Marie-Claude L'Homme, Université de Montréal

A Computerized Model for Processing Lexical Combinations in Technical Language

Abstract

This paper will address questions related to the processing of lexical combinations in computerized tools. The focus is on word combinations found in special languages (more precisely, technical language), and their relevance for specialized translation. A model for the implementation of lexical combination data in existing terminological databases is proposed. This model links databases describing polysemic lexical units and terminological databases (existing term banks). The link is performed using a conceptual coding of nominal units reproduced in all databases. The focus will be on the combinations studied up to this point, i.e., verb + term and deverbal noun + term combinations.

1. Introduction

Terminological databases (term banks) are widely used by specialized translators to access the equivalents of technical terms or decode their meanings. However, term banks do not provide information on lexical combinations comprising terms. These are complex units translators need to access in order to provide idiomatic translations of longer expressions.

The representation of specialized lexical combinations in reference works (computerized or not) enables translators:

a) To reproduce a particular specialized usage in a target language;
b) To delimit more specifically the different meanings of a polysemic unit; e.g.

release (gas)	libérer
release (bolt, screw)	desserrer
release (key, button)	relâcher
release (brake)	desserrer
etc.	

c) To delimit more accurately differences between languages; e.g. the verb *fail* will have different French equivalents according to the "entity" that fails (e.g. *pump, condenser, fuse*).

Even though computerized tools, and more specifically databases, present several advantages (over paper references) for storing specialized lexical combinations, there is no general agreement as to how these complex units should be processed. Some proposals have been made (Fontenelle (1994), Gouadec (1992), Heid (1992), and Heid & Freibott (1991), among others), but none of them seems to apply to the combinations we have studied.

This paper presents a model for the storage and retrieval of specialized lexical combinations in a computerized tool. The model provides for a link with existing terminological databases designed for translation purposes. The examples provided below are based on English-French lexical combinations, but the approach can easily be extended to other language pairs. A model has already been developed for V + N groups. We have recently developed an extension for French deverbal nouns + term combinations. The presentation will thus focus on these groups. We are currently studying other combinations comprising a terminological unit (adjective + term; non-deverbal noun + term) in order to store them in the computerized model.

We will provide a very general description of the model. Other specifications are discussed in L'Homme (1995) and L'Homme (1996). These papers deal with verb + term combinations. A thorough discussion of the extension to deverbal noun + term combinations will be available shortly.

2. Background: Specialized Lexical Combinations

The lexical combinations our computer model is concerned with have the following characteristics:

a) They are used in a special language (namely technical language);

b) Both units in the combination are linked grammatically and their use in a given context is determined by usage within a group of speakers (in this case, a group of specialists);

c) combinations comprise a terminological unit (TU) which is nominal in nature and another word (a verb or a noun). The combinations examined in this paper contain the following units:

A verb and a terminological unit (V + TU; TU + V; V + prep. + TU: *traiter de l'information, boot a computer*);

A deverbal noun and a terminological unit (N + prep. + TU: *traitement de l'information, access to a file*).

d) The lexical unit with which the term is used is generally polysemic;

e) Both units (the term and the lexical unit it is combined) are autonomous (they can be used in several other contexts, and can appear in other combinations);

ea) A lexical unit can be combined with several terms. In the following example, the verb *release* can be combined with *screw, gas*, etc.

adjust	
apply	*screw*
compress	*gas*
hit	*brake*
pack	*part*
press	*key*
RELEASE	*data*
secure	
tighten	

eb) A term can be used with several lexical units. In the following example, *data* can be combined with several lexical units.

ACCESS		
apply	*screw*	
COMPRESS	*gas*	
hit	*brake*	
PACK	*DATA*	
press	*key*	
PROCESS		
release		*PROCESSING (of)*
SAVE		*PACKING (of)*
secure		*ACCESS (to)*
tighten		

ec) A lexical unit (with a given meaning) can be used with a wide range of terms that belong to the same concept class. In the following example, the verb *release* can be used with all sorts of "fasteners".

RELEASE	*SCREW*
	SET SCREW
	SOCKET HEAD SCREW
	SQUARE HEAD SCREW
	SQUARE HEAD BOLT
	etc.

Each unit involved in combinations like those described above has a high degree of combinability. For this reason, we will refer to them as *lexical combinations* rather than *collocations*. The latter term is generally used to designate groups in which the association of the units involved cannot be determined by the meaning or the syntactic properties of either unit (Mel'cuk *et al.* 1995).

3. A Model for the Storage of Specialized Lexical Combinations

3.1 General Objectives

The computer model proposed has the following technical objectives:

a) Provide access to lexical combinations from either unit in the group;

b) Provide access to lexical combinations from both units used in the group;

c) Provide access to lexical combinations from the standard terminological record; conversely, permit access to term records from the lexical combination data;

d) Provide additional information on both lexical components in the combination (the information related to the terminological unit is provided in the term record); additional information can be added to other lexical units.

3.2 Terminological and Verbal Records

In order to meet the objectives listed above and to reproduce the characteristics proper to specialized lexical combinations, we proposed the following scenario. First, separate databases are created for the description of each lexical unit involved in the combination. One is

dedicated to the description of terminological units (the actual term bank). Other databases are created to describe the other linguistic units used with the term (currently verbs and deverbal nouns): each part of speech is described in a separate database. Then, a link is established between the entries contained in these databases and the records contained in a terminological database. The entries are supplied with all relevant information (definition, context, etc.). We have reproduced, below, an example of a verbal entry and a terminological entry.

Terminological entry

Entry	data file
Def.	A file containing data records.
Cont.	In other systems, programs and data are read into data files...
Entrée	fichier de données
Déf.	Fichier qui comporte un ensemble de données.
Cont.	Tant qu'un fichier de données se sera pas ouvert, les options...
Dom.	informatique

Verbal entry

Entry	delete
Struc.	V + O
Def.	To remove something that has previously been written or stored.
Cont.	...if you delete a page by mistake... / Whenever you erase (or delete) a file...
Entrée	effacer
Struc.	V + O
Déf.	Faire disparaître quelque chose qui a été préalablement inscrit ou stocké.
Cont.	...on efface toutes les données et programmes dont on n'a plus besoin... / Positionnez le curseur à l'endroit où vous voulez effacer une (des) ligne(s)...

The terminological record is typical of records that can be found in monoreferential terminological databases (definition, subject field, context, etc.). The verbal record contains a definition, the description of standard syntactic structures, a definition (each meaning is described in a different record), and one or more contexts.

3.3 Use of Concept Classes

Besides the typical lexicographic information, the records in both databases will contain tags used to link verbal and terminological records. Units that can be combined will convey the same tag. For instance, a given verb (*delete*) can be used with direct objects referring to information-related concepts (*data, data file,* etc.). It can also be used with subjects referring to humans (*user, operator*, etc.). We use a tag referring to the conceptual class to which these terms belong and supply it in the database. For example, the terms that belong the class of

PEOPLE can be used as subjects of *delete*; the terms that belong to the class of *ENT REP* can be used as direct objects of *delete.* The tags are supplied in fields indicating the syntactic functions of the terms that can be combined with a verb. All the units appearing in the terminological database are tagged using the same system. A simplified example is reproduced below.

Terminological record

Entry data file
Def. A file containing data records.
Entrée fichier de données
Déf. Fichier qui comporte un ensemble données.
Dom. informatique
Concept ENT REP

Verbal record

Entry	delete
Struc.	V + O
Def.	To remove something that has previously been written or stored.
Subject	PEOPLE
Object	ENT REP
Entrée	effacer
Struc.	V + O
Déf.....	
Sujet	PEOPLE
Objet	ENT REP

The conceptual classification made in the terminological database and in the other databases is hierarchical. For example, the terminological unit *data* is defined as a "content-only" concept. "Content-only" is a concept class comprised in a wider class, namely "convey", and so on and so forth.

DATA < CONTENT-ONLY (class that comprises concepts referring to realities used for disseminating of information that are considered according to their content);

CONTENT-ONLY < CONVEY (class that comprises all concepts referring to realities used for for disseminating information);

CONVEY < REPRESENTATIONAL ENTITY (class that comprises all concepts referring to realities used for transmitting, reproducing, or disseminating information);

REPRESENTATIONAL ENTITY < ENTITY (class that comprises all concepts referring to individual objects).

The hierarchical system enables the coding of concept classes on different levels. A verb, for instance, may combine with terms belonging to very large concept classes or, on the contrary, very restricted classes.

3.4 Deverbal noun + term combinations

Deverbal nouns have been examined separately since the combinability of several deverbals can be described according to the verb from which they are derived. Deverbals are also described in a separate database but much of the information supplied is transferred from the verbal database.

To illustrate this, we will discuss an example with the verb *démarrer* and the deverbal *démarrage*. We will first reproduce the verbal entry *démarrer*.

démarrer	
Struc.	V + O
Def.	Mettre quelque chose en état de fonctionner.
Subject	PEOPLE, COMPUTERS, PROGRAMS
Object	COMPUTERS
Cont.	Si on démarre un ordinateur après avoir débranché le clavier....

It is said here that *démarrer* can be used with direct objects referring to COMPUTERS (e.g. *portable computers*, *microcomputers*, etc.) and with subjects denoting PEOPLE, COMPUTERS and PROGRAMS (e.g. *user*, *system*, *software*). *Démarrage* can also be used with terms that belong to the same concept class but the syntactic structure of the whole will be different. The object of the verbal structure will be transformed into a prepositional phrase, and the subject into an additional prepositional phrase in which the noun denotes an agent (*démarrage de l'ordinateur par l'utilisateur*). We simply transfer the tags from the verbal record into appropriate fields in the deverbal noun database. (The agent will not appear in the deverbal record, since this type of data will not be requested by users).

démarrage	**démarrer**	
	Struc.	V + O
déf. [.....] ◄	Def.	Mettre quelque chose en état de fonctionner
	Sujet	PEOPLE, COMPUTERS, PROGRAMS
complément1 ◄———	Objet	COMPUTERS

More information can then be supplied in deverbal records, such as contexts and the standard syntactic structure. The definition provided for

a verb can also be transferred and preceded by a phrase such as [Process of...]. We have reproduced, below, an example of a more complex deverbal noun entry. In this entry, part of the definition, and the conceptual classes the object of the verb (ENT REP) and the adjunct (storage) were transferred from the verbal entry *charger.*

chargement
Struc. N de C1 + C2(en, dans)
Déf. [Activité qui consiste à] Placer, disposer quelque chose dans un endroit où on peut l'utiliser.
Complément1 ENT REP
Complément2 STORAGE
Contexte Le mode Assistance apparaît à l'écran lors du chargement de dBASE III Plus... / ...chargement d'un fichier du disque en mémoire principale...

Verbe associé charger V + O + A(en, dans)

4. Conclusion

Tests were conducted on databases describing different lexical units in the language of computing. Although the corpus may not yet be representative of technical language as a whole and problems remain to be solved, the computer program based on the method described in this paper appears to be an efficient technique for storing accessing, retrieving, and managing specialized lexical combinations.

Of course, all terms in a terminological database must be marked with tags referring to conceptual classes, but the remainder of the term bank structure is not altered. Moreover, new term bank models conceptual networks take into account. The tagging will be done eventually one way or the other. Verbal, adjective, and nominal databases designed for lexical combination retrieval could use any particular tagging represented in these term banks.

The model described in this paper presents a certain number of other advantages. First, all lexical units have a number of descriptions corresponding to their different meanings. There is no useless duplication of lexical forms in different records. Secondly, the use of tags representing conceptual classes prevents the reproduction of all terminological units that might be combined with another lexical unit. Finally, all databases can be updated separately (a term record can be added to the terminological database; if a tag referring to its conceptual class is added to this record, all other lexical units containing that particular tag will automatically be associated with the new term).

References

Benson, M., E. Benson, R. Ilson 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*, Amsterdam/Philadelphia, John Benjamins.

Cohen, B. 1992. «Méthodes de repérage et de classement des cooccurrents lexicaux», in *Terminologie et traduction* 2/3, pp. 505–511.

Collins Cobuild. English Language Dictionary 1987. London/Glasgow, Collins.

Fontenelle, T. 1994. "Towards the Construction of a Collocational Database for Translation Students", in *Meta* 39-1, pp. 48–56.

Gouadec, D. 1992. «Terminologie et phraséologie», in *Terminologie et traduction* 2/3, pp. 549–563.

Heid, U. 1992. «Décrire les collocations. Deux approches lexicographiques et leur application dans un outil informatisé». in *Terminologie et traduction*, 2/3, pp. 523–548.

Heid, U. 1994. "On the Way Words Work Together – Topics in Lexical Combinatorics", in Martin, W. et al. 1994. *Euralex '94 Proceedings*, Amsterdam, pp. 226–257.

Heid, U., G. Freibott 1991. «Collocations dans une base de données terminologique et lexicale», in *Meta*, 36-1, pp. 77–91.

L'Homme, M.C. 1992. «Les unités phraséologiques et leur représentation en terminographie», in *Terminologie et traduction* 2/3, pp. 493–503.

L'Homme, M.C. 1995. "Processing Word Combinations in Existing Term Banks", in *Terminology* 2-1, pp. 141–162.

L'Homme, M.C. 1996. «Méthode d'accès informatisé aux combinaisons de mots (cooccurrents) en langue technique», *Quatrièmes Journées scientifiques. Lexicomatique et dictionnairiques*, Université Lumière Lyon-2 (Lyon), 28–30 sept. 1995, to be published in the proceedings.

Mel'cuk, I., A. Clas, A. Polguère 1995. *Introduction à la lexicologie explicative et combinatoire*, Duculot, Louvain-la-Neuve (Belgique).

Pavel, S. 1993. «Vers une méthode de recherche phraséologique en langue de spécialité», in *L'actualité terminologique* 26-2, pp. 9–13.

Picht, H. 1987. "Terms and their LSP Environment – LSP Phraseology", in Meta 23-2, pp. 149–155.

Roberts, R. 1993. "Phraseology: The State of the Art", in *Terminology Update* 26-2, pp. 4–8.

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

Sager, J.C., K. Kageura 1994–1995. "Concept Classes and Conceptual Structures – Their Role and Necessity in Terminology", in *Terminology and LSP Linguistics. Studies in Specialized Vocabularies and Texts.* Actes de Langue française et de linguistique (ALFA) 7/8, pp. 191–216.