Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 140 – 148

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Context-dependent variation of LSP collocations: A corpus-based analysis

Laura Giacomini*

Department of Translation and Interpretation, University of Heidelberg, Plöck 57a, Heidelberg 69115, Germany

Abstract

This paper concerns a corpus-based study on LSP collocational variation, a phenomenon that is largely underrepresented in bilingual LSP lexicographic resources. Terminological variation on the formal (i.e. paradigmatic/syntagmatic) and semantic level is analysed in texts belonging to the technical domain (building and energy subfield). With the help of a balanced monolingual LSP corpus, variation is observed as a context-dependent phenomenon, with context interpreted as the combination of topic, genre-specific and communicative functions related to the LSP vertical dimension. A model entry from an English-Italian technical dictionary shows how the treatment of LSP collocations as autonomous terms and a corpus-based description of their variational behaviour can provide users with a means for efficiently tackling text production tasks, both in the native and in the foreign language.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: collocations; variation; LSP; LSP corpus; e-lexicography; technical domain.

1. Introduction

The topic of variation in specialised language has been analysed, both from a lexicological and a lexicographical standpoint, in a number of recent studies that have mainly focused on cross-domain variation of single terminological items and terminological word combinations. This paper, which presents information collected during

* Corresponding author. Tel.: +49-6221-547244.
E-mail address: laura.giacomini@iued.uni-heidelberg.de

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.430

specialised translation tasks at the Department of Translation of Heidelberg University and parallel research, concentrates on the topic of variation inside the same (sub)domain and, from this perspective, attempts to capture and describe recurrent patterns in collocational behaviour. Whereas the introductory section of the paper illustrates the phenomenon of variation in LSP and suggests a specific description model for coexisting collocational variants, section 2 focuses on its connection to the topics of concept-term-relations and standardization. Section 3 deals with variation typology and exemplifies the main variational patterns and subpatterns in collocation. These findings are accounted for in a model entry for a bilingual LSP dictionary covering the technical domain, which is illustrated in section 4. The final section, section 5, makes some concluding remarks on the topic and suggests possible developments for future research.

Terminological items are affected by variation in several ways, e.g. when different levels of specialization are expressed (cf. *pulmonologist* vs. *pulmonary specialist*) or when borrowings with and without adaptation coexist in the same domain (cf. in the medical language *ECG* vs. *EKG*, *Polkissen* vs. *polar cushion cells*). Variation also involves units above the word, such as terminological collocations, which have so far been analysed in particular from the point of view of a comparison between LGP and LSP (Kerremans, 2010; Gledhill, 2000). In what follows, we will distinguish two types of collocational variation:

- Collocational variation on the formal level
  Different combinations with comparable meaning: This type of variational model includes the subtypes 'paradigmatic', with variation taking place on the lexical level (e.g. in the Italian IT language *disco rigido / disco fisso* for *hard disk*), and 'syntagmatic', in which variation affects the syntactic structure of a word combination, producing a set of semantically equivalent terminological items (e.g. *brain disorders*, *brain-related disorders*, *disorders of the brain*).

- Collocational variation on the semantic level
  One combination with different meanings: Given the consideration of the same (sub)domain, these meanings are usually quite close to each other.

2. The concept-oriented approach and the role of standardisation

In order to ensure conceptual coherency in variational description, variation has been looked at from the point of view of the relation that can be traced between concepts and designations (cf. Roelcke, 2013; Magris et al., 2002). The reference work is the E DIN 2342:2004-09 norm published by the German Institute for Standardisation under the name "Begriffe der Terminologielehre" (Concepts of terminology), in which *Begriffe* (concepts) and *Bezeichnungen* (designations) are paired as the central elements of terminology. Designations can take the form of terms, symbols, formulas or proper nouns. In light of the present study, it needs to be said that terms are to be intended not only as single terminological items but also as terminological multiword expressions or collocations.

This study also takes into consideration recent developments in terminography, with a concept-oriented approach being increasingly adopted in the implementation of monolingual and multilingual resources such as terminology databases of computer-assisted translation tools. According to the German terminology association (DTT), terminology management should adhere to the following 5 principles (cf. DTT's *Terminologiearbeit: Best Practices* 2014): 1) conceptual orientation; 2) elementariness (different types of information should not be assigned to the same attribute); 3) granularity (attributes should be as specific as possible); 4) designation autonomy (synonyms and abbreviations of a term should be treated as autonomous items); and 5) restricted number of mandatory fields. Although these principles are intended to provide guidelines for the creation of termbases, they could also be fruitfully applied to the representation of collocations in an LSP e-lexicographic tool, as the proposal for a lexicograhic entry in section 4 will show.

The classification of collocational variation proposed in section 1 runs counter to the idea of univocal relations among concepts and terms (among others, the idea of terminological monosemy) because it implies that a concept

may be expressed by more than one term, and that a term may refer to more than one concept. The lack of univocity can be observed at varying degrees depending on specific factors such as language, domain, or topic. Generally speaking, however, it is closely connected to the level of standardization of an LSP, i.e. the extent to which the terminology of a specialised field has been subject to rules that have as their purpose the conformity of that terminology with a standard (Quirion, 2014; Roelcke, 2010). Less standardized, i.e. newer and promptly changing domains such as the technical domain, display a terminology with a higher degree of variation and instability. These domains may allow for several terms referring to the same concept and capable of being classified along manifold criteria, as the following, self-evident example from IT language demonstrates:

- diachronic variants (German *Rechner* vs. *Computer*)
- commercial variants / company-internal terminology (*add-ons* [Firefox, IE] vs. *extensions* [Safari, Chrome])
- graphical variants (*online*, *on line*, *on-line*)
- synchronic borrowings (German synonymic calque *herunterfahren* vs. homonymic calque *downloaden*).

In this study, the terminology of the technical domain is at the centre of collocational variation analysis exactly because of its general resistance towards normativity and its great adaptability to the continuous sector's developments.

3. Variation in the technical domain: context-specific collocations

The study concentrates on the technical subdomain at the intersection between the building and the energy sector, with topics ranging from energy-efficient buildings, components and materials to insulation technologies and heating systems. Collocational analysis has been carried out on an Italian LSP corpus of around 15m words providing a sample of subdomain-specific textual genres: handbooks, technical data sheets, specialised magazines, product catalogues and marketing texts.

Context, which will serve as a key criterion in establishing variants distribution, will be defined as the combination of

- topic
- textual genre
- position in the vertical dimension

For the purpose of this study, the vertical dimension (cf. Kalverkämper and Baumann, 1996; Scarpa, 2008) has been segmented into three broad categories of texts, depending on the communicative relation intervening between author and recipient: scientific texts, didactic texts and popular scientific texts. This broad categorization intentionally does not take into account more fine-grained text typologies and avoids, for the moment, to connect the vertical status of a text with related textual genres: in fact, it is not meant to be exhaustive but only to provide a general framework against which the lexicographic treatment of vertical elements can be tested.

Collocations extracted from the Italian corpus have been contrasted with their English equivalents from lexicographic resources and parallel texts. Technical dictionaries, glossaries and databases (e.g. *Marolli Grande Dizionario Tecnico Inglese 2014* and the online IATE Building and public works sections) have also been consulted to evaluate the treatment of the Italian variants. Unfortunately, many variants are not recorded in these resources and contextual differences are not marked within the same subdomain, even if they are clearly relevant for variants distinction. This will be exemplified in the last section of the paper.

The Italian collocations have been grouped in two sets, matching the previously introduced groupings in *formal variation* (A) and *semantic variation* (B):

A – Formal variation
one concept : many terminological collocations

CONCEPT	TERMINOLOGICAL COLLOCATION
HEAT LOSS	*dispersione termica*
	dispersione di calore

B – Semantic variation
one terminological collocation : many concepts

CONCEPT	TERMINOLOGICAL COLLOCATION
WOOD SHEATHING	rivestimento in legno
WOOD PANELING	

3.1. Formal variation

The following tables collect the extracted Italian collocations according to their variation typology. Table 1 contains collocations of the class N+A or N+PP with their paradigmatic variants; the underlying grey shades indicate different subtypologies of collocational variation, which may occur at the level of N (light grey), A/PP (middle grey) or may involve more than one collocational component (dark grey). Variants of collocations in the form of single terminological items (cf. *coibentazione*) or compounds (cf. *termoconduttore*) have in this representation the same status as collocations because, independent of their morphology, they can be ascribed to the same concept.

The columns on the right offer information on the frequency and distribution of a given collocation in the corpus. +F signals the highest relative frequency of a collocation among the extracted variants, whereas the abbreviations sc (scientific text), dd (didactic text) and psc (popular scientific text) indicate in which textual category in the vertical dimension a given collocation was to be found. An empty field in the last column (e.g. *conduttore termico*) stands for a similar distribution at all levels of the vertical dimension. As anticipated in the previous section of the paper, even in the absence of an explicit association with textual genres this kind of distributional description discloses valuable information for correctly embedding collocational variants in texts according to current usage of terminology.

Table 1. Formal variation of collocations: paradigmatic variants

CONCEPT	**TL COLLOCATION (ITALIAN)**	**frequency**	**distribution**
THERMAL EFFICIENCY	**rendimento** termico	+ F	
	efficienza termica		sc / dd
THERMAL CONDUCTIVITY **k, λ**	**conducibilità** termica	+F	
	conduttività termica		sc / dd
HEAT LOSS	dispersione **termica**		sc / dd
	dispersione **di calore**	+F	psc
THERMAL CONDUCTOR	conduttore **termico**	+ F	

	termoconduttore		sc
THERMAL INSULATION	**isolamento** termico	+F	psc
	coibentazione		
SOUNDPROOFING	**isolamento acustico**	+ F	
	insonorizzazione		
	protezione contro i rumori		psc
ELECTRIC POWER CONSUMPTION **KWh**	consumo di **corrente elettrica**	+ F	psc
	consumo di **energia elettrica**		psc

CONCEPT	**TL COLLOCATION (ITALIAN)**	**frequency**	**distribution**
THERMAL INSULATION	**isolamento** termico	+F	psc
	coibentazione		
THERMAL EFFICIENCY	**rendimento** termico	+ F	
	efficienza termica		sc / dd
THERMAL CONDUCTIVITY **k, λ**	**conducibilità** termica	+F	
	conduttività termica		sc / dd
HEAT LOSS	dispersione **termica**		sc / dd
	dispersione **di calore**	+F	psc
THERMAL CONDUCTOR	conduttore **termico**	+ F	
	termoconduttore		sc
SOUNDPROOFING	**isolamento acustico**	+ F	
	insonorizzazione		
	protezione contro i rumori		psc
ELECTRIC POWER CONSUMPTION **KWh**	consumo di **corrente elettrica**	+ F	psc
	consumo di **energia elettrica**		psc

None of the presented collocational sets is meant to exhaustively cover the number of possible verbalisations of a given concept, or to account for all possible equivalents in the selected language pair. Distributional patterns of formal variants cannot be justified by frequency evidence alone and, for this reason, only a relative frequency indication has been recorded in Table 1 to provide a comparison between collocations which tend to appear in the same position in the vertical dimension. Distributional patterns can be explained by taking into consideration the role of contextual factors (i.e. topic, genre, relation between message sender and recipient) and external factors such as a company's terminological preferences. Moreover, symbols and measurement units, for instance in the case of

the concepts of thermal conductivity and electric power consumption in Table 1, can be of help in assessing terminological equivalence of two or more variants.

Table 2 displays examples of formal variation on the syntagmatic level, to which the same distributional analysis as in Table 1 has been applied. In this case, variation primarily concerns prepositions in N+PP combinations. Different from other technical subdomains, the absence of the preposition introducing a prepositional phrase, e.g. *impianto aerazione* for *impianto di aerazione* (*ventilation system*), is generally infrequent and seems to occur only in specific textual genres such as data sheets (for instance in tables or similarly condensed texts). This kind of syntagmatic structure lacks systematicity and can be considered a marginal phenomenon (cf. the opposite case of IT language, with systematic patterns like *trasferimento dati* / *data transfer*, *archiviazione dati* / *data storage*, etc.).

Table 2. Formal variation of collocations: syntagmatic variants

TL COLLOCATION (ITALIAN)	**frequency**	**distribution**
rivestimento **in** legno	+F	
rivestimento **di** legno		dd / psc
protezione **contro i** rumori		psc
protezione **dai** rumori		
protezione **dal** rumore	+ F	
condizionatore **a** parete		
condizionatore **da** parete	+F	dd / psc

From a bilingual perspective, formal variation coincides with different kinds of equivalence, as the following examples concerning the language pair Italian-English show:

a) IT dispersione termica / ~ di calore : EN heat loss
b) IT inerzia acustica : EN acoustic inertance / ~ mass
c) IT isolamento termico / coibentazione : EN thermal insulation / heat ~
d) IT protezione dal rumore / ~ contro il rumore : EN sound protection / noise ~

Examples a) and b) are of a many:1 equivalence type, examples c) and d) imply a many:many equivalence. Moreover, a-c) are examples of paradigmatic variation, whereas d) displays syntagmatic variation in the Italian collocations.

3.2. Semantic variation

Table 3 applies to semantic variation of collocations, with one terminological collocation matching more than one concept. Due to the overall cross-genre relevance of these combinations, distributional evidence in terms of vertical positioning has turned out to be redundant and has not been integrated in the table. The indication concerning the relative frequency of a collocation in the corpus, on the contrary, is shown. The symbol = indicates that a collocation can be attributed with a comparable frequency to all its possible meanings.

Table 3. Semantic variation of collocations

TL COLLOCATION (ITALIAN)	**CONCEPT**	**frequency**

impianto **termico**	HEAT / COLD	+F
	HEAT	
isolamento termico	RESULT OF INSULATING	+F
	PROCESS OF INSULATING	
rivestimento in legno	SHEATHING	=
	PANELING	=
energia elettrica	ELECTRIC CURRENT	=
	ELECTRIC ENERGY	=

The column with the concepts contains the part(s) of the concept which is subject to variation and which is verbalised by the component of the collocation appearing in bold characters, e.g. in the collocation *impianto termico* (N+A), the adjective *termico* may be referred either to heat and cold at the same time, or to heat only.

Motivation for semantic variation can sometimes be found in the morphology-related semantics of a collocation's component, for instance in the case of *isolamento* and *rivestimento*, with the ending -mento referring either to a process or to its result. Generally speaking, it can be found in the polysemy of a collocation's component. Also the relevance of a certain reading depends on the context, this time not intended as a specific genre and/or vertical status but as the specific text topic.

4. A proposal for variational encoding in an LSP lexicographic entry EN-IT

The corpus findings, their analysis and categorization according to the categories described in sections 3.1 and 3.2 have been incorporated into a model entry of an LSP e-dictionary for the language pair English-Italian. The entry is meant to be hosted in the section of the dictionary with Italian as the target language and to serve active translation purposes for native speakers of English. The proposal concerns the collocation *thermal conductivity*, which has been first of all looked for in the available lexicographic resources. *Marolli* (2014) records *thermal conductivity* only under the lemma conductivity and provides only one Italian equivalent, *conduttività termica*. The IATE database records the collocation as an independent entry, provides the variant *heat conductivity* and two Italian equivalents, *conducibilità termica* and *conduttività termica*. *Conducibilità termica* is marked as "preferred". This is indeed a much better result in terms of coverage of collocational variation in the two languages. However, a non-native speaker of the target language obtains no information concerning the distribution of the two equivalents and the motivation for the suggested preference.

Table 4 shows the microstructural items of an integral LSP lexicographic entry. The entry consists of three distinct layers, the purpose of which is to account for conceptual, cross-linguistic and language-specific features of a collocation and its variants, also paying attention to their contextual distribution.

Table 4. Microstructure of an LSP lexicographic entry

LAYER	ITEM	ADDITIONAL INFORMATION
1) CONCEPT LAYER	terminological concept	cross-reference to related concepts
	symbol/formula	
2) DEFINITION LAYER	definition (L_1 and/or L_1)	
3) DESIGNATION LAYER	designation L_1	for each designation:
	(designations ordered by rel. frequency)	pragmatic labels *
	other designation(s) L_1 (variants)	corpus examples

	reciprocals **
	collocations **
designation L_2	for each designation:
other designation(s) L_2 (variants)	pragmatic labels *
	corpus examples
	reciprocals **
	collocations **

* level of specialization, company-internal preferences

** cross-reference to other entry

The type and number of items actually shown during consultation (for instance, definitions in which language) depends on the function for which the lexicographic resource is being used at a specific moment. The following scheme displays a model entry intended for active translation into Italian, with a user looking up the English collocation *thermal conductivity*. For demonstration purposes, the three layers have also been numbered in the model entry.

1) THERMAL CONDUCTIVITY
λ
k

2) DEF: The rate of heat flow, under steady conditions, through unit area, per unit temperature gradient in the direction perpendicular to the area.
UNITS (SI): W/(m K)

3) EN: **thermal conductivity**
heat conductivity

IT: **conducibilità termica** (sc dd psc)

EX:
Il poliuretano espanso rigido è il materiale isolante che, a parità di spessore, garantisce le migliori prestazioni. Il suo valore di **conducibilità termica** stabile nel tempo (lD) è compreso tra 0,024 e 0,028 W/mK in funzione del tipo di schiuma e del tipo di rivestimento. (www.poliuretano.it)

Il coefficiente di **conducibilità termica** del vetro è molto elevato (λ = 1 kcal/h · m · ° C), è quindi necessario aumentare la resistenza termica fornita dal tamponamento utilizzando vetri doppi con intramezzata una camera d'aria… (Mottura/Pennisi 2014)

COLL:
valore di ~
coefficiente di ~

conduttività termica (sc dd)

EX:
Schede tecniche. Dop - Isover EPS ... **Conduttività termica** a 10°C, λD W/(m·K), λD W/(m·K), λD W/(m·K). 0,038, 0,036 ... Dimensioni, EPS 038, EPS 036, EPS 035. (isover.it/pannelli-isolanti-e-feltri/isover-eps)

Analisi della **conduttività termica** di progetto dei materiali da costruzione (Muscio 2012)

REC: resistività termica

Underlined text refers to hyperlinked data, such as cross-references to autonomous entries of reciprocals and collocations, or corpus evidence for the pragmatic labels regarding the level of specialisation (sc, dd, psc). Hyperlinked corpus evidence is crucial to enable the user to find suitable usage contexts by filtering first the desired level of specialisation. However, a small set of corpus occurrences (2 in the presented model entry) should be provided by default after each designation variant. In order not only to account for variation, but also to clearly

attribute variants to the correct concepts, it is necessary to link the concept of thermal conductivity to related concepts (i.e. thermal conduction, thermal conductance C, thermal transmittance U, thermal resistance R and thermal resistivity), possibly with the help of an interactive graph.

5. Concluding remarks

The study demonstrates how variational models in collocations play a central role in LSP textual practices and shows that existing lexicographic resources mostly fail to account for this phenomenon. It is, of course, necessary to stress the importance of corpus analysis focusing on domain-related context variation, which can contribute significantly towards detecting systematic vs. non-systematic aspects of collocational variation. The aim of future work in the field of collocational variation in LSP should be to find suitable descriptive models to support users during active translation and, in general, text production. The relevance of context-based analysis of collocational variation in LSP also needs to be pointed out as a key aspect, especially by considering context as the combination of several textual factors, namely topic, textual genre and the status of a text in the vertical hierarchy. Moreover, as exemplified in section 4, LSP collocations should be treated as autonomous terminological entities and linked to a coherent conceptual system.

The level of standardisation of a specific domain is certainly another crucial criterion on which to base any lexicographic and terminographic description (Arntz, Picht, and Mayer, 2002). In particular, an in-depth investigation should be carried out into the terminology of relevant technical subdomains (e.g. IT, building, electrical and mechanical engineering) which are involved in a large range of translational activities as well as in text production, and in which, from the point of view of variational encoding, the lack of adequate, up-to-date reference resources is particularly evident.

References

Arntz, R., Picht, H., and Mayer, F. (2002). *Einführung in die Terminologiearbeit* (Studien zu Sprache und Technik vol. 2). Hildesheim.

DTT Deutscher Terminologie-Tag e.V. (2014). *Terminologie. Best Practices*. Cologne.

Kalverkämper, H., and Baumann, K.-D. (1996). Fachliche Textsorten. Komponenten – Relationen – Strategien. *Forum für Fachsprachenforschung 25*. Tübingen.

Kerremans, K. (2010). A Comparative Study of Terminological Variation in Specialised Translation. In C. Heine, and J. Engberg (Eds.), *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009*. Aarhus.

Quirion, J. (2014). La mesure de la variation terminologique comme indice de l'évolution des connaissances dans un environnement bilingue. In R. Temmerman, and M. Van Campenhoudt (Eds.), *Dynamics and Terminology. An Interdisciplinary Perspective on Monolingual and Multilingual Culture-bound Communication* (pp. 281-302). Amsterdam/Philadelphia.

Magris, M. et al. (2002). *Manuale di terminologia*. Milano.

Marolli, G. (2014). *Grande Dizionario Tecnico Inglese*. Milano.

Roelcke, T. (2010). *Fachsprachen*. Berlin.

Roelcke, T. (2013). *Definitionen und Termini. Quantitative Studien zur Konstituierung von Fachwortschatz*. Berlin/Boston.

Scarpa, F. (2008). *La traduzione specializzata*. Milano.