CHAPTER NINE

A MULTIPERSPECTIVE APPROACH TO SPECIALIZED PHRASEOLOGY: INTERNET AS A REFERENCE CORPUS FOR PHRASEOLOGY

GUADALUPE AGUADO DE CEA

0. Abstract

Following the growing interest in Language and the Internet, this article tries to give a comprehensive overview of specialized phraseology and its multifaceted approaches in relation with the Internet. Starting from its pragmatic nature in specialized languages and analysing the importance of these lexical combinations from the descriptive approaches to terminology, this paper describes the rich and varied landscape that specialized phraseology offers to LSP practitioners, terminologists, technical translators as well as communication mediators. These professionals require mastering not only the domain terminology but the contextualised use of specialized verbal collocations to produce technically-stylistic natural texts. However, specialized dictionaries do not often include this type of verbal phraseology that can provide the reader with linguistic knowledge as well as domain knowledge, nor do they include many of the metaphorical uses. So technical translators often take up the Internet as a reference corpus, with the advantages and disadvantages it may have as referring to lexical reliability and language use for translation purposes. Finally, some of the approaches to phraseology from the computational point of view are presented to illustrate and highlight the importance of these lexical combinations in natural language processing studies and applications

1. Introduction

In the last twenty years, scientific interdisciplinary research has evolved more rapidly than ever due to information and communication technologies (ICT) and the meteoric development of the Internet. This fact has given rise to

new communicative situations, new cultural models, new varieties of language and discourse, a new field of research, Netlinguistics (Posteguillo 2003), and a flurry of new terminology and phraseology. This phraseology, originated in English as the 'lingua franca' of science and technology, is constantly growing and its presence is pervasive in many professional texts. Mastering it can become a rather difficult issue for different professionals: technical translators, terminologists, ESP teachers, communication mediators, engineers and scientists who need it for several purposes. Furthermore, these lexical combinations have acquired a relevant role in other fields such as natural language processing applications since they do not only provide knowledge on the domain but also help extract conceptual relations about the specific terms of a discipline. Thus, depending on the purpose and the needs all these professionals have, phraseology can be seen as having different dimensions that can be approached transdisciplinarily.

In order to frame the scenario of what follows, I resort to the notion of languages dealing with specific areas of human activities. The object of study is the same i.e., the language used in the different fields of specialized activities and professions or 'special subject languages' (LSP), as Sager *et al.* (1980: 38) call them, but different viewpoints can be adopted. In teaching English as a foreign language it is possible to distinguish between English for General Purposes (EGP) and English for Special Purposes (ESP), which appears as a convenient division for designing syllabuses and course outlines that suit students' needs. Therefore, when referring to English for Specific Purposes (ESP) the focus is on the **teaching of language** to specialists, mainly non-native, who need it for some occupational, professional, or academic purposes, whereas when emphasizing the **language as a vehicle for the transmission of specialized knowledge** other aspects are highlighted and other terms are preferred: *langue de spécialité, technolekt, lenguaje profesional o lenguaje académico (*Alcaraz 2000).

In this latter view, special or specialized languages can be characterised by the following pragmatic features: (a) the subject field, (b) the users participating in the communicative act: expert to expert, expert to semi-expert, semi-expert-to semi-expert, expert to layman, semi-expert to layman, and teacher to pupil; (c) the situation or the context in which the communicative act takes place; (d) the purpose of the communicative function; (e) the discourse features that may vary depending on the different scientific and technical genres; and (f) the presence of specific terminology and phraseology.

The aim of this chapter is to give an overview of different approaches to phraseology in specialized texts. The examples in this article are based mostly on work carried out on an English-Spanish corpus on computing and the Internet. Following this introduction, I have tried to delimit the concept of

specialized phraseology. The next sections, 3 to 5, are devoted to the different perspectives of this topic according to the specific needs of study: technical translators and terminologists, scientific writers and other professionals in academic environments that deal with specialized languages in their work. The emphasis on phraseology by these professionals can shift from a more terminological angle to a more academic one. However, all of them require expertise not only in the domain terminology but also in the contextualised use of specialized collocations to produce technically-stylistic natural texts. Notwithstanding, ICT dictionaries do not often include verbal phraseology as lexicographic works have traditionally focused on nouns as being the main components of terminology. Thus, all these professionals often take up the Internet as a reference corpus in search of the right collocation, with the advantages and disadvantages it may have as referring to lexical reliability and language use for translation and production purposes. In section 6, I would like to mention why collocations are of great interest in several natural language processing applications. Finally, the last section describes the possibilities that Internet as a reference corpus for phraseology offers to all these different LSP practitioners.

2. Defining specialized phraseology

Phraseology is not easily defined as it combines grammar, lexis and semantics. Broadly speaking, the term 'phraseology' can be understood in two ways: as the linguistic discipline that deals with the combination of words, and in this sense it is comparable with lexicology as the science of lexis, and as the set of phraseological units or phrasemes. These phraseological units are formed by two or more words that tend to co-occur together, i.e. they collocate. This idea of 'collocation' represents a wide framework in which many different combinatory units can be included.

According to the Russian tradition of phraseology (Cowie 1998:4), they form a *continuum* or a cline with free combinations ('open collocations') at one end and restricted combinations at the other pole. A 'free word combination' can be described using general rules; that is, in terms of semantic constraints on the words that appear in a certain syntactic relation with a given headword. The other extreme can be represented by 'idioms'. Somewhere between these two poles are phraseological units. In this view, 'fixedness' is an important issue. But 'fixedness' is not the sole criterion that can help differentiate them.

For the British contextualism (Sinclair 1966, Halliday 1966), 'frequency' became a crucial feature to characterize a 'collocation', formed by a 'node' or 'base' and a 'collocate', i.e. the 'node' prefers the company of a 'collocate' rather than that of other lexemes that can be almost synonymous. The

availability of large machine-readable natural language corpora helped to describe the collocational meaning of these lexical units in terms of statistical results. These two basic axes, 'fixedness' and 'frequency', along with 'acceptability', that is, the reproducibility of collocations with some semantic specialization and its acceptance by community speakers, as well as 'idiomaticity' (or non-compositionality, as the meaning of the lexical unit cannot be deduced from the sum of its components) - have been generally accepted by most linguists.

When describing these units, researchers have also taken into account the morpho-syntactic, semantic and pragmatic features that collocations present Thus, depending on the combination of criteria adopted in collocational studies, the approaches comply mainly with syntactic criteria (Benson *et al.* 1986), lexico-syntactic criteria (Corpas 1997), semantic criteria (Firth 1957, Halliday 1966, Mel'cuk et al. 1995, Corpas 1997) and pragmatic criteria (Sinclair 1966, Kjaer 1990, Heid 1992, Gledhill 2000), though other criteria such as statistical (Krisnamurthy 2004) and conceptual (Riabtseva, 1992, Meyer & Mackintosh 1996, Montero-Martínez *et al.* 2002,) have also been applied.

In general language, 'phraseological units', 'phrasemes' and 'collocations' are some of the recurrent expressions used to refer to many of these multi-word expressions and phrases such as *idioms* ("raining cats and dogs"), *catchphrases or proverbs* ("value for money", "one swallow does not make a summer"), *formulae* ("in this day and age", "mind the step"), *compounds* ("air-conditioned system"), *free collocations ("*preliminary report") and *bound or restricted collocations* ("grind one's teeth", "meet the requirements") and so on.

For the purposes of this work, idioms, catchphrases and pragmatic formulae will not be considered since their presence is not so common in specialized domains and they can be found in many general dictionaries. Neither will free collocations be addressed as the countless combinations language produces can be more relevant in literary studies. According to this, compounds, - as combinations of several lexical units representing a concept and functioning as a categorization element,- and restricted collocations or phrasemes will constitute the core of our analysis, although some grammatical collocations will also deserve our attention. In sum, the approach here undertaken is influenced by my own academic experience and research in the different links between the English language and Computer science.

(a) First, I focus on those combinations of words that recurrently appear in real ICT texts and have become an important category of lexical patterning in this domain: compound terms ("downward compatibility", "uplink communication") as well as SPUs in which a verb or deverbal noun appears ("to boot a computer", "data processing"). All these terminological phrasemes are of

great importance for translators, communication mediators, terminologists and lexicographers.

(b) Secondly, efficient communication in LSP involves more than terms and so I would like to refer also to other types of lexical combinations very much used in scientific and academic contexts, as these 'collocations' are the 'preferred way to say things in a particular discourse' (Gledhill 2000: 1) or 'discourse community'. They are also cultural and domain-dependent as will be shown later. Take, for example, some collocations used in research articles: "Any shortcomings which remain are our own" and this Spanish pragmatic equivalent: "*Cualquier error (...) debe imputarse al autor del artículo*", etc. Evidence suggests that this kind of collocations follow certain conventions, naturally agreed by specialists, though in many cases not overtly put forward. Thus, they deserve greater attention in LSP contrastive scenarios considering its important role in discourse and genre analysis. Engineers, researchers and communication mediators need mastering these lexical combinations to produce texts that are accepted by their colleagues, as professionals throughout the world use English (online) on a daily basis.

(c) Thirdly, since collocations are perceived as semantic linguistic units, they have been featured in natural language processing (NLP) for a great variety of applications. Collocation acquisition and extraction falls within the general class of corpus-based approaches to language. Once extracted, collocations can be useful for several NLP applications: sense disambiguation, language generation systems, information retrieval, as well as for term extraction with lexicographic and terminographic purposes.

Given the interdisciplinary approach adopted in this chapter, the terms 'collocations', 'terminological phrasemes' and 'specialized phraseological units' (SPUs) will be used interchangeably as, depending on the perspective described, one or the other are preferred. In the following section, we will see the close relations of phraseology and terminology in the light of the latest approaches to terminology.

3. Specialized phraseology from the terminological perspective

Terminology is the fundamental part of a knowledge domain as knowledge is transferred by using linguistic units such as terms and phrasemes as well as non-verbal symbols, such as icons, mathematical formulae, graphics, pictures etc. These linguistic units do not appear as content-independent labels but they rather acquire the terminological 'status' in texts, within a contextual communicative framework. Both terms and specialized phraseological units have a referential role and they constitute the nodes from which the knowledge

domain is structured, that is, they denote objects, processes, states and relations specific of a knowledge domain or a specialized activity.

Specialized texts are then taken as the basis of specialized communication, and they are considered to be the suitable environment to analyse terms as well as the other combinatory expressions that accompany them. The study of the 'ecological' system of these specialized phraseological units, that is, the study of context, provides real evidence that the term is used in a certain field. In other words, the presence of a term in a field is related to a conceptual content, and it accounts for the possible relations established between terms and concepts that are specific of that subject field.

This view of specialized phraseology is based on the new approaches to the theory of terminology, Communicative Theory of Terminology (CTT) (Cabré 1999, 2003), the Sociocognitive Approach to Terminology (Temmerman 2000), the socioterminology (Gaudin, 2003) and the textual perspective, (Bourigault y Slodzian 1999). According to all these approaches, terminology can be considered as an interdisciplinary subject linked with other areas such as linguistics, translation, computer science, and information and cognitive science.

From this terminological perspective, specialized phraseological units (SPUs) can be characterized by three components: (I) a linguistic component, (II) a cognitive component, and (III) a socio-communicative component.

(I). **Among the linguistic features**, we include grammatical combinations of word classes, their variations and the semantic categories they represent (processes, events, entities, or relations). Lets us see some of these characteristics:

(a) SPUs include at least one term or a terminological unit in the combinatory group. Some of these combinations allow for certain variations as in examples (3) and (6)

(1) N+N: web site, web page

(2) Adj. + N: intelligent building, relational database management systems

(3) V. + N: to transmit a frame, (frame transmission) to run a command.

(4) V + Adv.: to create dynamically, to manipulate interactively

(5) Adv. + Adj.: strongly typed wrappers, digitally distributed environments

(6) N + Adj. + N.: computer-readable format, machine-readable technologies

machine-accessible /computer-accessible information

(b) They show some degree of idiomaticity. For instance, it is said "to kill a process", "to abort a process" but not "to assassinate a process". The degree of fixedness in SPUs varies as some words can be included between the two main elements, allowing for some spans, as in example 8.

(7) "a classic Macintosh application which will not **run natively on** Mac OS X, but runs inside the Classic environment, or a Win16 application running on Windows XP using the Windows on Windows feature in XP.

(8) "Native mode is used in computing to describe something **running** on a computer **natively** or in **native mode** meaning that it is running without any external support as contrasted to running in emulation". (Wikipedia)[1].

(c) Although it is generally accepted that connotation is irrelevant in terminology and consequently in SPUs (Thomas 1993: 47), in the ICT domain there are several examples that account for connotational references. Let us consider the term "hacker" that, depending on the cultural and professional context, it can bear favorable or denigrating connotations, ranging from a highly skilled programmer to somebody that breaks into a system disabling security measures to cause damage[2].

(d) As we have just mentioned, phraseological variations are frequent in ICT texts. Sometimes these variations can be used interchangeably and preferences for one or the other are not clear as in example (9), or there can be several terms used as synonyms, as in example (10). though they do not really refer to the same results of the action performed. Speaking loosely, a "file" can be deleted, cut, erased, wiped, removed, or overwritten and users refer basically to the same action, i.e. to make a file disappear from the screen or from a drive, but in a more precise way, the results are not the same. At this point what counts is just the action of making it disappear rather than the final results or whether the file can or cannot be retrieved anymore.

(9) EN: neural networks

SP: *redes neuronales, redes de neuronas, sistemas neuromórficos, sistemas de procesamiento paralelo*, etc.

(10) EN: delete, cut, erase, wipe, remove, overwrite a file

---

[1] Wikipedia includes a second meaning for 'native' when applied to an operating system, or an instruction set, etc. In this case it means that the corresponding item was implemented specifically for the given model of the computer or microprocessor, as opposed to emulation or compatibility mode.

[2] In *The Jargon File*, by Eric Raymond, the reader can find a wide variety of examples with different types of connotations. http://catb.org/~esr/jargon/html/online-preface.html

SP: *borrar, cortar, eliminar, quitar, machacar un fichero*

(e) SPUs can also express intensifying functions, as in the following example: "screaming fast network server connection", where screaming intensifies the speed of the connection. Following the Meaning-Text Theory, (Mel'cuk *et al.* 1995), L'Homme (2002) has studied this type of functions and states that lexical functions allow to capture semantic relations between terms, such as taxonomic relations as well as other lexical functions.

II. **The conceptual or cognitive features** include the place these phraseological units occupy in a conceptual structure, the relations they show in the conceptual map and the way they are used to transmit knowledge.

(f) SPUs refer to one concept in a domain: "download a file", "initialize a printer", "restore the printer" (send printer initialization files) "capture a screen (shot)", etc.

(g) SPUs can be crucial in organizing the conceptual map of a domain and in establishing relations between concepts. For example, a "heavyweight ontology" is a hyponym of an "ontology", and it gives us a hint that a "lightweight ontology" can also exist, as a brother concept. Thus, the two subconcepts or hyponyms inherit the general characteristics of the hypernym. Similarly, these compounds inherit collocational patterns of the main concept, and just as an ontology is designed, used, modified, opened, pruned, mapped, etc., the subconcepts inherit this phraseology.

(h) ICT texts are flooded with metaphorically-based SPUs. In ICT texts, the majority of new terms are metaphorical extensions of the words used in the general language. It is very common to ascribe personal features to hardware or software: programs go mad, buildings and clothes are intelligent, terminals can be smart or dumb, surfers surf the web, spaces are collaborative, interfaces are intuitive, etc. If Artificial Intelligence (Pavel 1993:25) was paved with theatre-related metaphors, such as frames, scripts, scenarios, actors, thematic roles, settings, etc, the highly figurative phraseology of the Internet nourishes from nature and society: oceans, traffic, violence, fire, etc.

(i) SPUs also provide interesting information about the collocates and thus they allow for a classification of collocates in semantic classes, as can be seen in Table 9-1.

Verb Noun	Terms	Comments
Support (V)	- Software (Java, Linux, etc...) - a language, - language features - upgrades - serial ports	This verb collocates with (a) hardware devices that are compatible or that can be used with other hardware elements; (b) with software (programs,

	-	booting from the CD-ROM	communications, ...)
	-	multimedia communications	
Transmit (V)	-	wireless video	This verb collocates with SW or the like (signals, video, music...)
	-	wireless power	
	-	wireless signals	
	-	wireless data packets	
	-	wireless information	
Run (V)	-	an update on the booting	This verb collocates with HW and SW elements. Run on a processor/ Unix systems, Run under Windows, Unix, Be run from
	-	a program	
	-	the system	
	-	on Linux,	
	-	Linux on your system	
	-	a command on a machine	

Table 9-1

III. **The socio-communicative perspective** reflects the pragmatic features mentioned in section I: subject field, users, situation, purpose, etc of specialized communication. -

(j) Some geographical preferences can be seen in the use of *computador/ ra* versus *ordenador* in Spanish, as the first is preferred in all South American countries, whereas the latter is more common in Spain, except in University Departments, where they speak about *Arquitectura de Computadores* (Computer architecture) *Tecnología de computadores* (Computer technology).

4. Specialized phraseology from the translation perspective

It is commonly accepted that one of the main obstacles in producing a good technical translation is to find the right equivalents of the source terms in the target language. In the traditional theory of terminology, the lexical problems of technical translation were reduced to finding a mere substitution of the source-text term by a target-text term. This is a serious matter indeed. But very often one important difficulty resides in finding the right lexical combinations that make the text appear as written by an expert and not by a layman. More than ever it can be said that we are witnessing what Pavel (1993: 21) describes as "terminology-in-the-making" process since ICT technologies are continuously creating a glut of new terms and phraseology due to their development in interdisciplinary fields. Thus, ICT translators from English into Spanish have to

face several problems partly derived from the lack of updated reference sources in the target language (Spanish). This leads to the overuse of loan words (*anglicismos*) with the morphological, syntactical and collocational problems this entails Some of the most important difficulties translators have to face are the following:

(a) The number of English loan words *(anglicismos)* is enormous as new technologies are developed and consequently new terms and phraseology are constantly appearing. Let us mention just one example: grid computing, and a host of larger compounds and phraseology derived from this one: grid computing developments, resources, solutions, applications. There have been some attempts to translate it: *rejilla informática, computación distribuida*, but up to now, none of these solutions has been widely accepted. Other English terms have almost consolidated their presence by themselves and in compounds: hardware, software, byte, bit, web, applet, cookie, blog, as well as in some collocations "create a blog", "retrieve a cookie", "create a cookie", etc.

(b) Adjectives with well established general meanings become part of new lexical combinations and express new meanings in this domain by metaphorical use. These words go through a process of terminologization, as in the use of adjectives such as legacy[3], proprietary, collaborative, interactive, intuitive, user-friendly, etc., illustrated in the following examples. In some of them human qualities are attributed to machines and programs.

(11) EN: legacy system, legacy data, legacy programs run on obsolete hw.

SP: *sistemas legados, datos legados, programas legados que funcionan o corren sobre / con hardware obsoleto.*

(12) EN: Interactive system, program, environment, media, game, process, interface, visualization...

SP: *atlas, mapa, aplicación, cuento, sistema, juego, entorno interactivo.*

[3] Legacy is used to refer to old, 'antiquated' systems, but the connotations of 'old' and 'antiquated' are not advisable from the marketing point of view, so this term acquires some positive connotations. In such cases, 'legacy' has an elegant flavour that 'antiquated' and 'old' do not provide. In the on line BNC corpus and the Cobuild Corpus it does not appear with this new meaning, neither does it appear the collocation "sistema legado" in the Spanish CREA. However, in the on-line Webster dictionary this adjective is specifically referred to computing: "of, relating to, or being a previous or outdated computer system". Data from these corpora were obtained in October, 15th, 2006.

(13) EN: collaborative[4]: environment, method, interact in a collaborative space,

SP: *método, entorno colaborativo, interactuar en un espacio colaborativo*

(14) EN: proprietary: systems, software[5], formats,

SP: *sistemas propietarios como* Windows, software propietario (although the right equivalent is *patentado, de marca registrada*)

(15) EN: design an intuitive[6]: interface, program, application

SP: *diseñar interfaz intuitiva, programa intuitivo, aplicación intuitiva.*

(c) As above mentioned, in ICT texts it is common to have several terms and collocations for the same concept, object or function, many of them based on metaphorical or metonymical uses. A well-known example is surf, navigate, cruise, ride, browse the web or 'google' as a verb[7]. Most of these terms are related to the metaphorical concept of "Internet is an ocean" or "Internet is a highway". However when translating into Spanish, some of these metaphorical senses disappear and the equivalents range from maintaining the same metaphors related to the ocean, as in *navegar por la web* o *navegar por Internet, or surfear,* to other synonyms of the verb to browse, and the metaphorical concept of "Internet is a bookstore or a library": *curiosear (*to browse*) por la web, curiosear por Internet, ojear (*to have a look at*) la web, ojear Internet, hojear* (to glance through, to leaf through) *la web, hojear Internet*[8].

(d) The ICT world is probably one of the most active in creating neologisms for fun, social communication and technical debate. As *The Jargon*

---

[4] This adjective has reached greater importance with the advent of the Internet and the Computer Mediated Communication (CMC), and once more a human quality is attributed to machines. In Spanish it is a neologism.

[5] It is a synomym of non-free software. The new collocative meaning derives in different oppositions from the original meaning. In this field, proprietary software is opposed to free software.

[6] The evolution of command-based programs such as DOS systems to menu-driven ones has influenced the application of the adjective intuitive to qualify easy-to-use programs rather than saying user-friendly programs

[7] In Aguado (2006) the reader can find a greater variety of metaphorical examples and their translations.

[8] The results of a recent search for Spanish texts in Google (October, 20, 2006) gave the following hits: surfear la web, 189 pages, surfear Internet, 65 pages; curiosear la web, 19 pages, curiosear por Internet, 76 pages; ojear la web, 51 pages and ojear Internet 58 pages; hojear la web, 29 pages and ojear Internet 93 pages.

File[9] explains, hackers incorporate "their own myths, heroes, villains, folk epics, in-jokes, taboos, and dreams". This cultural world poses many problems in translation. A well-known example is "spam" and "spamming". Originally it had a negative connotation in America and Great Britain as it represented some sort of cheap, canned meat, produced during the war and later popularized by the British Monty Python group as something undesirable.

(e) Textual information can help translators to understand and determine the relations between several collocations, especially when there is a meronymical relation in the equivalents selected. Moreover, they can provide cohesion by means of encapsulation and prospection, as these two cohesive devices can be linguistically realized by means of hypernyms in technical texts (Alvarez de Mon 2000a. 2000b). For instance, a 'pendrive' is translated as *memoria* flash, *memoria* USB, *llavero* USB, *lápiz informático*, disco duro portátil, etc. In each of them a different trait is highlighted: the type of port, USB; the function, *memoria, lápiz*; the shape, *llavero*. Let us have a look at the following examples in which by means of the definition or an explanation in English, the translator can learn about this device:

(16) a **pendrive** es **una unidad de disco duro intercambiable** de hasta 1 GB de capacidad y del tamaño de un pulgar, con conexión USB.

(17) Una **memoria USB** o *Pendrive* (*Universal Serial Bus*) (en inglés *USB flash drive*) es un pequeño dispositivo de almacenamiento que utiliza memoria flash para guardar la información sin necesidad de pilas. (Wikipedia)

(18) Los **receptores externos** vienen en forma de **lápiz informático (conexión del tipo USB)** o **de tarjeta estándar** para ordenadores portátiles.

(19) Flamante **llavero USB** con antivirus. El U3 Smart Drive, de la marca Verbatin, le pone un toque de distinción a las **memorias USB**.

(f) Some morphemes (prefixes and suffixes) acquire a new value when translating some collocations. For example, screen shot, screen dump, or snapshot are generally translated by one-word term, *pantallazo*. The suffix '-azo' is used in Spanish with an augmentative or derogatory meaning, as well as to refer to a knock or a blow given with an object. *Pantallazo* has neither of these meanings, but denotes the action of inserting or capturing a screenshot.

(g) Technical documentation and user manuals can consolidate expressions in Spanish that were not normally used to express the same actions. For example

[9] http://catb.org/~esr/jargon/html/introduction.html

'press a key' (sometimes 'push a key' is also used) had always been translated by *pulsar, presionar* or *apretar una tecla*, but in recent searches on the Web, the documents retrieved show a high frequency of the verb *oprimir* (to 'squeeze') accompanying *tecla, botón, interruptor,* or an adverb, *oprimir aquí,* as the phraseological expression for the English equivalent 'press a key', 'press a button', or 'press here'. It could be argued that it is incorrect, but when translators make a search on the Web this is one of the commonest translations they will find.

(h) Some of the difficulties in translating collocations are derived from the semantic overload of some terms such as platform, environment, support, development, cluster, interoperability, (as in the case of semantic interoperability) used many times as 'cliché words' and the translator does not find the right verb that accompany as those collocations are not included in dictionaries. Let us see some phraseological units in Spanish with the noun *plataforma* (platform) and the verb *soportar* (support). In this case, the verb *soportar* can be classified as a phraseological verb, according to Lorente (2002).

(20) **Compilar** en una plataforma Windows

(21) Los servidores **trabajan bajo la plataforma** Unix

(22) La **plataforma** Unix **incorpora** funcionalidad on demand

(23) La **plataforma** Unix **soporta** X dominios

(24) La versión Server **soporta** un nº ilimitado de usuarios

(25) El gestor de tipografías **soporta** OpenType

(26) Java **soporta** las plataformas Windows NT y Mac

(27) El producto ya disponible soporta una amplia variedad de protocolos

Sometimes verbs have been specifically created in one field in one language, such as to input, to download, to upload, to debug, in computing and, in that sense, they can be considered as domain-specific terminological verbs because of their semantic specificity. However, the equivalents in Spanish *introducir, descargar o bajar, subir o colgar, depurar*, do not have that semantic specificity, but they acquire it by way of collocations, or the collocates they go with. In this case they can be considered as phraseological verbs.

We have seen that terms and SPUs are important to achieve an effective communication in LSP contexts, but there are other types of phraseological structures which, according to Nattinger and DeCarrico (1992:11), play a significant role in the rhetorical construction of academic and scientific texts. The following section will deal with these phraseological units.

5. Phraseology in academic texts

As mentioned before, specialized phraseology can also be approached from a more discourse-based and academic perspective (Howarth 1996, Gläser 1998, Gledhill 2000). In the last ten years the academic perspective has prioritized the study of the great variety of expressions and lexical combinations that a discourse community uses according to their rhetorical aims. Academic discourse, especially in the engineering domains, is founded on a system of preferred expressions, i.e., some recurrent lexical patterns that professionals prefer to use in order to communicate in a more effective and precise way. Collocations are then a fundamental mechanism that allows for new formulations to take place throughout the text.

Some authors state (Riabtseva 1992: 380) that, in academic style, collocations are: communicatively obligatory, rhetorically relevant and phraseologically bound. Furthermore, Riabtseva contends that different languages generate similar, but not identical collocations for the same conceptual patterns. Therefore, collocations can not be translated literally but need to be reinterpreted considering the target language, thus confirming that phraseology is domain-dependent and language-bound[10].

In ICT English-written academic papers, the IMRD structure (Introduction, Method, Results, Discussion/ Conclusions) (Swales, 1990) is generally adopted (Posteguillo 1999), and each rhetorical section shows a particular phraseology in the sense that certain collocations are preferred to others. Verbs used in these rhetorical sections are called 'discursive verbs', as they allow to organize the information and provide metadiscursive data, such as the author's stance, or the communicative function. In the following examples, some phraseology used in the different sections of a research article in ICT texts can be seen:

(a) In "Introductions", several synonyms can be used when introducing a topic:

(28) EN: The aim/ goal/ purpose /objective of this study was to.....

[10] After comparing the macrostructure of 30 articles written in English, corresponding to the artificial intelligence field, and 30 articles written in Spanish, corresponding to the mercantile law field, it is clear that the macrostructure and the lexical expressions used in academic articles is more language and cultural-dependent when the topic is more national-constrained, even though in both cases they may include a comparative study of the problem. For instance, to refer to the introductory part, in AI articles, all of them started with 'Introduction', whereas in the articles on Mercantile Law, several synonym expressions were used: *consideraciones generales, premisa, planteamiento del problema, presupuestos, introducción*, or they directly started with a general statement presenting the problem.

This paper describes, presents, deals with, shows, surveys, proposes,

This paper is focused on...

A tough challenge for the NLP community is...

(29) SP: *Este método describe, este artículo presenta, se ha aplicado el método, se han efectuado pruebas...*

(b) In the "Results" section, the word 'results' is usually accompanied by different verbs or adjectives:

(30) EN: Results confirm/ provide/ show/ suggest/ indicate/ imply/

Results obtained/ recorded/ expected[11].

(31) SP: *Los resultados confirman, implican, sugieren*

La hipótesis demuestra, ratifica

(c) In the last section, "Conclusions"[12], the following collocations can be found:

(32) EN: Anyone can draw his own conclusions from this difference ...

the conclusions he draws from the findings ...

...and derives some significant conclusions.

In a nutshell

(33) SP: *se pueden extraer las siguientes conclusiones*

en pocas palabras

(d) In the "Acknowledgments" section, some differences can also be found in Spanish and English ICT papers.

(34) EN: to thank somebody **for** their comments **on** an **earlier** version/draft of this **paper**

(35) SP: *Agradecer **a** alguien sus comentarios **a** una **primera** versión de este **artículo.***

---

[11] More examples of the different phraseology used in other sections according to the discursive functions can be found in Bowker & Pearson (2002), Posteguillo (1999) and Fortanet *et al.* (2002) although in this latter book they analyse research articles with a macro-structure approach.

[12] However, in the articles on Mercantile Law there is not always a section named Conclusions, but there are other elements that identify the section as such: a *modo de recapitulación, ofrecemos seguidamente una propuesta.., las reflexiones que anteceden tienen como objetivo..., Consideraciones recapitulativas, Consideraciones finales, A modo de conclusión, una nota final sobre..., Epílogo.*

(36) EN: Any **shortcomings** which remain **are** our own.

(37) SP: *Cualquier **error (...)** debe **imputarse** al autor / a los autores del artículo.*

The intrinsic nature of these discursive verbs is not a sufficient requirement for collocational restrictedness, but it is the N+V collocability that is of most interest. According to Howarth (1996: 91) some discursive verbs maintain their literal meaning, such as: associate, define, include, involve, mention, represent, suggest, understand, to mention just a few. However, taking a comparative view of both languages, it is interesting to verify that very often the grammatical components are not the same in both languages: apply for a job, *solicitar un trabajo.* Considering this, grammatical items are fundamental for phraseology.

Moreover, similar technical genres show that the presence of these discursive verbs can vary in the two languages. The analysis of one of the conventional protypical genres in software engineering, the Requirements Specification Document, in English and Spanish, gave different results. After analysing with the WordSmith Tools four documents in each language, (2168 tokens in English and 2890 tokens in Spanish), the first ten verbs in number of occurrences in English were: describe (75), specify (16), include (16), present (16), design (14), process (14), illustrate (11), obtain (7), indicate (7), accomplish (6). The verb *describer* (describe) and other word-types in Spanish, *descripción, describe*) appeared the first one in the ranking in both languages, in Spanish (58) but some of the other verbs were different. In Spanish, the next in number of occurrences were *definir* (36) and some variations (*definición, define), realizar,* (36*), introducir* (24), *especificar* (22), *cancelar* (22), *comprobar* (22), *clickear (*16*), abordar* (10), *identificar* (10). This shows that analysing phraseology from a contrastive point of view is extremely interesting since non-native engineers and scientists have to write both in English and in their mother tongue with the frequent results that both languages are influenced by the collocations usual in the other.

In short, it is clear, then, that phraseology is closely tied to particular subject areas and topic-specific collocations are a major defining aspect of these fields. That is why it is important to aim at raising students' awareness of those typical collocations in their domain fields.

6. Specialized phraseology in NLP applications

In the previous sections we have highlighted the different dimensions that a collocation has in specialized texts. In computational terminology, engineers, and computer scientists have also realized that the formalization of specialized

texts is necessary to build useful applications. Hence, depending on their interests they have worked on formalizing terminological data (Gillam & Ahmad 1996) or developing software to extract terms and collocations from texts (Bourigault 1994, Smajda, 1993). Other researchers have focused on combining ontological and linguistic tools (Faber & Jiménez, 2002) or on building ontologies of collocations (Milkov 2000). Some other important applications have centred on automatic abstracting (Voorhees & Buckland 2005). Many applications are dictionary-oriented and in this respect, some interesting attempts have been made in automatic compilation of specialized collocation dictionaries (Wanner et al. 2005).

Generally speaking, NLP applications for term and collocations extraction fall under two research approaches: a statistical approach which is beyond the scope of this article and a lexical approach. However, combined approaches (statistical and linguistic) have also been developed (Daille 1994). The lexical method has followed two main lines, very much in the sense that we have presented in section 3 and 5.

The first one, a term-based line, tries to detect and extract compound terminological units. In this respect, according to Kageura & Umino (1996: 260), two notions are relevant: **termhood** and **unithood**. Termhood refers to the degree that a linguistic unit is related to domain-specific concepts and it is relevant to simple and complex units. Unithood refers to the degree of stability in collocations, and this concept is relevant to simple and compound terms as well as other grammatical collocations.

The second line, more linguistically-based, is interested in basic linguistic structures that provide knowledge patterns, as Meyer calls them (2001: 290). Strictly speaking these linguistic structures do not fill in the slot of specialized collocations. They can be considered as grammatical collocations (in the sense of Benson et al. 1986/1997) but they are of great interest since they provide knowledge on conceptual relations.

Following this line, in computational terminology, knowledge extraction tools are based on the assumption that conceptual relations are linguistically realised by means of certain predictable, recurrent lexico-syntactic patterns. These patterns provide knowledge on some kinds of relations, such as hyperonym-hyponymy[13], or class-subclass_of (*is a, classified as*, etc), meronymy, (*contains, is a part of*) or other non-hierarchical relations, (function: *designed for*, *used for,* cause-result, *caused by,* etc). These relations have been mainly studied for English (Kavanagh 1995), though there are also some works

[13] This method has also been employed to identify some geographical classes and entities with regard to an ontology (Cimiano & Staab 1992) by using Hearst's patterns (1992: 540): "...such as..."; "...or other...", "...including...", "especially...".

carried out for the part_of relation in Spanish (Climent, 2000, Diez Orzas, 1999).

With regard to taxonomical relations, in Aguado & Alvarez de Mon (2006: 495)[14] we analysed the phraseology of the class_of relation in Spanish, with the practical aim of extracting knowledge for an ontological application. The first linguistic results of the study showed that several kinds of lexical units could provide some conceptual knowledge. First, lexical patterns based on some verbs, such as *clasificar* (classify), *distinguir* (distinguish), *figurar* (figure) and other synonyms. The following examples are some of the lexical patterns with the verb *clasificar*

(38) Según/ de acuerdo con (criterio), las/los H se clasifican en X, Y, Z, etc.

(39) ... se clasifican en H de X o de Y

(40) Los/las H se clasifican en los siguientes grupos:

(41) se clasifican (generalmente / básicamente/ comúnmente) en diversos/varios tipos

(42) Los/las H se clasifican en X o Y

Besides these knowledge patterns, there are other lexical units that can provide information on hierarchical relations. These are generic nouns that indicate hyperonymy, such as *categoría,* (category), *clase, (*class) *tipo,* (type) *grupo,* (group) *etc.* Thirdly, certain types of words, for example, determiners such as *los siguientes, (*the following), other numerical expressions or other lexical units that indicate the end of the hyponyms: *y, así como.* All of them conveyed the idea of hyperonymy. Finally, some paralinguistic elements, mainly punctuation, semi-colon, colon, could also help identify hyperonymy as well as some terms in bold.

In sum, both collocations and phrasemes seem to offer an interesting potential field of research and application in many different NLP areas such as information retrieval, information extraction, machine translation, word sense disambiguation, name entity recognition and knowledge acquisition.

7. The Internet as a reference corpus for phraseology

As becomes clear from the preceding passages, we have presented different approaches to specialized phraseology. In this last section, our aim is to show

[14] This analysis has been carried out within the research Project SEMANTIC SERVICES: "Infraestructura tecnológica de servicios semánticos para la *web* semántica", Plan Nacional de I + D + I 2004-2007, 2660, funded by the Spanish Government.

why the Web can be considered as a corpus and more specifically as a reference corpus for all these LSP professionals.

In corpus linguistics, a discipline much older than the Web, a collection of texts was considered as a 'corpus' if it complied with the following requirements: sampling and representativeness, finite size, machine-readable form and a standard reference (McEnery & Wilson 1996: 21). More recently, Bowker and Pearson (2002:10) also proposed four characteristics to define a 'corpus' that partially coincide with the above mentioned: authenticity, electronic format, largeness and specific criteria. In this view, two of the former features, 'sampling' and 'standard reference', have been replaced by 'authenticity' and 'specific criteria'. These definitions have been more loosely interpreted in other disciplines, since the purpose of the corpus and the research that can be carried out on it are some critical issues when using the Web as a corpus.

According to this, in the field of natural language processing (NLP), Manning & Schütze (1999:120) claim that, for statistical NLP purposes: "One commonly receives as a corpus a certain amount of data from a certain domain of interest without having any say in how it is constructed". Following this line, Kilgariff & Grefenstette (2003: 334), in an special issue on the Web as corpus of the *Computational Linguistics* journal, asserted that the *web is a corpus* and defined a 'corpus' just as "a collection of texts", thus adopting a broad and more practical view of the concept. They argued that McEnery and Wilson mixed the question "What is a good corpus?" with "What is a good corpus for?". Put it at its simplest, the purpose of the research is the rationale that guides the validity of a corpus.

However, the concept of 'Web as corpus' can be understood in at least two different ways: first, the Web **for** corpus, i.e. the source to compile machine readable texts, and second, the Web **as** corpus, i.e. the body of freely available online documents accessed directly as a corpus.

There are several reasons to use the Web **for** creating a corpus with online materials. Fletcher (2005) points out the following: (a) freshness and spontaneity, as webpages are authentic examples of current language; (b) completeness and scope, as existing corpora may lack a text genre or content domain of interest, or else may not provide sufficient examples of an expression or construction easily located online; (c) linguistic diversity, as languages and language varieties for which no corpora have been compiled are found online; (d) cost and convenience since the web is virtually free and webpages are available to researchers and professionals alike; and (e) representativeness, as the proportion of information, communication and entertainment delivered via the Net is constantly growing and increasingly reflects new uses of language.

Undoubtedly, these are in fact very good reasons to take the Web as an almost immeasurable repository **for** creating a corpus, and at the same time these grounds can also be applied for using the Web **as** corpus (Volk 2002). In this sense, the Web has not the disadvantages of compiling a traditional corpus: a lot of effort, a time-consuming task, difficult to update, etc. On the contrary, the Web allows downloading all sorts of documents in electronic form, with different varieties of language. Besides, they are authentic texts, representative of the language types companies need to handle. Hence, they turn into a reliable testbed to be used in preparing students as prospective professionals. From this wide perspective, the Web opens a wide and generous universe to carry out research from the different perspectives and it is an instant help to technical translators and writers, professional communicators and NLP researchers.

In the last decade, the use of the Web **as** corpus has taken a great relevance in different fields. But, can the Web be considered as a *reference corpus*? According to the *Merriam Webster on line Dictionary*, 'reference' is the act of consulting, and the Web is unquestionably the best resource for consulting at a click of the mouse. In this sense it is a reference corpus[15]. Then again, we can put forward the following question. What do LSP practitioners need this kind of reference corpus for?

For LSP teachers, the Web provides a wealth of specialized technical materials that otherwise would be difficult to collect. As a consequence, finding and classifying collocations, definitions, classifications among other lexical and terminological issues can be carried out easily. An additional point of interest of the Web **as** corpus is by using different search engines, Google, AltaVista, Yahoo, Excite, MSN Search, among others, to search for certain specialized collocations or lexical combinations to check their usage and the contextual domain in which they appear.

Translators and terminologists have found the Web as the most exciting and challenging tool for looking up a word, a terminological unit or a specialized phraseme. By using advanced search facilities and restrictions for particular languages, to find neologisms in context, not included in paper or electronic dictionaries, has turned into an easy task. It can also be used to validate the frequency or the possible combinations of specialized collocations and compound terms. Another important issue for translators is addressing the problem of parallel corpora. In this regard, Resnik (1999) developed a method

[15] What purists consider a reference corpus (Baker 2006: 30) is a large corpus, usually consisting of millions of words from a wide range of texts and which is representative of a particular language variety, as the British National Corpus (BNC) for English or the *Corpus de referencia del español actual* (CREA), and the *Corpus Diacrónico del Español* (CORDE).

to find parallel texts in the Web by using a search engine and (Baroni *et al.* 2006) designed a web tool to build instant corpora.

However, some linguistic precautions should be taken and some filtering is necessary in order to achieve some coherent results when searching for some specific information. Quite often the hits retrieved are repetitive or not representative of any reliable text, some texts are of unknown authorship, and some are merely captions or fragments mixed with personal opinions. Nevertheless, comparing frequencies is a good means of evaluating the vitality of some terminological and phraseological expressions and their uses, but this method has to be contrasted also by checking other documents. By way of example, in a web search carried out to verify the usage of the Spanish verb *clasificar* (to classify), to detect the conceptual relation subclass_of, the first ten hits referred to Real Madrid, a Spanish football team. This proved that, although web searches can be really helpful in some cases, on other occasions they are completely misleading and statistics are not reliable.

On the other hand, from the production point of view the Web can be used to confirm or reject intuitive decisions taken on certain words and select phraseological units that help produce natural-sounding text. Notwithstanding these contributions, it is necessary to bear in mind that, for instance, many of the English texts retrieved are not written by native speakers and they become part of what is known as 'Global English', that is, English used as an international means of communication. This does not mean that all the hits retrieved are good examples of a language, as in a way, the Web gives preference to content rather than to perfect stylistic chunks of language combinations.

In summary, the Web offers huge possibilities as has been pointed out, first to gather a great variety of documents in a few seconds, representative of new genres or subgenres that may not be found in a standard corpus, and second, to look up other on-line corpora in search of a certain linguistic phenomenon. On the other hand, with the advent of electronic corpora and the Internet, some of these tasks have been facilitated mainly for pedagogical purposes. In the Internet there are several programs to extract concordances on line from different mini corpora online and thus the samples retrieved are usually restricted to 50 cases[16].

[16] There are many programs that allow the user to search online for a certain collocation in different corpora, with a variable size. For instance, http://www.lextutor.ca/, http://www.natcorp.ox.ac.uk, http://www.webcorp.org.uk/, provide teachers and students with many profiles to perform a great range of activities. A complete web site on corpus-based linguists can be found at http://devoted.to/corpora, with tools, articles, corpus in different languages, etc.

8. Closing comments

In this chapter I have attempted to give a brief outline of phraseology in some of the fields in which its presence is relevant for several purposes. It is clear that SPUs and collocations represent one of the major difficulties in theoretical linguistics as they cannot be studied in any one of the traditional divisions of linguistics (i.e. neither in semantics, nor in syntax), precisely because of their irregular semantic and syntactic nature. However, nowadays, lexical knowledge has assumed the central role it deserves both in linguistics and in computer applications.

As we have seen SPUs cause serious problems for translators and technical writers, since producing and translating them into the target language is rendered difficult by the syntactic and lexical characteristics of each domain and each language. There is a need to combine or reconcile the linguistic constraints imposed by a particular language with the expectations found in a particular field. Many translations of specialized texts are characterized by the use of these phraseological units and, although they observe the rules of the language system they are not in conformity with standard LSP usage. The notion of native-like proficiency in a language depends crucially on the stock of those phraseological units that are specific of a certain field. Hence, language-awareness is important for these professionals if they want to be able to identify and reproduce these lexical units in order to communicate efficiently.

SPUs as well as other lexico-grammatical collocations cannot be conceived as frozen, static items but as dynamic representations within a conceptual network and they can be analysed and explored using computational methods. In fact, LSP professionals, terminographers and computational linguists can get the most in knowledge acquisition value by using computer-readable corpora. In this sense, harvesting the web for terminological and linguistic purposes will greatly increase since the web has proved to offer many possibilities for research in all these fields. One aspect that has not been studied in this article, for the sake of space, is phraseology in terminographical tools, i.e. on-line mono- and multilingual ICT dictionaries, though they are generally characterized by a lack of explicit information on word combination possibilities. That is why developments on specialized dictionaries of collocations are in great demand for these professionals.

9. References

Aguado de Cea, G. and I. Alvarez de Mon. "Estructuras de clasificación en español. Terminología y adquisición de conocimiento explícito para la

Web semántica" in Pérez-Llantada, C., R. Plo and C.P. Neumann *Actas del V Congreso Internacional AELFE*, Zaragoza, pp. 492-499, 2006.

Aguado de Cea, G. "De *bits* y *bugs* a *blogs* y *webs*: aspectos interdisciplinares, socioculturales y lingüísticos de la terminología informática" in Gonzalo C. and P. Hernúñez (eds.) *CORCILLVM. Estudios de traducción, lingüística y filología dedicados a Valentín García Yebra.* Madrid: Arco/Libros, pp. 693-720, 2006.

Alcaraz, E. *El inglés profesional y académico.* Madrid: Alianza Editorial, 2000.

Alvarez de Mon, I. *La cohesión del texto científico-técnico. Un estudio contrastivo inglés-español.* Ph. Doctoral Dissertation. CD-ROM. Universidad Complutense de Madrid, 2000a.

Alvarez de Mon, I. "Cohesion in written texts: An Analysis comparing English and Spanish" in *International Journal of Translation*, Bahri Publications, Vol. 12, 1-2, pp. 81-96, 2000b.

Baker, P. *Using corpora in discourse analysis.* Continuum: London/New York, 2006.

Baroni, M., A. Kilgariff, J. Pomikálek and P. Richlý. "WebBootCat: a Web tool for instant corpora", *Proceedings EURALEX*, 2006, Turin, Italy, 2006.

Benson, M, E. Benson and R. Ilson. *The BBI Combinatory Dictionary of English. A guide to word combinations.* Amsterdam/Philadelphia: John Benjamins Pub. Co., 1986/1997.

Bourigault, D. *LEXTER, un logiciel d'Extraction de TERminologie. Application à l'acquisition des connaisances à partir de textes.* PhD tesis. Paris: École des Hautes Études en Sciences Sociales, 1994.

Bourigault, D. and M. Slodzian. "Pour une terminologie textuelle". *Terminologies Nouvelles,* nº 19. pp. 29-32, 1999.

Bowker, L. and J. Pearson. *Working with specialized language: a practice guide to using corpora.* London-New York: Routledge, 2002.

Cabré, M.T. *La terminología: representación y comunicación.* Barcelona: IULA, 1999.

—. "Theories of Terminology. Their description, prescription and explanation". *Terminology* 9:2, pp. 163-199, 2003.

Cimiano, P. and S. Staab. "Learning by Googling". *SIGKDD Explorations Newsletter*, 6 (2): pp. 24-33, 2004.

Climent Roca, Salvador. *Individuación e información Parte-Todo. Representación para el procesamiento computacional del lenguaje.* Estudios de Lingüística Española. Vol. 8. http://elies.rediris.es/, 2000.

Corpas, G. *Manual de fraseología española.* Madrid: Gredos, 1997.

Cowie, A. P. (ed.). *Phraseology: Theory, Analysis and Applications.* Oxford: Clarendon Press, 1998.

Daille, B. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filters linguistiques*. PhD dissertation. Paris: Université Paris VII, 1994.

Diez Orzas, P. L. "*La relación de meronimia en los sustantivos del léxico español: Contribución a la semántica computacional*". Tesis Doctoral. http://elies.rediris.es/, 1999.

Faber, P. and C. Jiménez (eds.). *Investigar en terminología*. Granada: Ed. Comares, 2002.

Firth, J.R. *Papers of Linguistics 1939-1951*. Londres: OUP, 1957.

Fletcher, W.H. "Concordancing the Web: Promise and Problems, Tools and Techniques". To appear in Hundt, M., N. Nesselhauf and C. Biewer, C. (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi. [Available 22 July, 2006. http://kwicfinder.com/FletcherConcordancingWeb2005.pdf], 2005.

Fortanet, I. (coord.). *Cómo escribir un artículo de investigación en inglés*. Madrid: Alianza Editorial, 2002.

Gaudin, R. *Socioterminologie. Une approche sociolinguistique de la terminologie*. Bruxelles: Duculot, 2003.

Gillam, L. and K. Ahmad "Knowledge engineering terminology (data)bases" in *Proceedings of the 4th International Congress on Terminology and Knowledge Engineering* (TKE) pp. 205-214, 1996.

Gläser, R. "The Stylistic Potential of Phraseological Units in the Light of Genre Analysis" in Cowie, A.P. (ed.). *Phraseology: Theory, Analysis and Applications*. Oxford: Clarendon Press, pp. 125-144, 1998.

Gledhill, Ch. *Collocations in Science Writing*. Tübingen: GunterNarr Verlag, 2000.

Halliday, M. A. K. "Lexis as a linguistic level", in Bazell, C., J. C. Catford, M. A. K. Halliday and H. R. Robins (eds.). *In Memory of J.R. Firth*. London: Longman, 1966.

Hearst, M.A. "Automatic acquisition of hyponyms from large text corpora". *Proceedings of the 14th conference on Computational Linguistics*, USA: New Jersey, Morristown, pp. 539-545, 1992.

Heid, U. "Décrire les collocations". In *Terminologie et Traduction* 2 (3), pp. 523-548, 1992.

Howarth, P. A. *Phraseology in English Academic Writing*. Tübingen: Niemeyer, 1996.

Kjaer, A.L. "Methods of describing word combinations in language for specific purposes". *IITF Journal*, Vol. 1 (1-2): 3-20, 1990.

Kageura, K. and B. Umino. "Methods of automatic term recognitions: A review" *Terminology*, Vol. 3(2) pp. 259-289, 1996.

Kavanagh, J. *The Text Analyzer: A tool for knowledge acquisition from texts.* M.Sc. thesis. University of Otawa, 1995.

Kilgarriff, A. and G. Grefenstette. "Introduction to the Special Issue on the Web as Corpus", *Computational Linguistics*, 29, (3), 333-347, 2003.

Krisnamurthy, R. (ed.). *English Collocation Studies. The OSTI Report*, Continuum, 2004.

L'Homme, M.C. "Functions lexicales pour represénter les relations sémantiques entre termes". *TAL*. Vol.43, nº 1/2002, pp.19-41, 2002.

Lorente, M. "Verbos y discurso especializado" *Estudios de Lingüistica Española (ELIES)*16 [Avaliable at http://elies.rediris.es] ISSN 1139-8736, 2002.

Manning, Ch. and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, 1999.

McEnery, A. and A. Wilson. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

Mel'cuk, I., A. Clas and A. Polguere. *Introduction à la lexicologie explicative et combinatoire,* Louvaine-la-Neuve: Duculot, 1995.

Meyer, I. "Extracting knowledge-rich contexts for terminography". In Bourigault, D., C. Jacquemin and M. C. L'Homme. *Recent advances in computational terminology*. Amsterdam/Philadelphia: John Benjamins Publishers, 2001.

Meyer, I. and K. Mackintosh. "Refining the terminographer's concept-analysis methods: How can phraseology help?" *Terminology* Vol 3(1) pp.1-26, 1996.

Milkov, N. "Logico-Linguistic moleculism: Towards an Ontology of collocations and other language patterns". In Simov, K. and E. Kiryakov (eds). *Ontologies and Lexical Knowledge Bases, Proceedings of Ontolex*, 2000, pp. 82-94, Sozopol, Bulgaria, 2000.

Montero-Martínez, S., M. García de Quesada and P. Fuertes_Olivera. "Terminological phrasemes in Ontoterm", *Terminology* 8:2, pp. 177-206, 2002.

Nattinger, J.R., J. S. DeCarrico. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press, 1992.

Pavel, S. "Neology and Phraseology as Terminology-in-the-Making". In Sonneveld, H. B. and K. L. Loening (eds.). *Terminology: Applications in interdisciplinary communication*, 21, 1993.

Posteguillo, S. "The Schematic structure of Computer Research Articles" *English for Specific Purposes*, Vol 18, 2, pp.139-160, 1999.

—. *Netlinguistics. Language, Discourse and Ideology in Internet.* Castelló: Universitat Jaume I, 2003

Raymond, E. *The Jargon File* [Available at

http://www.catb.org/~esr/jargon/html/], 2006.

REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [on line]. Corpus de referencia del español actual. <http://www.rae.es> [Available 1 october, 2006].

Resnik, P. "Mining the Web for bilingual text", in *Proceedings of the 37th Meeting of ACL*, 527-537, College Park, MD, June, 1999.

Riabtseva, N.R. "Metadiscourse Collocations in Scientific Texts and Translation problems: Conceptual Analysis". *Terminologie et Traduction* 2 (3) pp. 375-385, 1992.

Sager, J.C, D. Dungworth and P. McDonald. *English Special Languages. Principles and Practice in Science and Technology.* Wiesbaden: Brandstetter, 1980.

Sinclair, J. "Beginning the study of lexis". In Bazell, C., J. C. Catford, M. A. K. Halliday and H. R. Robins (eds.). *In Memory of J.R. Firth.* London: Longman, 1966.

Smadja, F. "Retrieving collocations from Text: Xtract", *Computational Linguistics,* 19 (1), pp.143-177, 1993.

Scott, M. Lexical Analysis Software. *WordSmith Tools* 4.0, 1996/2006.

Swales. J. *Genre analysis. English in academic and research settings.* Cambridge: Cambridge University Press, 1990.

Temmerman, R. *Towards new ways of Terminology Description: The sociocognitive approach.* Amsterdam/Philadelphia. John Benjamins Pub., 2000.

Volk, M. "Using the Web as corpus for linguistic research" in Renate Pajusalu & Tiit Hennoste (eds.) *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Oim.* Publications of the Department of General Linguistics, 3 University of Tartu. [Available 14 March, 2006], 2002.

Voorhees, E.M. and L. P. Buckland. *Proceedings of the Fourteenth Text Retrieval Conference.* Gaithersburg, Maryland. [Available 27July, 2006], 2005. http://trec.nist.gov/pubs/trec14/t14_proceedings.html, 2005.

Wanner, L., B. Bohnet, M: Giereth and V. Vanessa. "The first steps towards the automatic compilation of specialized collocation dictionaries" *Terminology.* Vol. 11 (1) pp. 143-180, 2005.