Marie-Claude L'Homme
Observatoire de linguistique Sens-Texte (OLST)
Département de linguistique et de traduction
Université de Montréal
Montréal (Québec), Canada

DEFINING CONCEPTS: STARTING FROM CONCEPTS THEMSELVES OR FROM EMPIRICAL DATA. A COMMENT ON EKATERINA MHAANNA'S ARTICLE ENTITLED: "PROCESSUAL CONCEPT RELATIONS BY METHODS OF TERMINOLOGY"

1 INTRODUCTION[1]

The article by Ekaterina Mhaanna presents part of the author's thesis devoted to the modelling of relations between concepts. As the author rightly points out, the issue of relations is central in terminology (as it is in many other disciplines dealing with relational data). The author focuses on complex and non-hierarchical relations – such as those involving an entity and a process – defines them, and proposes a model for representing them formally. This work aims to contribute to a better understanding of the set of relations in which concepts appear (and eventually lead to descriptions of the relations between terms that denote these concepts).

The relations considered are particularly relevant, since it can be assumed that they can be found in several different fields of knowledge. Surprisingly, they have not been dealt with extensively in terminology, contrary to other well documented relations, such as hyperonymy-hyponymy, meronymy-holonymy, and cause-effect. Process relations have been taken into account by Feliu (2004) who has analyzed their application in the field the human genome. They have also been characterized in other theoretical frameworks (e.g., in lexical semantics, Fillmore 1968; FrameNet), although for a different purpose (describe lexical units in terms of argument structures).

The author's concern for implementing formal definitions of these relations in existing hierarchies of relations (based on work by Nistrup Madsen et al. 2001) is particularly relevant, especially when considering that ontologies or other forms of knowledge representations are being designed in various different fields. There is an acute need for formal definitions of concepts and of the relations in which they can be found.

The author is very careful and advises readers that models previously designed are open to discussion and that some of their components can be revised. In fact, it would probably suffice to say that conceptual models are always the reflection of a given point of view on data or on knowledge. My comment will illustrate another perspective which is again only the reflection of a specific point of view on terminological data.

STARTING FROM CONCEPTS OR STARTING FROM DATA?

Process relations are different from other central relations considered up to now in terminology in the following ways:

- They involve two units that belong to two different classes (often, an entity and an activity);
- They are non-hierarchical;
- The activity can be realized linguistically by a noun, but also by a verb.

The approach taken by the author consists in defining the relation and the concepts it involves very rigorously. The subsequent aim is to investigate corpora to analyze how these concepts and relations are realized in text.[2] I will take a different approach and raise a few questions on the potential findings it can provide. The general idea is to try to find out if data leads us to a list of relevant relations instead of starting the other way around.[3]

My perspective is based on the following:

- Looking at the issue of concepts (perhaps, I should say senses) and relations using as a starting point texts and the data they contain; i.e., linguistic forms of senses are the starting point of this analysis;
- The observation of interactions of linguistic forms with others lead to the delimitation of their senses; they will also lead us to a set of relations that can differ from one unit to another.

In other words, my perspective is corpus-based and relies on structural lexical semantics. To illustrate its application, I used an example borrowed from the field of computing.[4]

a) Linguistic forms as a starting point

If the starting point of our analysis are linguistic forms found in specialized corpora, a selection can be made in order to retain forms that are relevant for a given terminology project (that can be the design of a specialized dictionary, the enrichment of a term bank, the population of an ontology, etc.). In text, different linguistic forms can be used to designate specialized entities. For example, the following linguistic expressions can refer to "a peripheral used for producing text of images on paper from data contained in the computer": *printer, printing peripheral, device used for printing.*[5]

Terminologists (or other specialists concerned with the collection of terms) can decide to take into account only part of these linguistic forms. They can set aside printing peripheral if they consider that it can be interpreted as a generic form when taken out of context. They can also set aside the phrase device used for printing which can be considered as a defining paraphrase. Other analysts could determine that all these forms are relevant for they can be used to access information about senses in texts (e.g., as in computational terminology where term variants have been characterized in order to devise techniques to handle them automatically (Daille et al. 1996)).

b) Interaction of linguistic forms

A corpus will reveal a number of paradigmatic relations, such as those listed in Table 1. Some of these relations can be found in text using linguistic markers. A lot of work has been carried out in this area during the past decade (Ahmad & Fulford 1992; Meyer 2001, among others).[6]

Table 1. Units interacting with printer (paradigmatic relations)

peripheral	Hyperonym (generic term)
laser printer, network printer, color printer, portable printer	Hyponyms (expressed by printer + a noun or an adjective)
scanner, modem, monitor	Co-hyponyms: (based on shared collocates)
printhead	Meronym
user	Agent
data, files	Patient
hard copy	Result
paper (placed in the printer); toner (in a laser	Other paradigmatic relations

printer), etc.	

A corpus will also lead us towards syntagmatic relations, such as those listed in Table 2.

Table 2. Units interacting with printer (syntagmatic relations)

(expressed by a noun) printing; (expressed by verbs) (someone prints something with a printer) print	Typical activities done by the agent
connect (to connect the printer to a computer), configure a printer, turn on a printer	Other activities done by the agent
the printer prints something, the printer crashes	Activities done by the printer itself

REPRESENTING RELATIONS BETWEEN LEXICAL UNITS

The paradigmatic and syntagmatic relations listed in the previous section can be represented using formal apparatuses, some of which are increasingly used in terminology work. For most paradigmatic relations (hyperonymy, meronymy), ontology editors offer mechanisms to represent them and take into account most of their specificities. However, other tools must be used to represent syntagmatic relations.

Syntagmatic relations (i.e., collocations) require a description of the syntactic structures in which the key word (in this case, printer) participates (for example, to distinguish the two structures in which it appears when it interacts with the verb print), the argument structure of the unit (in order to know which arguments are involved in the meaning of a collocate) and the meaning the collocate.7 The example below shows how the two senses of print can be represented in terms of their combination with printer.

> Define a printer as: A peripheral used by a USER to produce a hard copy of DATA stored in the computer.
>
> > the USER acts on DATA with the printer : the user prints1b data with the ~ the printer acts on DATA: the ~ $prints_{1a}$ data

CONCLUDING REMARKS

In this very short comment, I have tried to show how the issue of relations can be examined from the perspective of empirical data rather that from the abstract concepts some of these units denote. This account should not be interpreted as a form of criticism of the work presented in Mhaanna's paper. It simply aims to show that, according to the perspective chosen to look at a problem, data can reveal different facets.

In order to account for what has been said above about the analysis of "printer" and the heterogeneity of the data as it is found in running text, I allowed myself to add a new level to the 3 levels representation reproduced in Mhaanna's article. The original representation accounts for the relations between the objects of the world, the abstraction and generalization of these objects in terms of concepts, and the representation of concepts in terms of linguistic forms. It would be more appropriate to say that this new level is a specification of level 3, hence a subdivision of this level.

Table 3. An addition to the 3 levels representation in terminology (based on the Table provided in Mhaanna, in this volume)

Level 1	Level 2	Level 3	Level 3a
The concept refers to the object in the world	The concept refers to the abstraction in the form of a concept	The concept refers to the representation of the concept	Linguistic units used in running text to express knowledge
Object	Concept	Designation	terminological units (base form and variants)
Property	Characteristic	feature specification	the meaning of units is the result of their interaction with other units
ontical system	concept system	concept diagram	terminological units share paradigmatic and syntagmatic relations with other units
ontical relation	concept relation	relation specification	Each relation can be specified with a formal apparatus

I also believe that the two perspectives illustrated in Mhaanna's article and in this comment are useful for terminology work but on different levels. Corpus-based studies such as the one I illustrated cannot simply take as a starting point the data as it presents itself in text. They need to be supported by a theoretical framework (be it conceptual of semantic). Terminologists, when looking at linguistic data in text, must define the object they are analyzing in the most precise manner. Similarly, relations must be characterized before being investigated in text.

On the other hand, abstract models such as that defined in Mhaanna's article become entirely useful once they have been validated on a significant amount of data. And this data, in terminology, remains, what can be found in specialized texts.

This short comment simply shows, once again, that terminological data, terminological concepts, and terminology work cannot be considered from a single point of view to account for their complexity. This has already been pointed out by many scholars and is once more demonstrated here.

[1]I would like to thank Ekaterina Mhaanna for sending me a revised version of her article. I will also take this opportunity to mention that I have not consulted the thesis of the author and that my comment is based exclusively on the material available in the article.

[2]This part of the work, however, is not reported in the article.

[3]Of course, this analysis is not as thorough as that of the author's or as that of the work cited by the author. The examples provided below are simply given to illustrate my point of view and can serve as a basis for discussion.

[4]The example is based on work carried out at the Observatoire de linguistique Sens-Texte (OLST). We are currently developing the English version of a dictionary on computing and the Internet. Thanks are extended to Louis-Philippe Dargis who has contributed to the compiling of the corpus and the analysis of terms.

[5]The example given here does not involve a polysemic unit. In this case, the analysis will have to proceed to some form of disambiguation (e.g., page: ”a Web page,” ”a division of the memory,” or ”a division of a document”).

[6]The paper by Sergio Barrios (in vol 19, 2008) also lists a number of linguistic markers in a corpus of law in Portuguese.

[7]This model is based on lexical functions (Mel’cuk et al. 1995).

REFERENCES

AHMAD, K. & H. FULFORD (1992). Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology. Report CS-92-07. Guildford: University of Surrey.

DAILLE, B., B. HABERT, C. JACQUEMIN & J. ROYAUTÉ (1996). Empirical Observation or Term Variation and Principles for their Description. Terminology 3(2), 197-257.

FELIU, J. (2004). Relations conceptuals i terminologia : anàlisis I proposta de detecció semiautomatica. Barcelona: Institut Universitari de Lingüística Aplicada (IULA).

FILLMORE, C.J. (1968). The Case for Case. In Bach, E. & R.T. Harms (eds.). Universals in Linguistics. New York: Holt, Reinhart and Winston, 1-88.

FrameNet. http://framenet.icsi.berkeley.edu/. Accessed 8 November 2007.

MEL'CUK, I., A. CLAS & A. POLGUÈRE (1995). Introduction à la lexicologie explicative et combinatoire. Louvain-la-Neuve (Belgique): Duculot / Aupelf - UREF.

MEYER, I. (2001). Extracting Knowledge-Rich Contexts for Terminography. In Bourigault, D., C. Jacquemin & M.C. L'Homme (eds.). Recent Advances in Computational Terminology. Amsterdam/Philadelphia: John Benjamins, 279-302.

NISTRUP MADSEN, B., B. SANDFORD PEDERSEN, H. ERDMAN THOMSEN (2001). Defining Semantic Relations for OntoQuery. In Anker Jensen & P. Skadhauge (eds.). Ontology-based Interpretation of Noun Phrases. First International OntoQuery Workshop. University of Southern Denmark, Department of Technical Language, Communication and Information Science, 57-88.