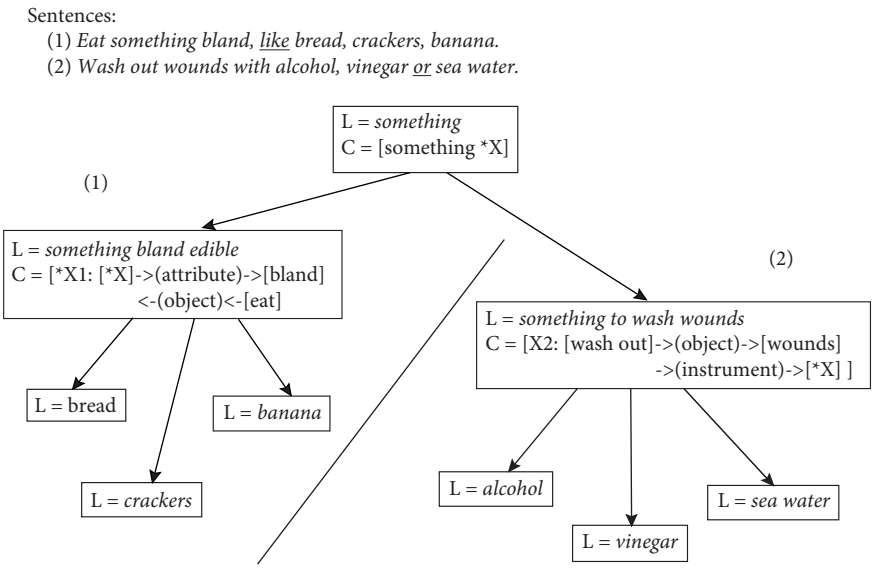

Figure 7. Action centered unlabeled concepts.

Online dictionary gives *material* as its hyperonym, our daily usage conceptualization is more in relation to *eating*, making this CG: [eat] → (object) → [*X] a closer representation to our concept of food.

Although obvious groupings are found, a difficulty remains about the placement of such groupings in a concept hierarchy. The relation to the verb is actually the strongest shared properties among the subclasses, and therefore a concept node representing the superclass must give this information, the labeling is not important. In Figure 7, the nodes are placed under [something] but this is certainly inadequate. We leave it to future research to look specifically at this type of verb-related unlabeled concepts.

In this section, we presented, via manually analyzed examples, the validity of our concept node representation formalism and we will, next, describe a procedure for doing so semi-automatically.

5. Steps for building the concept hierarchy through corpus analysis

The present section examines the required steps to perform a semi-automatic construction of a concept hierarchy from corpus analysis. The present system

used in our laboratory, SeRT (Semantic Relation in Texts), has been designed to perform semi-automatic identification of semantic relations, term extraction, as well as storage of terms and their interrelation in a predicate format (e.g., *hyperonym(pipeline inspection, overhead environment diving)*) as well as visualization of the predicates in a graphic form. The ideas presented in this article emerge more from our wish-list for SeRT than from its present capabilities. The limitations of the representation capability encouraged us to investigate a conceptual approach and to move toward a conceptual graph representation of information in future versions.

The text-to-concept hierarchy process can be broken down into five stages (1) identification of semantic relations in text including hyperonyms and other defining relations; (2) identification of nominal entities involved in the semantic relation; (3) transformation of the nominal entities into Conceptual Graphs (CGs); (4) disambiguation of the concepts and relations within the CGs; (5) positioning of the concept node in the concept hierarchy. We now examine each step:

1. *Identification of semantic relations in text*

This is the only step actually performed in SeRT. It works with a flexible definition of linguistic patterns. Using a mix of lexical and part-of-speech tags allows for the discovery of semantic relations. For a good study of the limitations of knowledge patterns, we refer the reader to Meyer (2001) who looks in details at types of relations and knowledge patterns to find them. Many have studied the hyperonym relation and many linguistic patterns are known and can be used. Other relations also have specific patterns, but more exploration is required, especially as there is a debate regarding the domain-specificity of such patterns. The two most important problems (in our view) are the ambiguity of some knowledge patterns (they can lead to different semantic relations) and their low precision (they can lead to other type of information). To limit human intervention, non-ambiguous and high precision patterns should be identified and used first.

2. *Identification of the concepts involved in the semantic relation*

The automatic identification of the concepts linked by a semantic relation requires a partial parsing to identify the nominal entities (NPs) right and left of the relation. This follows a limiting hypothesis of concepts being expressed by NPs. We argue in this research for a larger view of expression of concepts, but we think that, for automatization purposes, it would be a natural first step before trying to find concepts that are expressed by sentences or VPs. Parsing

technologies are not 100% accurate but have made steady progress in the recent years with the availability of larger dictionaries as well as large corpora of text to help derive probabilistic grammars which can assist in resolving ambiguities. A current popular parsing methodology is based on the Link Grammar (Sleator and Temperley 1993).

3. *Generation of Conceptual Graphs from NP*

The generation of the Conceptual Graph corresponding to a NP requires a NP parser as well as a set of rules to generate a CG from the parse tree. A subset of these rules, defined in previous work, can transform dictionary definitions (taken from the American Heritage First Dictionary) into conceptual graphs (Barrière 1997). Any set of rules must be defined with a set of relations in mind. Sowa (1984, appendix) has defined a set of relations, so have many other researchers. There is no agreement among researchers, but as we have argued elsewhere (Barrière 2002), the variations often emerge from a variation in the granularity of the expression and using a hierarchy of relations could bring them together. A hierarchy would also allow for variations in the more specific relations (finer granularity) as influenced by the domain.

4. *Disambiguation of the Conceptual Graph*

Disambiguation is the most difficult part; ambiguity residing both in the relations and the concepts within the CG. Let us look at the concepts first. A CG should be an integration of multiple concepts, but after step (3), the CG is simply an integration of words since it has been generated from the NP found in the text. Each word can have multiple meanings and the finding of the appropriate meaning for this particular CG is problematic. Word sense disambiguation has always been and still is a major concern in Natural Language Processing (NLP) research. It is less problematic in a terminological setting as most disambiguation problems arise from high frequency general language words. Under the hypothesis that within a very specific domain, word senses are more restricted, this disambiguation problem is less of an issue.

The disambiguation of relations is also a difficult problem. Relations are inferred from textual cues (knowledge patterns) in text which can be ambiguous (especially prepositions). Moreover, if we wish to generate CG from noun compounds (as we mentioned in Section 3) to explicitly express compositionality, then there are not even any knowledge patterns to rely on, and the semantic relation must be inferred solely from the concepts involved in it (Barker and Szpakowicz 1998). This is still a research issue. In our own research, we have looked at both semantic relation and concept disambiguation (Barrière and Popowich 1996; Barrière 1998) in the restricted context of a children's dictionary.

5. *Positioning of the concept*
Each new concept node needs a label, a Conceptual Graph representation and a position within the concept hierarchy. The position, if not obviously given by the sentence from which the node was constructed, will then be dependent on the concepts from which the CG is made up. A comparison with other CGs in the concept hierarchy will be required. A CG platform which can handle the known CG operations such as internal join, external join, maximal common subgraph will therefore be required. There are at least two popular platforms of this type available: Cogitant (Genest and Salvat 1998) and Prolog+CG (Kabbaj et al. 2001).

5.1 Issues and possible problems

Although we have not included it as a necessary step above, the labelling of concepts might be interesting in a Natural Language Generation application. We emphasized the idea of unlabeled concepts by focusing on the importance of the CG representation. Some CG representations have a stable term as a label (or could even have different term variants as a set of labels (Jacquemin 2001)), other CGs will be unlabeled in which case a label 'generator' would have to be used to produce an automatically labelling from the CG (Nicolov et al. 1995).

A second issue relates to our hypothesis of "relation-first" in complement to the "term-first" extraction. This hypothesis could lead to the extraction of concepts having no connexion to important terms in the domain. To avoid such situations we could automatically detect isolated clusters of unlabeled concepts not connected (through references) to the main concept hierarchy.

6. Conclusions and future work

Our work applies corpus analysis methodology to discover concepts in text. While most researchers in computational terminology search exclusively for terms, we believe that concepts are not only mental constructs accessible through introspection but also an integral part of text. As such, they are discoverable through extensions to existing corpus analysis techniques, by finding surface patterns indicative of hyperonym relations independently of terms.

Although unlabeled concepts can be related to the lexical gaps apparent in machine translation (comparing texts in different languages immediately brings forward the intricacy of the word/concept relation) our hypothesis is that unlabeled concepts can be discovered within a monolingual environment and

be used to anticipate lexical gaps even in the absence of the comparative setting provided by two languages. Unlabeled concepts serve a mediating role in knowledge organization by allowing for nuance and refinements in the representation of the elements of a domain, leaving space for unlabeled groupings of elements.

We suggested that concept nodes within the lattice have a definite structure made of label, superclass, definitional characteristics and examples (subclasses). As superclass and subclass are implicit in a hierarchical organization, each node can therefore be reduced in the lattice to a tuple: label and a conceptual representation. The conceptual graph formalism has been chosen for its flexibility and expressiveness and we have shown through different examples how it is adequate as the conceptual representation of each node. Because of its compositionality, it is also adequate for the positioning of concept nodes in the concept hierarchy.

We plan to further study the organization of the concept lattice, especially the interesting cases of unlabeled concepts which have no superordinate and therefore no obvious place in the hierarchy. We also plan to further study the relation between labels and conceptual representation. Much work needs to be done in the automatic generation of conceptual representation from multiword labels to disambiguate the compositional meaning. Much work needs also to be done in the opposite direction, in the generation of a natural sounding descriptive label from a conceptual representation.

We believe that TKBs based on concept lattices would better cover the important information — the concepts — in a domain and would therefore better fulfill the goals of the terminological processes that use them. Also, for any knowledge modeling exercise, the construction of a concept lattice, emphasizing both labels and conceptual representations, would provide a very valuable and thorough framework for the organization of the information.

Notes

* This research was made possible by the financial support of the National Science and Engineering Research Council of Canada. It has been conducted while the author was at the School of Information Technology and Engineering, University of Ottawa. The author would like to thank Terry Copeck for the software development of SeRT. The authors would also like to thank Elizabeth Marshman and Ingrid Meyer, School of Translation and Interpretation, University of Ottawa, for giving us access to their scuba-diving corpus.

1. Interestingly, *cupboard* is one of those semi-transparent names which started with a compositional meaning but extended into a slightly different meaning. The Online Etymology Dictionary (www.etymonline.com) gives: *Cupboard* (c.1325) was originally a board or table to place cups and plates on; sense extended 1530 to "a closet or cabinet for food, etc."

References

American Heritage First Dictionary. 1994. Boston, Mass: Houghton Mifflin.

Barker, K. and S. Szpakowicz. 1998. "Semi-Automatic Recognition of Noun Modifier Relationships." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*. 96–102. Montreal, Canada.

Barrière, C. 1997. *From a Children's First Dictionary to a Lexical Knowledge Base of Conceptual Graphs.* Ph.D. Thesis, Simon Fraser University.

Barrière, C. 1998. "Redundancy: helping semantic disambiguation." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*. 103–109. Montreal, Canada.

Barrière, C. 2001. "Investigating the Causal Relation in Informative Texts." *Terminology* 7(2), 135–184.

Barrière, C. 2002. "Hierarchical Refinement and Representation of the Causal Relation." *Terminology* 8(1), 91–111.

Barrière, C. and T. Copeck. 2001. "Building a domain model from specialized texts." In *Proceedings of Terminologie et intelligence artificielle (TIA 2001)*. 109–118. Nancy, France.

Barrière, C. and M. Hermet. 2002. "Causality taking root in Terminology." In *Proceedings of Terminology and Knowledge Engineering (TKE 2002)*. 15–20. Nancy, France.

Barrière, C. and F. Popowich. 1996. "Building a noun taxonomy from a children's dictionary." In Gellerstam, M. et al. (eds.). *Proceedings of Euralex '96.* 27–35. Göteborg, Sweden.

Barrière, C. and F. Popowich. 2000. "Expanding the type hierarchy with non-lexical concepts." In *Proceedings of Canadian AI.* 53–68. Banff, Canada.

Bowden, P.R., P. Halstead and T.G. Rose. 1996. "Extracting Conceptual Knowledge From Text Using Explicit Relation Markers." In Shadbolt, N., K. O'Hara and G. Schreiber (eds.). *Proceedings of the 9th European Knowledge Acquisition Workshop, EKAW'96.* 147–162. Nottingham, United Kingdom.

Cabré Castellvi, M.T., R. Estopà Bagot and J.V. Palatresi. 2001. "Automatic term detection: a review of current systems." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds). *Recent Advances in Computational Terminology.* 53–87. Amsterdam/Philadelphia: John Benjamins.

Cartier, E. 1997. "La définition dans les textes scientifiques et techniques: présentation d'un outil d'extraction automatique des relations définitoires." In *Actes des deuxièmes rencontres — Terminologie et Intelligence Artificielle, TIA'97.* 127–140. Toulouse, France.

Chung, T.M. 2003. "A corpus comparison approach for terminology extraction." *Terminology* 9(2), 221–246.

Condamines, A. and J. Rebeyrolle. 1998. "CTKB: A Corpus-based Approach to a Terminological Knowledge Base." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds.). *First Workshop on Computational Terminology (Computerm '98)*. 29–35. Montreal, Canada.

Cruse, D. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Davidson, L., J. Kavanagh, K. Mackintosh, I. Meyer and D. Skuce. 1998. "Semi-automatic Extraction of Knowledge-Rich Contexts from Corpora." In Bourigault, D., C. Beauchemin and M.-C. L'Homme (eds). *Computerm'98*. 50–56. Amsterdam/Philadelphia: John Benjamins.

Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Garcia, D. 1997. "Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes." In *Actes des deuxièmes rencontres — Terminologie et Intelligence Artificielle, TIA'97*. 7–26. Toulouse, France.

Genest, D. and E. Salvat. 1998. "A platform allowing typed nested graphs: How cogito became cogitant." In *Proceedings of the Sixth International Conference on Conceptual Structures (ICCS-98)*. 154–161. Berlin: Springer Verlag.

Hearst, M. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora." In *Proceedings of the International Conference on Computational Linguistics (COLING-ACL '92)*. 539–545. Nantes, France.

Jacquemin C. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge: MIT Press.

Jacquemin, C. and J. Royauté. 1994. "Retrieving terms and their variants in a lexicalized unification-based framework." In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 132–141. New York/Heidelberg: Springer-Verlag.

Kabbaj, A., B. Moulin, J. Gancet, D. Nadeau and O. Rouleau. 2001. "Uses, improvements, and extensions of Prolog+CG: case studies." In *Proceedings of the 9th International Conference On Conceptual Structures (ICCS'2001)*. 346–359. Stanford University, California.

Kageura, K. and B. Umino. 1996. "Methods of automatic term recognition: a review." *Terminology* 3(2), 259–290.

Lehrer, A. 1974. *Semantic Fields and Lexical Structure*. Amsterdam/London: North-Holland Publishing.

Merriam Webster Online Dictionary. http://www.m-w.com/cgi-bin/dictionary.

Meyer, I. 2001. "Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds.). *Recent Advances in Computational Terminology*. 279–302. Amsterdam/Philadelphia: John Benjamins.

Meyer, I., K. Mackintosh, C. Barrière and T. Morgan. 1999. "Conceptual sampling for terminographical corpus analysis." In *Terminology and Knowledge Engineering (TKE'99)*. 256–267. Innsbruck, Austria.

Nicolov, N., C. Mellish and G. Ritchie. 1995. "Sentence generation from conceptual graphs." In Ellis, G., R. Levinson, W. Rich and J. F. Sowa. (eds.). *Conceptual Structures: Applications, Implementation and Theory, Third International Conference on Conceptual Structures, ICCS '95*. 74–88. Santa Cruz, United States.

Pearson, J. 1998. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

Resnik, P. 1995. "Using information content to evaluate semantic similarity in a taxonomy." In *Proceedings of the 14th* International Joint Conference on Artificial Intelligence *(IJCAI)*. 448–453. Montreal, Canada.

Sager, J., D. Dungworth and P. F. McDonald. 1980. *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

Sleator, D. and D. Temperley. 1993. "Parsing English with a link grammar." In *Third International Workshop on Parsing Technologies*. 277–192. Tilburg, Germany.

Sowa, J. 1984. *Conceptual Structures in Mind and Machines*. Reading, Mass.: Addison-Wesley.

Trimble, L. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Vanderwende, L. 1994. "Algorithm for automatic interpretation of noun sequences." In *Proceedings of International Conference for Computational Linguistics (Coling 1994)*. 782–788. Tokyo, Japan.

Wüster, E. 2003. "Historical readings in terminology: The wording of the world presented graphically and terminologically." [translated by Juan C. Sager]. *Terminology* 9(2), 269–297.

Wüster, E. 2004. "Historical Readings in Terminology: The structure of the linguistic world of concepts and its representation in dictionaries." [translated by Juan C. Sager]. *Terminology* 10(2), 281–306.

Author's address

Caroline Barrière
Institute for Information Technology
National Research Council Canada
Université du Québec en Outaouais
Pavillon Lucien Brault
100 rue St-Jean-Bosco
Gatineau (Québec) J8Y 3G4
Canada

Caroline.Barriere@nrc-cnrc.gc.ca

About the author

Caroline Barrière has a doctorate in Computational Linguistics from Simon Fraser University, as well as a master's degree in Electrical Engineering and a bachelor's degree in Computer Engineering from École Polytechnique de Montréal. Her research fields are computational terminology and lexicography and their applications for machine translation and language learning. Her work focuses on knowledge extraction from dictionaries and corpora and its conceptual representation.