Collocational Information for Terminological Purposes

Elsabé Taljard
Department of African Languages, University of Pretoria
e-mail: elsabe.taljard@up.ac.za

Abstract

Traditionally, collocational information was sparsely, if at all, provided in terminological tools. This is particularly true for LSP dictionaries in paper format. One possible reason for this state of affairs is that LSP dictionaries had in the past often been compiled by subject field experts, who did not deem this kind of information as useful from a terminological point of view. This paper critically considers the importance of collocational information in terminological tools such as terminological databases, term banks and LSP dictionaries. It is argued that collocational information can assist the user on both conceptual and usage-related level and that such information is therefore of critical importance. Secondly, two methodologies for the identification of collocates, viz. introspection and corpus-based methods are compared, and it is concluded that access to huge amounts of real-life language data, also for LSP purposes, leads to insights which surpass those arrived at through introspection.
Keywords: collocational information; corpus-based terminology; terminological database

1 Introduction

Traditionally, collocational information was sparsely, if at all, provided in terminological tools for LSP purposes. Even currently, it is the exception rather than the rule to find collocational information in paper LSP dictionaries. The reasons for this state of affairs are not clear; however, it can be speculated that subject field specialists who played a pivotal role in the compilation of LSP dictionaries, simply did not deem this kind of information necessary in such a dictionary. Secondly, space constraints also contributed to the neglect of collocational information. The major reason however, was probably the traditional view on terminology, i.e. that terms are context independent and that any kind of contextual information, which includes collocational information, is of lesser importance for LSP purposes. The notion that terms should be studied within context is one of the major principles on which the modern (corpus-based) approach to terminology is based and access to and use of corpus material has afforded terminologists and lexicographers the opportunity to access huge amounts of terminological data from which terminologically relevant data can be extracted semi-automatically by means of corpus-query tools. Furthermore, the advent of e-lexicography, also for LSP purposes, has opened up new possibilities, one of these being the notion that a single database can be the source of multiple terminological products or tools, e.g. a variety of LSP dictionaries, aimed at different users, with different functions and of various levels of complexity, and online term banks.

The aim of this paper is twofold: first, to critically consider the importance of the inclusion of collocational information in terminological databases, whether such databases are used as data sources for online term banks or as a data source for the compilation of (a variety of) either online or paper LSP dictionaries. Secondly, a comparison is made between introspective and thus to a certain extent intuitive identification of collocations of terms by subject field experts versus corpus-based

extraction of collocational information. For this purpose, a small case study was done, using academic vocabulary as a case in point.

2 On the Nature of Collocations

The concept of collocation is notoriously difficult to define, even though, or perhaps because, as Evert (2007) points out, it is based on a widely-shared intuition that certain words have the tendency to co-occur in natural language. A distinction is made between empirical and phraseological collocations, where the former is defined as "the recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages", cf. Smadja (1993), i.e. "empirical statements about the predictability of word combinations" (Evert, 2007). Prototypical examples are *salt* and *pepper*, and *day* and *night*. Phraseological collocations are defined as "semi-compositional and lexically determined word combinations" (Evert, 2007) and are also called multiword expressions, e.g. *grant a request* and *put in an appearance*. A range of subcategories is subsumed under the latter term, including completely opaque idioms to combinations which are subject to arbitrary lexical restrictions. It is argued that both types of collocations are relevant for terminological purposes, especially in view of the fact that a database can be the data source for a variety of terminological tools, which may have different functions, such as cognition, text production, text reception as well as translation. Empirical collocations are relevant on the conceptual level, since members of a collocational set may be conceptually related – the reason why the terms *proton*, *neutron* and *electron* are a collocational set, is precisely because these concepts are conceptually related – they are all subordinates appearing in a part-of relation to the concept *atom*. In any terminological tool which has cognition as a function, this information would be extremely useful. Phraseological collocations are relevant on a more pragmatic, usage-related level, especially for the functions of text production and translation. This issue is further discussed in the following paragraph.

3 Relevance of Collocational Information in a Terminological Database

If text production and translation are seen as potential functions of a terminological tool, provision of collocational information becomes imperative, especially when the text production takes place in a language other than the L1 (dominant language) of the potential user of the tool. As indicated by L'Homme (2006:186), collocations are often unpredictable combinations, even in specialized language, and should therefore be provided for in the database used as data source for the eventual terminological tool. Translation of collocations is potentially problematic, since they are often idiosyncratic and language specific.

Secondly, collocations are domain dependent, which implies that collocations of a word in general language with which a user may be familiar, will be different from those of a term in a specific subject field, and secondly, that collocations of polysemous terms in different subject fields would also differ. To illustrate this principle, two special purpose corpora were compiled, one on academic vocabulary containing 3.3 million tokens (LSP_1 corpus) and one on climate change (LSP_2 corpus), of 290 000 tokens. Collocations for the lemma *data* which showed up as a KeyWord in both corpora, were extracted from both these corpora by means of *SketchEngine* and the results compared to those obtained from a similar exercise, using the enTenTen (2013) corpus as an example of a language for general purpose (LGP) corpus. For the purpose of this exercise, function words were not taken into

consideration, since the emphasis here is on finding collocation candidates which are conceptually related to the search node. The statistical measure used to compute the strength of the collocational relationships is the Mutual Information Score (MI). The results are tabled below:

LGP corpus enTenTen Sample: 4300 KWIC lines	**LSP_1 corpus** Academic vocab 4319 KWIC lines	**LSP_2 corpus** Climate change 392 KWIC lines
in-memory	collection	supplementary
NoSQL	analysis	anomaly
warehousing	collected	sources
unstructured	qualitative	climatic
biometric	quantitative	U
raster	data	core
retrieves	collect	overlap
BLS	obtained	UAH
MDM	analysed	comparable
aggregated	Table	adjusted

Table 1: Top ten collocations for *data* arranged according to MI score.

As can be seen from Table 1, there is no overlap between the collocations for *data* in the LGP corpus and either of the two LSP corpora, neither is there between the collocations of the two LSP corpora. Differences in collocational behaviour are even more pronounced when data extracted from the WordSketches for the term *data* in the LGP corpus are compared to those in the two LSP corpora – only three instances of overlap present themselves (compare table 2):

LGP corpus	**LSP_1 corpus**	**LSP_2 corpus**
***data* as OBJECT OF**		
encrypt	standardize	transmit
store	extract	amplify
transmit	calibrate	be
collect	distribute	
benchmark	collect	
***data* as SUBJECT OF**		
suggest	afford	constrain
indicate	show	explain
reside	suggest	exclude
file	indicate	start
centre	provide	come
***data* as MODIFIER OF**		
Census	high-frequency	UAH
meta	IES	Tes
EXIF	SARS	Met
census	DCP	metereological
raw	usage	Office
empirical	survey	above

Table 2: Excerpts from WordSketches for *data* in three corpora.

In the third instance, collocations are assumed to be useful elements in the conceptualization of a

knowledge domain, as indicated by Fuertes-Olivera (2012). With the availability of software tools such as *GraphColl*, it now becomes possible to not only provide information on the conceptual relationships between collocates, but also to contextualize terms within a much larger collocation network. *GraphColl 1.0* is a free tool, downloadable from the website http://extremetomato.com/projects/graphcoll, designed in such a way as to be usable by both novice and advanced users for the building of collocational networks. As Brezina, McEnery and Wattam (2015: 141) argue, collocates of words do not occur in isolation, but "form part of a complex network of semantic relationships which ultimately reveals their meaning". The applicability of Brezina et al.'s model to terminology and the identification of collocation networks which reveal conceptual relationships between terms, has not yet been tested – such an undertaking would require a separate and detailed investigation, but for the purpose of this paper, a small experiment was carried out to test, albeit on a small scale, the operationalization of collocation networks via the *GraphColl* software. It must be emphasized that such a small scale investigation does not do justice to the complexity and possibilities offered by the software. Figure 2 below, for example, represents what is termed second-order collocates – the software offers the possibility of progressing beyond first order collocates to investigate connectivity between collocates at various levels of collocational relationships – in Figure 1, fourth-order collocates with *time* as initial node are represented. One of the main advantages of the *GraphColl* software, is the fact the collocational networks are visually represented, which makes for easy interpretation. Compare the following figure (Brezina et al, 2015:153):

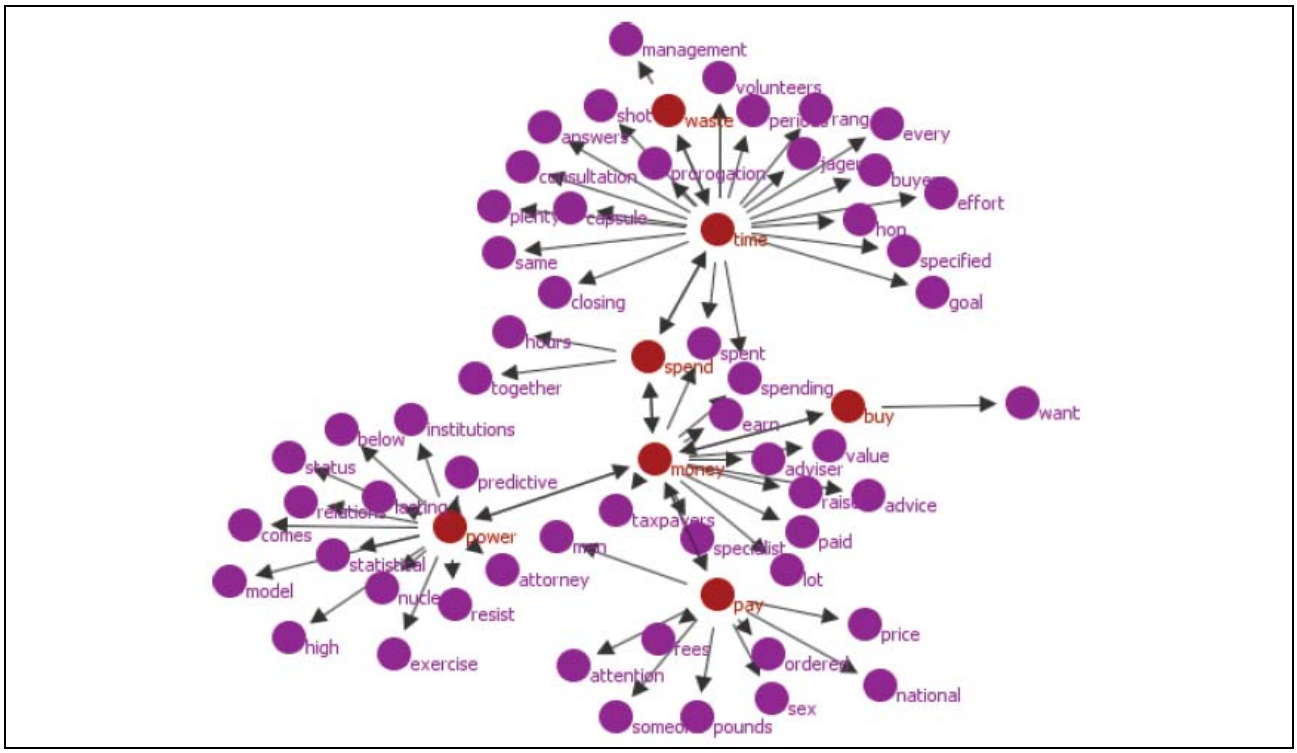

Figure 1: Fourth-order collocates for *time* as initial node.

For this experiment, the LSP_2 corpus on Climate change was utilized, the term *atmosphere* being used as a search node. From the figure below it can be deduced that a bi-directional relationship exists between *atmosphere* and *greenhouse*, in other words that in this corpus *atmosphere* co-occurs with *greenhouse* and that both concepts collocate with *carbon*. To what extent the graph gives an account of conceptual relationships is difficult to judge without the input of a subject field expert, but

it is a possibility which at least warrants further investigation.

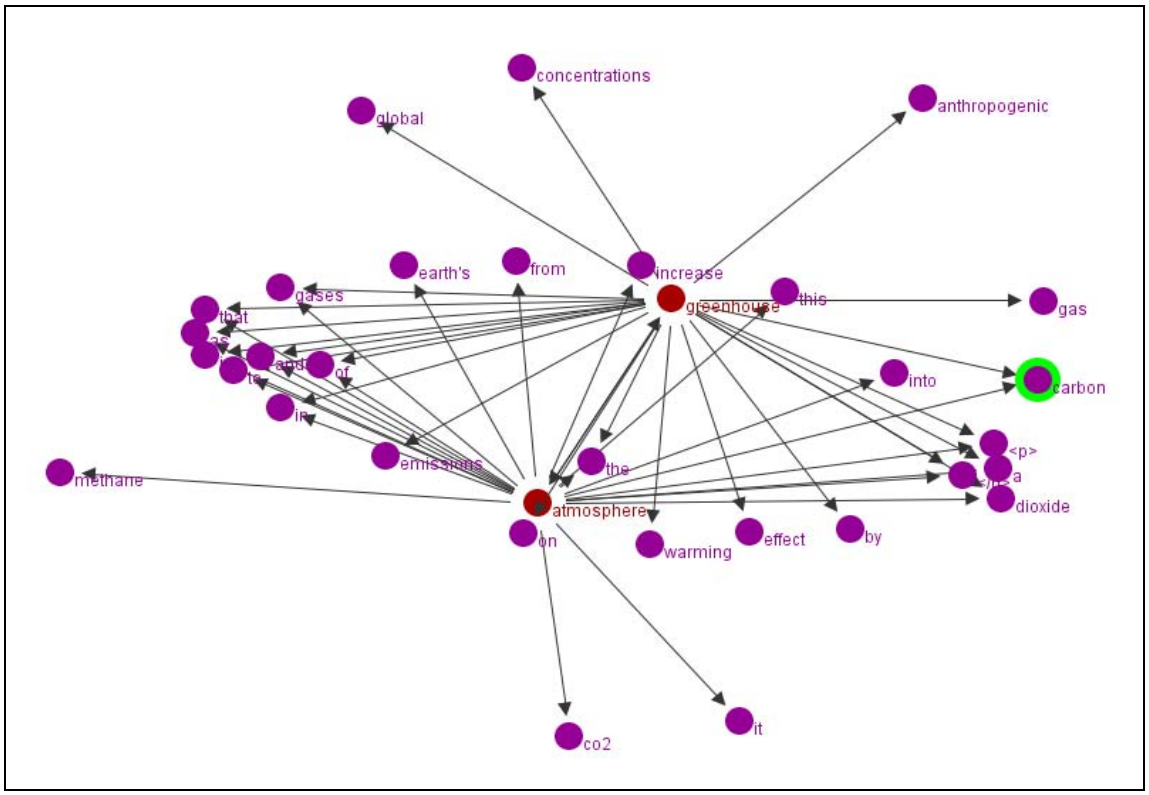

Figure 2: GraphColl for *atmosphere* in the LSP_2 corpus.

4 Introspection versus Corpus-based Identification of Collocational Information

Before the availability of electronic corpora and corpus query tools such as *WordSmith Tools* and *SketchEngine*, collocational information was mainly gleaned through introspective and intuitive means from subject field experts, bearing in mind that very little consideration was traditionally given to collocational information for terminological purposes. This practice was standard procedure in the so-called traditional approach to terminology, where terms were studied in isolation, without any thought being given to the context within which terms were used. In order to see to what extent introspective information correlates with corpus-based information and thus reflects authentic LSP language usage, a case study was done of collocations of a selection of so-called academic vocabulary. To this end, a list of sub technical academic terms, compiled by a team of English academics at a South African university was used as data source. The compilers of the list relied completely on their intuition and years of experience in academia in identifying term candidates. This list contains a total of 1 156 (mostly single word) items. Collocations were selectively supplied, apparently based on the judgement of the compilers as to whether such information was relevant. Ten items for which collocational information was supplied were randomly selected, and the collocates of these were compared to those generated computationally by means of *SketchEngine*, based on a 3.8 million word special corpus consisting of academic articles, dissertations and theses, covering a rage of academic disciplines. Compare Table 3:

		Introspective data	Corpus data		
	Term	Collocates	Raw freq	Freq/mil	Collocates
1	**abandon (V)**		121	29.53	
		(OBJECT)			(OBJECT)
		principle(s)	Does not meet minimum threshold		
		project	Does not meet minimum threshold		
			27	6.59	child
			3	0.73	newborn
			2	0.49	twin
2	**consent (N)**		918	224.08	
		(OBJECT OF)			(OBJECT OF)
		give	104	25.39	give
		(MODIFIED BY)			(MODIFIED BY)
		common	Does not meet minimum threshold		
		mutual	Does not meet minimum threshold		
		general	Does not meet minimum threshold		
			18.55	76	presumed
			191	46.62	informed
			44	10.74	written
3	**goal (N)**		854	208.46	
		(OBJECT OF)			(OBJECT OF)
		achieve	48	11.72	achieve
		(MODIFIED BY)			(MODIFIED BY)
		long-term	Does not meet minimum threshold		
		short-term	Does not meet minimum threshold		
			21	5.13	primary
			18	4.39	normative
4	**illustrate (V)**		674	164.52	
		(OBJECT)			(OBJECT)
			16	3.91	distribution
		point	8	1.95	point
			7	1.71	interaction
5	**issue (N)**		1881	459.15	
		(OBJECT OF)			(OBJECT OF)
			88	21.48	address
		raise	30	7.32	raise
		make (an issue) of	Does not meet minimum threshold		
6	**mobile (ADJ)**		482	117.66	
		(MODIFIES)			(MODIFIES)
		workforce	Does not meet minimum threshold		
		library	Does not meet minimum threshold		
			142	34.66	device
			119	29.05	website
			71	17.33	equipment
7	**null (N)**		154		

		(and) void	3	0.73	
					(MODIFIES)
			111	27.1	hypothesis
8	**option (N)**		420	102.52	
		(MODIFIED BY)			(MODIFIED BY)
		first	Does not meet minimum threshold		
		easy	Does not meet minimum threshold		
		soft	Does not meet minimum threshold		
			39	9.52	care
			24	5.86	alternative
			8	1.95	viable
9	**trend (N)**		428	104.47	
		(OBJECT OF)			(OBJECT OF)
		set	Does not meet minimum threshold		
			10	2.44	observe
10	**valid (ADJ)**		266	64.93	
		(MODIFIES)			(MODIFIES)
			16	3.91	marriage
			15	3.66	consent
		reason	14	3.42	reason
		argument	7	1.71	argument

Table 3: Introspective vs corpus-based collocates for selected academic terms.

As can be seen in Table 3, for four of the ten items, i.e. *abandon, mobile, option* and *trend*, there is no correspondence between the collocations identified by means of introspection, and those which are corpus-based. For the remaining six, various degrees of correspondence can be identified. It is significant that in the case of *abandon, consent, goal, issue, mobile*, *option* and *trend* some of the introspectively based collocates do not even meet the minimum threshold in order to be recognized as a collocate of a specific item. For *illustrate*, *issue* and *valid* a correspondence between the introspectively based and corpus-based collocations are noted; however, collocates which occur more frequently in the corpus were overlooked during the process of introspection. These results once more confirm the truism that access to large amounts of real-life language usage in the form of corpus data provides insights which are not possible through introspective analysis. With LSP corpora becoming more accessible and available, it therefore makes sense to complement introspective knowledge on collocations with corpus-based information, also for terminological purposes.

5 Conclusion

In this paper, the necessity of access to collocational information in terminological tools has been motivated. Collocational information in terminology can assist the potential user on both conceptual and usage-related level, therefore provision of collocational information for terminological purposes can no longer afford to be neglected. The advantages of access to LSP corpora, new developments in e-lexicography and the availability of sophisticated software make this a feasible undertaking.

6 References

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139-173.

Evert, S. (2007). *Corpora and collocations*. Accessed at http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf. [09/11/2015].

Fuertes-Olivera, P. (2012). Lexicography and the Internet as a (Re-)source. *Lexicographica* 28(1): 49-70.

L'Homme, M.-C. (2006). The Processing of Terms in Dictionaries: New Models and Techniques. A State of the Art. *Terminology* 12(2): 181-188.

Smadja, F. (1993). Retrieving Collocations from Texts: Xtract. *Computational Linguistics* 19(1): 143-177.