Specialized Lexical Combinations: Should they be described as Collocations or in Terms of Selectional Restrictions?

Marie-Claude L'HOMME and Claudine BERTRAND,
Montréal, Canada

Abstract

This paper is an attempt to show that specialized lexical combinations (word groups used in special languages) – even though they share some similarities with collocations (word groups used in general language) – do not behave exactly like them. Thus, they should not be described using the apparatuses lexicographers usually resort to. We will demonstrate that, in most specialized lexical combinations, co-occurrents can combine with small or large groups of terminological units and that these terms can easily be grouped within larger semantic classes. This demonstration is based on a study conducted at the University of Montreal [Bertrand 1999].

1   Introduction

Many terminologists and other lexicographers have been interested in describing word combinations in special languages [Bergenholtz & Tarp 1995, Cohen 1986, Heid 1994, L'Homme 1995; Thoiron et Biron et Béjoint 1989]. Combinations which have attracted their interest comprise two lexemes that are bound to one another: constraints related to conventions established within a given subject field makes lexeme1 prefer the company of lexeme 2 rather than that of other lexemes. Examples of specialized lexical combinations found in the literature are provided in (1):

(1) *créer un fichier: *établir un fichier* [Heid & Freibott 1991]
     *administrer un médicament: *donner un médicament* [Laporte et L'Homme 1997]

Typically, lexeme 1 is a term (defined as the keyword), a unit with special reference within a specialized subject field. In (1), for example, the terms are *fichier* (*file*) and *médicament* (*drug, medecine*). The other lexeme is often referred to as the *co-occurrent*. In (1), the co-occurrents are *créer* (*create*) and *administrer* (*administer*). The examples also show that terms, such as *fichier* and *médicament*, are preferably used with *créer* and *administrer* (in computer science and medical science) rather than *établir* or *donner* which could be considered as synonyms or near synonyms.

Specialists have often called these combinations *collocations*, a designation borrowed from general lexicography. Some have even used descriptive apparatuses developed for general language combinations (e.g. Cohen 1986 has described combinations used in the field of stock exchange in terms of lexical functions developed by Mel'čuk *et al.* [1984, 1988, 1992, 1995]).

497

However, it has not yet been proven that specialized lexical combinations behave like general language collocations. As will be discussed below, some studies have underlined the discrepancies between the word groups that have attracted the interest of lexicographers and terminologists. We will demonstrate that specialized lexical combinations cannot truly be described as prototypical collocations, since many lexemes defined as co-occurrents can combine with groups of semantically-related terms. For example, in medicine, *administrer* (*administer*) combines with *médicament* (*drug)* (see (1)), but it also co-occurs with other terms (e.g. *réserpine, morphine*). Similarly, in aeronautics, *piloter* (pilot) can combine with *aéronef* (aircraft) but also with *avion* (airplane) and *hydravion* (*seaplane).*

(2)   *administrer un médicament*
        administrer de la morphine
        administrer de la réserpine

        piloter un aéronef
        piloter un avion
        piloter un hydravion

Our demonstration is based on a study conducted by Bertrand [1999] that will be described further. First, we will discuss some common features shared by general and specialized word combinations that can explain why some specialists have envisaged them as word groups with identical behavior. For clarity, general combinations will be referred to as *collocations*; specialized combinations as *specialized lexical combinations* or *SLCs.*

2   Collocations and specialized lexical combinations: similarities

Specialists usually agree on the fact that both collocations and SLCs conform to a conventional usage within a community. [Mel'čuk *et al.* 1995] [1] mention that collocations cannot be accounted for in terms of regular syntactic or semantic rules. Bergenholtz & Tarp [1995] mention that special language users with insufficient linguistic knowledge will not be able to know whether a given word combination is correct in a particular field (e.g. in the field of molecular biology, which sequence among the following is correct: *to cut a gene, to cut out a gene, to cut off a gene, to break a gene*?*).*

Collocations are conventional within a given linguistic community; SLCs are conventional within a group of specialists. Learners of a language or a special language must acquire them as such since they are unpredictable. This "unpredictability" justifies their insertion in a reference tool.

These considerations have led to the compilation of general language dictionaries on collocations [Benson *et al.* 1986], general language dictionaries including collocations [Mel'čuk *et al.* 1984, 1988, 1992], and specialized language dictionaries on SLCs [Cohen 1986: dictionary on the stock exchange].

Based on this common feature, some specialists have taken for granted that collocations and SLCs behave the same way and could be described using similar descriptive models. Co-occurrents are simply listed under an entry defined as a terminological unit. The listing is sometimes further refined with a semantic classification of co-occurrents. Table 1 shows part of an entry extracted from Cohen [1986].

PRODUCTIVITÉ (*productivity*)	
Semantic category of co-occurrents	**co-occurrents**
DÉBUT (beginning)	*décollage, redresser*
CROISSANCE (growth)	*amélioration, accroître, élever*
INDÉTERMINÉS (indeterminate)	*évolution, évoluer*
DÉCLIN (decline)	*baisser, baisse, basse*
FIN (end)	none for *productivité*
AUTRES (others)	*avoir, enregistrer*

Table 1: Part of an entry extracted from Cohen [1986]

3 Co-occurrents combining with semantically related terms: previous studies

Beyond this apparent similarity, studies have shown that collocations and SLCs behave differently. Several authors have noted that co-occurrents in SLCs can combine with small or large groups of terminological units [Heid 1994; L'Homme 1995, 1997, 1998; Meyer & Mackintosh 1994, 1996]. Martin [1992] introduced the notion of "concept-bound" collocations in specialized languages (or sublanguages). According to the author, modifying concepts (i.e. co-occurrents) are often conditioned by some sort of "definitional knowledge" held by the head (i.e. terms) and are not strictly dictated by usage.

Heid [1994], in a study on the dictionary compiled by Cohen [1986], noted that terms denoting an "increase" or a "decrease" (e.g. *hausse, baisse, mouvement, progression, recul, repli, reprise*) have similar verbal co-occurrents as shown in Table 2.

Meaning conveyed by terminological units	Terminological units sharing common semantic features	Verbal co-occurrents
"increase"	hausse, mouvement, progression, reprise	s'amplifier, s'accélérer, s'accentuer
"decrease"	baisse, recul, repli	ralentir, limiter, freiner

Table 2: Common verbal co-occurrents shared by terminological units (cf. [Heid 1994, quoted in Bertrand 1999])

Heid [1994] suggests that SLCs can be classified into two different categories: lexical collocations (groups in which the co-occurrent combines with a single terminological unit) and concep-

tual collocations (SLCs in which the co-occurrent combines with several terminological units). However, the author does not give an approximation of the importance of the phenomenon.

L'Homme [1998] conducted research on specialized lexical combinations and the study evolved into a model for the description of specialized verbs in the field of computing [2]. Verbs were selected if a special meaning could be identified within the field of computing (e.g. *install: install an operating system; run: run a program on a computer*) [3].

This study demonstrated that verbs combine with several terminological units that share semantic properties. In fact, all the verbs described (over 200 French verbs and approximately 100 English verbs) show this property. The examples in (3) illustrate this observation: *install* combines with *operating system, Windows, package, Word, surfer*, terms that denote a "piece of software".

(3)   *Once the operating system is installed, you can install the drivers.*
        Users install Windows 98 on their portable computers.
        This package cannot be installed on 80486 computers.
        You can install Word on your computer from this CD-ROM.
        This routine will assist you in installing your web surfer.

The research conducted by L'Homme shows that the regrouping of terminological units within semantic classes provides a basis for a description of verbal and deverbal co-occurrents and confirms the observation made by Martin [1992] cited above. The other co-occurrents (adjectives and others nouns) have not yet been described. Furthermore, the model takes for granted that co-occurrents combine with classes of terms: this property has not been explored with a large corpus.

In a study on general language collocations, Mel'čuk & Wanner [1996] argue that there appears to be a correlation between the meaning of a lexeme and its restricted co-occurrence; lexemes with common collocates share semantic features. The authors studied German collocations that comprise keywords that denote an emotion (e.g. *Achtung*: *respect; Hass*: *hatred*; *Mitleid*: *compassion*) in order to find out the extent to which this is true and if so, to develop a more efficient descriptive model to be implemented in dictionaries. The conclusion is as follows:

> The treatment of lexical data as outlined above shows that significant correlations between restricted lexical co-occurrence and semantic features exist, and they allow for reasonable generalizations. At the same time, the correlations are far from absolute: idiosyncrasies in collocations abound and simply have to be listed [Mel'čuk & Wanner 1996: 211].

Hence, according to these observations, the generalization of keywords within larger groups of semantic classes appears possible, but cannot be systematically applied in general language. A typical collocation is semi-compositional: the keyword will combine uniquely with a given co-occurrent, whose meaning is altered within that specific combination.

4 A corpus-based study

As was shown in the previous section, the generalization of specialized co-occurrents as parts of a series of terminological units pertaining to the same subject field seems highly productive. This is hardly surprising since, in a given field, many terms share common semantic properties. For instance, in the field of medicine, a terminologist will list the names of different diseases; it is very likely that these terms will share many different co-occurrents. On the other hand, the generalization cannot be applied as systematically to general language collocations.

Even if several authors believe that SLCs can be described in terms of semantic classes, this property has not been explored using different types of corpora. Heid [1994] observed the co-occurrents listed in a reference work. L'Homme [1998] has taken this property for granted when describing specialized verbs. The work described in this section addresses this issue.

Using texts related to two fields of knowledge, namely aeronautics and philosophy, we extracted French specialized lexical combinations in order to measure the extent to which semantic classes in SLCs could be observed. The subject fields were chosen in order to determine if differences could be found between technical texts and texts relating to the humanities. Using the distinction made by Heid [1994] quoted above, we studied the proportion of lexical collocations (in which a co-occurrent with a given meaning combines with a single term) and conceptual collocations (in which a co-occurrent with a given meaning selects groups of terms) in both subject fields.

We extracted approximately 6000 SLCs (verb + noun (term), adjective + noun (term), and noun + noun (term)) from specialized texts. As a starting point, we chose terminological units that pertain to different semantic classes. Table 3 shows the terms selected from each field of knowledge. Examples of SLCs extracted at this level are provided in Table 4.

Field	Terminological unit	Semantic class
Aeronautics	aéronef (aircraft)	flying machine
	aéroport (airport)	installation
	piste (runway)	circulation installation
	vitesse (speed)	measuring unit
	vol (flight)	activity
Philosophy	amour (love)	disposition
	beauté (beauty)	property
	connaissance (knowledge)	moral entity
	être (being)	abstract entity
	vérité (truth)	principle

Table 3: Selected terms and their semantic class

We then extracted the SLCs in which these terms appeared from running text. The co-occurrents found in these SLCs were then used to find new SLCs : we selected 15 different co-occurrents for each term (5 co-occurrents per grammatical category). At this level, co-occurrents were selected according to their frequency in the corpus. In addition, if two co-occurrents were synonymous, we excluded one. Finally, if a verb and its nominalization (e.g. *exploiter* and *exploitation*)

Selected term	Pattern of SLC	Co-occurrent	Frequency
Aéronef (aircraft)	V + T	autoriser (authorize)	29
Aéronef	V+ T	exploiter (exploit)	42
Aéronef	N + T	contrôle (control)	2
Amour (love)	V + T	unifier (unify)	2
Beauté (beauty)	T + V	donner (give)	7
Être (being)	V + T	saisir (grasp)	35

Table 4: SLCs with selected terms extracted from the corpus

were both extracted as potential co-occurrents of a given term, we would exclude one of them. Table 5 gives examples of the new combinations found during this phase.

Selected term	Co-occurrent	New terms	Frequency
Aéronef	décoller (take off)	aéronef (aircraft)	19
		avion (airplane)	7
		hydravion (seaplane)	3
		vol (flight)	2
Être	saisir (grasp)	être (being)	30
		esprit (mind)	12
		âme (soul)	10
		moi (self)	5

Table 5: Selection of new SLCs

Finally, we examined the combination in order to determine: 1) if the co-occurrents with a given meaning combined with several different terms; 2) if a co-occurrent combined with a series of different terms, did the terms share semantic features? Figure 1 shows the steps of the study.

We observed that, in 86 % of the SLCs studied, the co-occurrents could be found in other combinations. Moreover, all the terms found in these SLCs belonged to the same semantic class. In other words, 86 % of the combinations were conceptual collocations. Surprisingly, the proportion is the same in both fields of knowledge. The remaining combinations (14 %) were SLCs in which a single term was found for a given co-occurrent. However, it is likely that the proportion of true "lexical collocations" could have been reduced if the observations had been made on a larger corpus. Table (6) presents a summary of the results obtained for the corpus on aeronautics; Table (7) shows the results for the philosophy corpus.

Even though the generalization of keywords as groups of terms seems to be highly productive as shown by the results obtained by this study, it should be pointed out that it cannot be applied systematically. Laporte and L'Homme [1997] observed that some co-occurrents combine with a generic term (e.g. *dangereux* – *dangerous* – in *médicament dangereux* – *dangerous medicine*), but not with hyponyms (*dangereux* does not combine with *aspirine* – *aspirin* – or *réserpine* – *reserpine* – **aspirine dangereuse*, **réserpine dangereuse*).
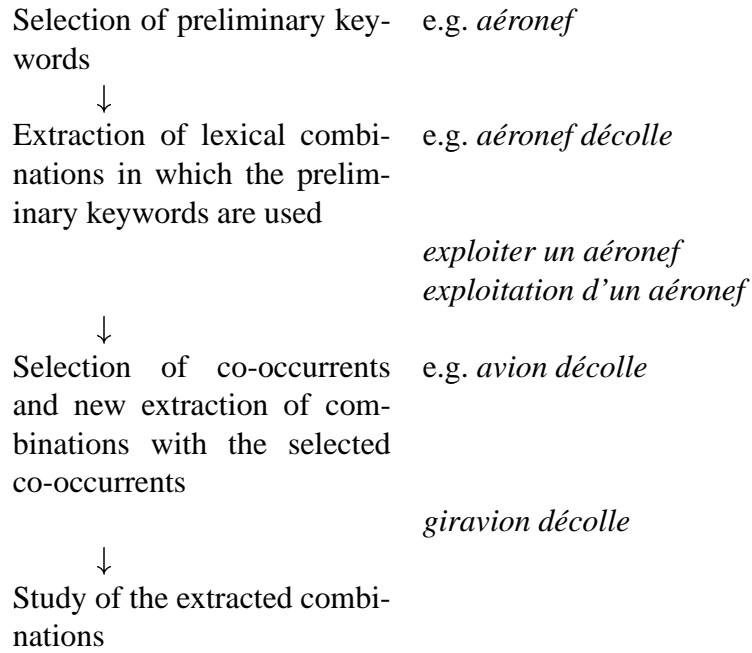
Selection of preliminary key-
words   e.g. *aéronef*

↓

Extraction of lexical combi-
nations in which the prelim-   e.g. *aéronef décolle*
inary keywords are used

exploiter un aéronef
exploitation d'un aéronef

↓

Selection of co-occurrents
and new extraction of com-   e.g. *avion décolle*
binations with the selected
co-occurrents

giravion décolle

↓

Study of the extracted combi-
nations

Figure 1: Steps of the study

Grammatical category of co-occurrent	% of specialized lexical combinations	Conceptual collocations	Lexical collocations
Verb	35 % (195)	88 % (170)	12 % (25)
Noun	26 % (145)	81 % (116)	19 % (29)
Adjective	39 % (220)	90 % (198)	10 % (12)

Table 6: Distribution of conceptual and lexical collocations in the aeronautics corpus

It is also worth underlining that, even though a co-occurrent can combine with several semantically-related terms, a specific term will be used more frequently in specialized texts (at least of technical nature). For example, in the aeronautics field, the verb *exploiter* (*operate*) can combine with the following terms: *avion* (*airplane*), *aéronef* (*aircraft*) and *giravion* (rotor-craft). However, specialists will use *aéronef* much more frequently in this combination. These observations are illustrated in Table (8). On the other hand, the lexical preferences are not easily noticeable in the philosophy corpus, whereas the co-occurrents can be found in a far broader spectrum of language, making the terminological lexicalization less evident.

5 Conclusion

The study described in section 4 confirms that conceptual collocations are highly productive in specialized languages. A given co-occurrent selects a group of terms which belong to a semantic class. Thus, specialized lexical combinations cannot be defined as true collocations which comprise two lexemes that combine to form a unique combination with a specific meaning. On the

Grammatical category of co-occurrent	% of specialized lexical combinations	Conceptual collocations	Lexical collocations
Verb	24 % (232)	88 % (204)	12 % (28)
Noun	28 % (280)	90 % (252)	10 % (28)
Adjective	48 % (450)	96 % (431)	4 % (19)

Table 7: Distribution of conceptual and lexical collocations in the philosophy corpus

co-occurrent	terminological units	frequency
exploiter (exploit)	aéronef (aircraft)	42
	avion (airplane)	2
	giravion (rotorcraft)	2

Table 8: Frequency of terms with the verb exploiter [Bertrand 1999]

other hand, SLCs are best described in terms of free lexical co-occurrence than restricted lexical co-occurrence if one admits that the "freedom" is limited to the boundaries of a specialized subject field.

Typical SLCs described in reference tools should take this property into account and define the selectional restrictions of co-occurrents rather than providing simple listings of co-occurrents. This feature is difficult to account for in paper dictionaries, but can be implemented in an elegant fashion in computerized reference tools.

However, it seems hazardous to over-generalize and attribute these properties to all collocations and all SCLs. It appears that it is possible to generalize about some collocations but not to do so systematically. Also, even though it appears that most SLCs can be accounted for in terms of selectional restrictions, some may still have to be listed as unique combinations in reference tools.

Notes

[1] Rien dans le sémantisme ou encore dans la syntaxe ne force ce choix [...] [Mel'čuk *et al.* 1995: 126]

[2] The descriptive model was later extended to verbal nominalizations (e.g. *installation, copy, formatting*) [L'Homme et Gemme 1997]

[3] The criteria on which the selection relies are developed in L'Homme [1998]

References

[1] ALONSO RAMOS, M. et S. Mantha (1996). "Description lexicographique des collocatifs dans un Dictionnaire explicatif et combinatoire : articles de dictionnaires autonomes?", Clas, A., P. Thoiron et H. Bon et H. Béjoint (éd.), *Lexicomatique et dictionnarique. Actes du colloque de Lyon 1995*, Beyrouth / Montréal : FMA / Aupelf-UREF, pp. 233-253.

[2] BÉJOINT, H. et Thoiron, P. (1992). "Macrostructure et microstructure dans un dictionnaire de collocations en langue de spécialité", *Terminologie et traduction* 2-3, pp. 513-522.

[3] BENSON, M., Benson, E. & Ilson, R. (1986). *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*, Amsterdam / Philadelphia, John Benjamins.

[4] BERGENHOLTZ, H. & S. Tarp (Eds.) (1995). *Manual of Specialized Lexicography*, Amsterdam / Philadephia : John Benjamins.

[5] BERTRAND, C. (1999). *Étude comparative des combinaisons lexicales spécialisées dans deux domaines de spécialité : collocations lexicales et collocations conceptuelles en aéronautique et en philosophie,* Montréal : Université de Montréal.

[6] COHEN, B. (1986). *Lexique de cooccurrents - Bourse et conjoncture économique*, Montréal, Linguatech.

[7] HEID, U. (1994). "On the Way Words Work Together - Topics in Lexical Combinatorics", Martin, W.*et al*. (Ed.), *Euralex '94 Proceedings*, Amsterdam, pp. 226-257.

[8] HEID, U. et Freibott, G. (1991) : "Collocations dans une base de données terminologique et lexicale", *Meta,* 36(1), pp. 77-91.

[9] LAPORTE, I. et M.C. L'Homme (1997). "Recensement et consignation des combinaisons lexicales en langue de spécialité : Exemple d'application dans le domaine de la pharmacologie cardiovasculaire, *Terminologies nouvelles* 16, pp. 95-101.

[10] L'HOMME, M.C. (1995). "Processing Word Combinations in Existing Termbanks", *Terminology*, 2(1), pp. 141-162.

[11] L'HOMME, M.C. (1997). "Organisation des classes conceptuelles pour l'accès informatisé aux combinaisons lexicales spécialisées verbe + terme", *Actes des deuxièmes rencontres Terminologie et intelligence artificielle,* 3-4 avril 1997, Université Toulouse-le-Mirail (Toulouse), pp. 161-174.

[12] L'HOMME, M.C. (1998). "Définition du statut du verbe en langue de spécialité et sa description lexicographique", *Cahiers de lexicologie* 73(2), pp. 61-84.

[13] L'HOMME, M.C. et R. Gemme (1997). "Modèle d'accès informatisé aux combinaisons lexicales spécialisées : verbe + nom(terme) et extension aux nom(déverbal) + préposition : nom(terme), Lapierre, L., I. Oore et H.R. Runte, *Mélanges de linguistique offerts à Rostislav Kocourek*, Université Dalhousie (Halifax, Canada) : Les Presses ALFA, pp. 89-103.

[14] MARTIN, W. (1992). "Remarks on Collocations in Sublanguages", *Terminologie et traduction*, n°s 2-3, pp. 157-164.

[15] MEL'ČUK, I. (1996). "Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon", Wanner, L. (Ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam /Phildelphia: John Benjamins, pp. 37-102.

[16] MEL'ČUK, I. *et al.* (1984) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques 1*, Montréal, Les Presses de l'Université de Montréal.

[17] MEL'ČUK, I. *et al.* (1988). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I1*, Montréal, Les Presses de l'Université de Montréal.

[18] MEL'ČUK, I. *et al.* (1992) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III1*, Montréal, Les Presses de l'Université de Montréal.

[19] MEL'ČUK, I., Clas, A. et Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve (Belgique), Duculot / Aupelf - UREF.

[20] MEL'ČUK, I. & Wanner, L. (1996). "Lexical Functions and Lexical Inheritance for Emotion Lex-emes in German", L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam / Philadelphia : John Benjamins, pp. 207-277.

[21] MEYER, I. & Mackintosh, K. (1994). "Phraseme Analysis and Concept Analysis in Exploring a Symbiotic Relationship in the Specialized Lexicon", Martin, W. *et al.* 1994. *Euralex '94 Proceedings*, Amsterdam, pp. 339-348.

[22] MEYER, I. & Mackintosh, K. (1996). "Refining the Terminographer's Concept Analysis Methods: How Can Phraseology Help ?", *Terminology* 3 (1), pp. 1-26.

[23] THOIRON, P. et Béjoint, H. (1989). "Pour un index évolutif et cumulatif de cooccurrents en langue techno-scientifique sectorielle", *Meta* 34(4), pp. 661-671.