CORPUS EXPLOITATION FOR TERMINOGRAPHICAL PURPOSES: A PROPOSED TERM EXTRACTION PROCESS FOR BILINGUAL SPECIALISED DICTIONARY ELABORATION

NURIA EDO MARZÁ
UNIVERSITAT DE VALÈNCIA

Abstract. The method and outcomes described in this research are part of a wider project aimed at creating a specialised, bilingual dictionary on industrial ceramics. One of the key stages in the elaboration of such a work is the term extraction process, since it determines which terminological units are part of the domain and should thus be included as dictionary entries. This paper deals with a proposal that is aimed at accurately detecting the units of specialised knowledge that constitute the terminology of a domain from a raw, bilingual corpus that was compiled *ad hoc*. This proposal is presented in the form of a series of stages in which prospective terms are progressively retrieved, identified and analysed with the concordance software program WordSmith Tools (WST) 5.0. In the method proposed here, the WST application WordList first generates monolexical frequency lists which are compared with the lists generated by the tool KeyWords, providing saliency data. Afterwards, Mutual Information lists provided once again by Wordlist constitute the first approach to the combinatorial aspect of terms. Subsequently, the WST application Concord, with its different options, provides further evidence on the way prospective terms collocate and combine as well as on their contextual nature, thus completing the term extraction process. This methodology, always combined with the terminographer's "manual" work, observation and intuition, has proved to be effective for the dictionary under development.

Keywords: *Corpus linguistics, Specialised Dictionary, Bilingualism.*

1 Introduction: a brief theoretical outline of what is to be done, how and why

The full significance of the research presented here becomes apparent when put into perspective as being part of a wider project in which a bilingual (English-Spanish, Spanish-English), corpus-based, active dictionary of the ceramics industry is being developed.[1] This process is currently in its final stage and the dictionary itself is intended to be published in the first half of 2011.

This article focuses on one of the key stages in the elaboration of this or any other specialised dictionary, i.e. that of term extraction. The term extraction process (TEP) here has its origins in a raw, bilingual corpus on industrial ceramics that was compiled *ad hoc*. Exploitation of the corpus (and thus retrieval and analysis of the terminological data contained therein) was carried out with the concordance software program WordSmith Tools (WST) 5.0.[2] WST is an integrated suite of programs for looking at how words behave in texts (Scott 1998), apart from providing varied corpus counts which may be used for different purposes.

[1] This paper originates from the project "*Lexicografía especializada para la creación de un diccionario inglés-español / español-inglés de terminología de la industria cerámica y azulejera*" awarded by the Generalitat Valenciana (Valencia, España); CODE GV05/121.

[2] There are many other alternative corpus query programs with similar applications and possibilities such as AntConc or MonoConc (free software packages quite easy to use conceptually) but, once you get used to it, the potential of WST is much bigger since it has more features.

The dictionary-making project as a whole saw the light as a proposal aimed at "filling a gap" in specialised communication by helping to improve the work of both translators and specialists/professionals in the field. From the very first moment and in accordance with the theoretical positioning adopted in this study, the use of a corpus was considered an imperative. As De Schryver and Prinsloo (2000: 292) stated "the intensified systematic exploitation of electronic corpora for lexicographic purposes has unmistakably revolutionised dictionary making". However, the use of electronic corpora for terminological/terminographical purposes has been accepted much more slowly. Historically, the scant use of corpora in terminology management may be better understood with Sager's (1990) words, even though he positions himself in favour of the utilisation of the "semasiological" process also for terminographical practice:

> "Traditional terminological theory identifies its approach as 'onomasiological', i.e. a 'naming' approach, because, in principle, it starts from concepts and looks for the names of these concepts. By contrast, the lexicographical approach is called 'semasiological', i.e. a "meaning" approach, because it starts from words and looks for their meaning. In reality, the onomasiological approach only characterises the scientist who has to find a name for a new concept (an invention, a new tool, measurement, etc.); the terminologist, like the lexicographer, usually starts from an existing body of terms to start with" (Sager 1990: 56).

In this light, the corpus has also and necessarily constituted the departure point for the TEP presented here. As Čermák (2002) states, the best information always comes from direct data and a major contribution of corpora may be seen in their offer of authentic and recurrent language combinations in context. Hence, the research presented aims to describe how the retrieval of terms from a corpus (made possible thanks to terminography) may be systematised for the wider task of elaborating specialised, user-oriented and user-friendly quality products in the form of dictionaries (Bergenholtz and Tarp 1995), thereby giving rise to what could more accurately be known as specialised lexicography.

In line with this positioning, the approach adopted for the creation of the bilingual dictionary of the ceramic industry has Cabré's (1999) Communicative Theory of Terminology (CTT) as one of its mainstays. Accordingly, the TEP proposed here utilises a corpus-based approach which allows terms to be analysed *in vivo* and characterised from the natural habitat in which they occur in specialised discourse. In this sense, as Cabré and Estopà (2002) concede, the analysis of terms in context opens the door to three important observations in the development of terminology as a field of study. Firstly, it allows the formal, conceptual and functional diversification of terminological units to be observed. Secondly, it enables other units of specialised knowledge to be detected beyond terminological ones. Thirdly, it enables the units of specialised knowledge to be placed within a multi-relational cognitive structure.

The method used in the development of the dictionary of industrial ceramics comprises 8 broad stages based on Auger and Rosseau's (1987) study *Méthodologie de la recherche terminologique* and the adaptation carried out on it by Gómez and Vargas (2002, 2003a, b) and Vargas (2005) so that it fits into current methodological trends and technological innovation in the field of corpus linguistics and terminographical research. The stages are the following:

1. Definition of the work
2. Work preparation and corpus compilation
3. Elaboration of the field diagram
4. Documentary corpus management
5. Term extraction

6. Data processing
7. Revision and normalisation
8. Editing

Of these eight stages, this paper focuses on the fifth one, term extraction, which implies the previous completion of stages one to four. Compiling an adequate corpus and rendering it compatible with the formats accepted by the program to be used subsequently (stages 1 to 4 above) are especially important for the corpus-based TEP proposed here. Hence, in the fourth stage, the bilingual corpus has to be digitalised (by means of an OCR process when necessary) and converted into a *.txt* file. The fifth stage, the one dealing with the term extraction process as such, comprises, broadly speaking, the stages shown in Table 1. These stages, further developed in section 3, have been approached in a sequential and complementary manner with WST. These "tools" of the program are called: WordList, KeyWords and Concord.

Outline of the term extraction process (TEP) undertaken with WordSmith Tools		
WordList (phase I)	Preselection (stage 1)	Comparison and combination of frequency and saliency results for obtaining a first list of prospective monolexical terms.
	Generation of monolexical frequency lists (stage 2)	
KeyWords	Generation of saliency lists (stage 3)	
WordList (phase II)	Generation of Mutual Information (MI) lists (stage 4)	First approach to the combinatorial aspect of terms: identification –thanks to MI scores – of two-word groups habitually collocating in the domain.
Concord	Clusters analysis: polylexical lists (stage 5)	Further analysis with polylexical lists and corroboration of the term character of previously retrieved units by means of contextual approaches. Identification of collocates, collocations and concordance lines. Characterisation of the terms for the subsequent elaboration of dictionary entries.
	Collocates analysis (stage 6)	
	Concordance analysis (stage 7)	

Figure 1: Outline of TEP undertaken with WST.

First of all, however, due to its essential role all through this process, section 1.1 accounts for the composition, size and statistics of the corpus of study, which is the basic, "raw material" for term extraction.

1.1. The corpus of study: composition, size and statistics

The corpus compiled for this study was an electronic, "raw", bilingual (English/Spanish), comparable written corpus of original texts, compiled *ad hoc* and made up of untagged and unmarked running text, that is, just plain running text without any tags or mark-up whatsoever.

Regarding the overall size of the corpus, after the aforementioned conversion to a text-only format, a simple count with WST's WordList tool ("Statistics" tab) revealed, apart from many other, the overall results shown in tables 1 and 3, that is, among other data, a total of 1,498,801 tokens or running words in the text.

The need to compile a balanced and representative corpus is a basic step for any accurate, relevant and succesful TEP despite the fact that the hybridity of some of the texts included with regard to subject matter will probably always cause some problems to

overcome with further research on the topic. Accordingly, as table 1 shows, the English and Spanish parts of the corpus were compiled to have approximately and proportionally a similar and comparable number of texual samples in each area and subarea of the field diagram elaborated, that is, to show an equilibrium in composition.[3]

MAIN AREAS AND SUBAREAS OF THE FIELD DIAGRAM ON INDUSTRIAL CERAMICS	**ENGLISH CORPUS** NUMBER OF TEXTUAL SAMPLES FROM THE AREA // TOTAL NUMBER OF WORDS (TOKENS)	**SPANISH CORPUS** NUMBER OF TEXTUAL SAMPLES FROM THE AREA // TOTAL NUMBER OF WORDS (TOKENS)
1.CHARACTERISATION OF RAW MATERIALS 1.1 Raw materials 1.2 Properties of the raw materials	11 samples // 121,421 tokens	13 samples // 134,003 tokens
2. PRODUCTIVE PROCESSES 2.1 Extraction of raw materials 2.2 Transformation process of raw materials 2.3 Productive processes for obtaining the product 2.4 Commercialisation 2.5 Security measures/occupational health 2.6 Environmental management	24 samples // 287,546 tokens	22 samples // 289,245 tokens
3. END PRODUCT TESTING	16 samples // 143,045 tokens	17 samples // 157,987 tokens
4. APPLICATIONS 4.1 Indoors applications 4.2 Outdoors applications 4.3 Decorative applications 4.4 Maintenance	12 samples // 119,097 tokens	15 samples // 123,876 tokens
5. ORGANISMS AND INSTITUTIONS	6 samples // 35,822 tokens	9 samples // 86,795 tokens
TOTAL COUNTS	**TEXTUAL SAMPLES: 69** **WORDS (tokens): 706,931**	**TEXTUAL SAMPLES: 76** **WORDS (tokens): 791,870**

Table 1: Corpus data regarding balance and composition.

In the same way, it is important that a balance exists regarding the size of the textual samples being part of the corpus. Table 2 presents, broadly speaking, the percentages of individual textual samples in the corpus placed according to the token size intervals shown in the left-hand column.

[3] The field diagram is a conceptual structure of the domain under study that allows every concept to be placed within an ordered, coherent structure or "skeleton". The field diagram allows the balanced inclusion of the necessary number and kind of texts dealing with each specific area or sub-area that go to make up the domain.

Intervals in number of tokens	Percentages of textual samples in the English and Spanish corpora
100-40,000 tokens	38% of the texts in the corpus
40,000-80,000 tokens	27% of the texts in the corpus
80,000-120,000 tokens	19% of the texts in the corpus
120,000-160,000 tokens	16% of the texts in the corpus

Table 2: Percentages of textual samples according to their number of tokens.

In this way, for instance, longer texts have not conditioned the results because a proportionally adequate amount of "long" and shorter texts has been included.

With respect to corpus composition from a genre perspective, the documents compiled in the corpus were mainly research articles from journals, manuals, text books, books on the speciality (mainly monographs), leaflets, webpages, norms/regulations from international organisms in charge of normalisation and standardisation, newspaper articles and magazines, all of them coming from what were considered to be reliable sources.

Apart from the counts presented in the previous tables, the statistical analysis obtained from the tab "Statistics" in the WST application WordList did also indirectly contribute to the quality of the TEP by corroborating or dismissing the suitability of the corpus regarding its specialisation level or size, amongst others.

Among the measures provided by WST that may help us to determine the degree of specialisation of a corpus as a whole (or of the different textual samples included in it), we find two main ratios: the Type/Token ratio (TTR) and the Standardised TTR. The TTR, normally expressed by means of percentages, is obtained by dividing the total number of types (different words in the corpus) by the total number of tokens (total number of words in the corpus). The higher the resulting value, the greater the number of different words contained in a corpus, so that a low figure normally indicates a high degree of repetitions or little variation in terms of vocabulary. This may therefore be interpreted as an indicator of the high level of specialisation of a given text.

Nonetheless, the TTR is sensitive to the extension of the textual samples and thus this ratio is not completely reliable for comparing texts with different sizes in a corpus. However, the Standardised TTR dilutes this influence exerted by extension as much as possible by not taking into account the repetition of the words that appear in other parts of the text, thus resulting in a higher mean value. Nevertheless, these values should only be taken into account for the comparison of texts/corpora with a similar size.

For instance, the statistical results obtained for the overall English and Spanish corpora compiled in this project were the ones shown in table 3.

CORPUS OF INDUSTRIAL CERAMICS	TOKENS (total number of words)	TYPES (different words)	TTR	STTR
ENGLISH CORPUS	706 931	23 470	3.32	40.91
SPANISH CORPUS	791 870	36 743	4.64	42.12
TOTAL	1 498 801			

Table 3: Overall corpus counts and statistical results regarding, tokens, types, TTR and Stnd. TTR.

At first sight, from these resulting TTRs below 5 (and considering the number of tokens of the corpora) it may be deduced that the English corpus is a little more specialised than the Spanish one, even though both of them present a considerable level of specialisation, as

required by a terminographical study.[4] Therefore, the corpora seem suitable for term extraction, but the fact of having compiled reliable texts from specialised areas should be enough to justify the quality of the corpus and its specialised character.[5]

2. The notion and implications of the semi-automatic term-extraction process

Before dealing with the process as such, it is necessary to delimit what is meant by the notion "term extraction process". Taljard and De Schryver (2002) define it as the process whereby computer software is used to automatically detect and extract potential terms from electronic corpora. However, in *all* approaches, humans remain the final arbiters and must decide whether or not the terms suggested by the software do indeed have term status (Taljard and De Schryver 2002: 46).

The retrieval and correct identification of these terminological units (TUs) is the key stage in order to determine which units ought to be included as entry terms and thus further characterised in the registries (prospective dictionary entries) created in the terminological database of the project. Nonetheless, as Cabré (1993) points out, not every single term appearing in the specialised text of a discipline must figure in the terminology that we want to study or deal with. Aims and prospective users are key aspects to be considered throughout the whole process in order to overcome problems or doubts. Moreover, terms do not belong exclusively to the nominal category, as many specialised dictionaries seem to indicate. There are far more categories than nouns in terminology. The stressed nominal character of specialised languages seems to be beyond any doubt (Sager *et al.* 1980, Cabré 1993, Lerat 1995 among many others) because it is a fact that the commonest grammatical category in which terms are given is, by far, that of the noun. However, verbs, adjectives and even adverbs are terminologically relevant, either as monolexical terms or, even more likely, as collocates of other units forming collocations and/or multi-word terms. It is absolutely crucial thus for this collocational behaviour of units to be borne in mind by the terminographer and reflected in any terminographical work.

Thus, there are many issues to be posed, agreed on and systematised prior to and while carrying out a project of this nature. Terms may be monolexical or polylexical; if they are polylexical, segmentation may also constitute a problem if not properly dealt with, especially for those who do not master the domain under study from a cognitive point of view. Moreover, terminological units are specialised *per se* but they may also present different degrees of specialisation or technicality. Thus, it is also necessary to know exactly what kind of terminographical work is to be developed, that is, the scope and "boundaries" of the area and the prospective users of the work. All these aspects are highly relevant for the TEP since they determine the kind of specialised lexical units to be retrieved and the way these units are to be presented in the dictionary.

3. Methodology for semi-automatic term extraction: stages

WST is a software concordance program made up of three applications: WordList, KeyWords and Concord. On the one hand, applications like WordList and KeyWords employ statistical methods that tend to produce large amounts of useless data or "noise". On the other

[4] This ratio below 5 has been proportionally established/considered as an indicator of a high specialisation level according to Edo's (forthcoming) findings on the topic.

[5] The use of the term *corpora* (in plural) aims to highlight the fact that the whole corpus of study is made up of two (sub)corpora, the English and the Spanish one.

hand, Concord, a tool which presents all instances of a lexical item within their immediate co-text also allowing the instances to be sorted in various ways. However, one danger of these text-oriented tools is the potential loss of valid data. In other words, they generate "silence". Therefore, since WST combines both statistical and text-oriented approaches, it is considered a complete, hybrid system requiring, nonetheless, in-depth human dedication in combination with purely automatic means. During term extraction, terminographical and technical knowledge start to merge and a systematic and reliable extraction process implies – apart from taking profit of terminotic tools – hundreds of hours of close observation and reflection on the part of the terminographer. Consequently, terminological extraction was approached here by following a series of stages intended at progressively identifying and analysing prospective terms until it becomes possible to determine whether they really have a terminological nature or not. In this paper in particular, even though the TEP implied the identification and analysis of all the terms of a domain, the focus was placed on the term *abrasion* from the moment it is retrieved by the program until the moment it is identified as a proper term, thus giving rise to multiple collocations and multi-word terms (subentries) derived from it.

Even though this paper focuses on the English part of the TEP, when the terminographical work being carried out involves more than one language to work with – in this case English and Spanish as the languages of the bilingual dictionary under development – a multiple (here, double) TEP is needed and comparison of the results in both languages is highly enriching and necessary.

As is explained in sections 3.1 to 3.4, the basis of the TEP here proposed is to retrieve, first of all, prospective terms which are considered to be so because of frequency and saliency criteria using Wordlist and Keywords. Afterwards, Concord will corroborate or dismiss their term status.

3.1. WordList (phase I)

Broadly speaking, the first approximation to the semi-automatic TEP consisted in using WordList to generate a monolexical frequency list of the words in the corpus in order to obtain a list of potential terms which are candidates – according to the frequency criterion – for confirmation as proper terms at the end of the TEP. However, the semi-automatic TEP with WST starts with a preselection process, the aim of which is to filter or eliminate clearly useless data from the frequency lists.

3.1.1. Preselection (stage 1)

Hence, in the first approximation to semi-automatic term extraction, WordList generates lists of monolexical units arranged either alphabetically or according to the frequency with which they appear in the corpus. In this initial stage, every single lexical unit in the corpus is retrieved by the program and needs to be screened or filtered before proceeding any further. Hence, two lexical filters known as stopword lists – one including English and the other one Spanish functional words (mainly pronouns, demonstratives, articles and prepositions) – were applied to WordList.[6] If not removed, grammatical words tend to occupy top positions in frequency lists and generate "noise", as shown in table 4, which compares the top 15 LUs in the corpus before and after preselection.

In fact, before preselection, in the English corpus we had to wait until position 24 in order to find what at first sight could be considered a potential term or lexical unit that was

[6] English stopword list obtained from: http://www.unine.ch/info/clef/englishST.txt.

directly related with the topic, and it was encouraging that this first non-grammatical word retrieved was *ceramic*. The adjective *ceramic*, as further term extraction has shown, is a basic collocate in the domain and is part of many multi-word units such as *ceramic batch, ceramic fibre, ceramic matrix composite (CMC)* or *ceramic veneer* among many others. In the case of the Spanish corpus, we had to wait till position 28 in order to find a potential term directly related with the topic, in this case, *esmalte* (glaze). This is also a highly expectable result and may be considered a good sign of the adequacy of the corpus. Therefore, frequency lists also provide data which, if correctly interpreted, indicate whether the corpus has been successfully designed with respect to the field under study and if it is comprehensive, representative and balanced enough regarding the subfields of the speciality.

At this point, clearly non-grammatical units were also removed, bearing in mind, however, one of the key aspects of the CTT (Cabré 1999), i.e. that lexical units activate their specialised character or not depending on the context, that is to say, on the communicative situation in which they occur. Accordingly, those lexical units which raised any doubts were left for further analysis in order to determine subsequently whether they had a terminological character or not.

3.1.2. Generation of monolexical frequency lists (stage 2)

After preselection, following general practice in the field of semi-automatic term extraction, the first step was to extract single-word terms computationally. Thus, after the preselection phase, monolexical frequency lists free from grammatical words were generated by WordList. This TEP sets out from the logical hypothesis that the lexical units with the highest probability of being terms are characterised (due to their representativeness of the field under study) by a high frequency of appearance in discourse. As sections 1 and 1.1 try to demonstrate, frequency is also a relative criterion, as it depends very much on the nature of the texts included in the corpus and on the number of texts on specific subfields included within it, that is to say, on the balance that is struck regarding the number of documents compiled on each sub-area. However, other criteria apart from frequency must be considered. That is why a list of keywords revealing saliency/keyness has been employed as the best complement to frequency data in these initial stages (see section 3.2). This is so since it may be the case that some terms which are of interest to and representative of the field of study may appear with a frequency of 1 and this phenomenon can be understood as being triggered by a series of circumstances such as the aforementioned nature of the corpus, its equilibrium or its size. A word that occurs only once in a single text or corpus is known as a *hapax legomenon* and, contrary to what could be thought because of its low frequency, these words should not be discarded without further consideration. In fact, hapaxes generally represent about 40% of the words in a corpus (Lardilleux and Lepage 2007). In this line, and using a corpus comparison method, Chung (2003) found that an important source of technical terms was those words which occurred only in the technical corpus even with a very low frequency. In the particular case of our English corpus on ceramics, 9,278 lexical units were instances of this linguistic phenomenon of hapax legomenon, including relevant terms such as *abrade*, *isotropy*, *xerography* and *zeolite*. Accordingly, not only high frequency items must be considered for term retrieval; other criteria must complement frequency results, which are, however, a good and logic starting point.

The top-frequency prospective TUs appearing in the monolexical lists generated after preselection are a series of units that, at least presumably, would be considered as prototypical of the ceramic industry domain in both languages. When the first unique *terms* on the WordList frequency list (or the ones on the KeyWord list) are observed, generally speaking it can be noticed that the core terminology of the specialised field at hand has been identified. These top-frequency units are the ones that would most likely come first to any

person's mind when industrial ceramics is suggested because they are among the terms considered to be central in the speciality. In addition to this, it is crucial to notice the combinatorial aspect of terms by observing how these highly-frequent, central units very often act as the "bases/nodes" of polylexical or multi-word terms or as collocates of other nodes. Following Ahmad and Rogers (2001), it could be said that these highly frequent monolexical units are the "mother terms" of a given speciality and as such they "engender" other terms through valid processes of formation and combination. For instance, the monolexical term *abrasion*, which this paper focuses on, is also the base of the polylexical term *abrasion hardness* and *temperature* is (as further research revealed) the base of *annealing temperature*, among many other cases. If table 4 is observed, the importance of a good preselection stage seems beyond all doubt and proves to save a lot of the terminographer's time and energy, apart from adding reliability to the analysis because there are fewer factors distracting his/her attention.

	ENGLISH *Before preselection*	***ENGLISH*** *After preselection*	***SPANISH*** *Before preselection*	***SPANISH*** *After preselection*
1	the	ceramic	de	esmalte
2	#	glass	la	cerámica
3	of	temperature	en	esmaltes
4	a	high	el	agua
5	in	materials	y	horno
6	to	surface	#	arcilla
7	is	material	que	temperatura
8	for	ceramics	se	pastas
9	are	tiles	a	óxido
10	by	tile	los	cocción
11	or	glaze	las	pasta
12	as	process	o	forma
13	be	clay	con	color
14	with	properties	del	baldosas
15	that	size	para	pieza

Table 4: 15 top-frequency lexical units in both corpora before and after preselection.

Of the monolexical units retrieved by the program after preselection, none of the LUs shown above would be refuted at first sight as non-terms. Hence, all these LUs which may be terms must be subjected to further analysis (further stages in the TEP) before they can finally be considered as terminological units (or not).

Additionally, WST offers the possibility of reducing these lists and meaningfully organising them by semi-automatically lemmatising items both in English and Spanish (apart from many other languages) so that in the lemmatised display, units with the same stem are grouped under the same lemma; for instance, *firing, fires* and *fired* are grouped under the lemma *fire*, that is, under a single position. With lemmatisation, the TEP may be meaningful and ordered for the terminographer since he/she can gather under the same lemma the spelling variant and morphological forms of the same word.

Apart from their importance as such, wordlists are also a necessary preliminary stage for generating keyword lists since these are obtained after WST compares the study corpus wordlist and a reference, general corpus wordlist. Therefore, this TEP departs from the idea that diverse techniques applied at progressive stages of a well-designed TEP (see figure 1) provide different but complementary data about the kind of terminological units to be found in a corpus. This TEP thus relies on: data obtained from frequency approaches (Wordlist), from saliency aprroaches (KeyWords) and from textual/contextual approaches (Concord).

Additionally, for the correct identification of relevant hapaxes and for the success of the TEP as a whole, human expertise in terminographical issues and in the field are critical aspects.

3.2. KeyWords: generation of saliency lists (stage 3)

As a matter of fact – and without dismissing their usefulness in any way – it is clear that a "simple" frequency list (even after preselection) tends to over-generate, i.e. to identify items which are not terms relevant to the specific subject field. As Taljard and De Schryver (2002) concede, reading through *top-frequency words* is obviously an unrefined procedure. In the TEP described in this paper, the best option was considered to be to combine this sort of unrefined but highly useful frequency approach offered by WordList with a more refined one – KeyWords – based on probabilistic calculus performed by the program in order to retrieve keywords, that is, words that are highly likely to be terms. Scott (1997: 236) defines the term *keyword* as "a *word which occurs with unusual frequency in a given text*". As Taljard and De Schryver (2002) go on to state, unusual frequency can be related to outstandingness and implies that a word has an unusually high (or unusually low) frequency in a text (or sub-corpus) in comparison to its occurrence in a *reference corpus* of some kind.

KeyWords thus offers a kind of complementary analysis to that offered by WordList frequency lists, since the former yields items with outstanding frequencies and not "top frequencies", as WordList does. Hence, the joint analysis of the results obtained with both applications results in a complete list of prospective terms, grounded both on the basis of their frequency and on their specialised, infrequent character if compared to the frequency wordlist of a general reference corpus. A keyword list therefore gives a measure of *saliency*, whereas a simple word list only provides *frequency* (Baker 2006). Considering both of them together gives as a result a list of lexical units that are highly likely to be terms and confirms/dismisses the preliminary term extraction obtained by means of frequency.

The results obtained with KeyWords, as Figure 6 shows, were a list of keywords, or words whose frequencies are statistically higher in the study corpus than in the reference corpus. As Berber Sardinha (2000) concedes, the software also identifies words whose frequencies are statistically lower in the study corpus, which are called "negative keywords", in contrast to positive keywords, which have higher frequencies in the study corpus. However, negative keywords will not be discussed in this paper and whenever a keyword is mentioned here it implies a "positive" value. Hence, a word will be a keyword if its frequency is either unusually high or unusually low in comparison to a reference corpus (Berber Sardinha 1999). In the case of this study, the reference corpus was the British National Corpus (a reference corpus of 100 million words of written and spoken general British English).[7]

[7] The downloadable BNC (English) word list was obtained from: http://www.lexically.net/downloads/version4/downloading%20BNC.htm.

	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	#	32.260	4,56	1.604.421	1,61	25.709,23	0,0000000000
2	CERAMIC	2.698	0,38	380		24.440,81	0,0000000000
3	CERAMICS	1.568	0,22	347		13.728,39	0,0000000000
4	GLAZE	1.435	0,20	162		13.171,73	0,0000000000
5	TILE	1.465	0,21	391		12.609,57	0,0000000000
6	TEMPERATURE	1.951	0,28	4.343		11.600,10	0,0000000000
7	TILES	1.478	0,21	1.155		11.049,91	0,0000000000
8	GLASS	2.202	0,31	9.352		10.696,56	0,0000000000
9	CLAY	1.357	0,19	1.588		9.402,83	0,0000000000
10	MATERIALS	1.799	0,25	6.753		9.121,84	0,0000000000
11	THERMAL	1.119	0,16	679		8.712,27	0,0000000000
12	REFRACTORY	909	0,13	69		8.507,67	0,0000000000
13	KILN	901	0,13	177		7.966,06	0,0000000000
14	SURFACE	1.719	0,24	8.898		7.755,27	0,0000000000
15	OXIDE	893	0,13	401		7.250,74	0,0000000000

Figure 2: Top 15 keywords as identified by KeyWords and compared to frequency lists.

The column "keyness" assigns a *keyness* value to each word; the higher the score, the stronger the keyness of that word, whereas the final column gives the p value of each word. As p is set so low here ($p < 0.000001$), almost all of the figures in this column are 0.000000. After comparing the 250 top-frequency LUs from the study corpus with the top keywords also retrieved by WST, a significant number of coincidences were noticed (89% of the 250 top-frequency units also appeared among the top 250 keywords) and a huge list of relevant terms for subsequent collocational analysis with Concord was obtained. Figure 2 shows the top 15 keywords retrieved by the program; the terms in the red boxes also appear among the15 top-frequency LUs and this comparison approach was the way the joint list of prospective monolexical terms was obtained.

In the specific case of *abrasion*, the term appears in KeyWords in position 125. In our corpus it occurs 227 times, compared to an occurrence of 71 times in the bigger reference corpus. Proportionally, however, its frequency is many times higher in the smaller corpus than in the 100 million-word reference corpus, thus constituting a keyword (a term) with a keyness value of 2,040.

Therefore, up to the moment we have a list of monolexical LUs (prospective terms) elaborated by the terminographer on the basis of both frequency and saliency lists generated by WST.

3.3. WordList (phase II)

At this point, the combinatorial aspect of terms must start to be considered; firstly, with the analysis of 2-word polylexical lists generated by WST on the basis of MI scores.

3.3.1. Generation of Mutual Information (MI) lists (stage 4)

Mutual information balances and contrasts the probability of two words occurring mutually joined with the probability of these words occurring independently. Apart from monolexical lists accounting for individual frequency of prospective terms and providing a first approximate step for term selection, WordList also offers the possibility of resorting to listings that can "measure" the strength or degree of mutual dependence – in this case measured in terms of Mutual Information (MI) – of LUs forming potential collocations.

In layman's terms, these listings are generated to detect the "strength" or the degree of mutual dependence of words that tend to appear together or in a close position, the ruling

principle being that the word that appears more frequently with another word than in any other position within the corpus tends to be considered a significant combination and therefore deserves special consideration. These "mutually dependent" LUs can be said to "collocate" or to appear in habitual company.

This option thereby constituted the first approach to the combinatorial behaviour of potential terms and is the first tool in the proposed TEP to provide explicit data on possible multi-word terms or collocations. Table 5 shows a small section of the Excel version of the huge MI list obtained for the English corpus and part of the information regarding the top-frequency term *abrasion* and its "degree of dependence" with respect to the most significant LUs (collocates) appearing nearby. The table shows, for instance, revealing MI scores for some potential collocations/multi-word terms such as a*brasion test*, *abrasion hardness*, *abrasion resistance*, *tiles resistant to abrasion*, *chemical abrasion*, and so forth, which further analysis with Concord will confirm or reject as real terms. These are all combinations that the program identifies as habitually collocating in the domain of industrial ceramics.

Word 1	Word 2	Texts	Gap	Joint	MI
ABRASION	CERAMIC	11	4	19	4,653893
ABRASION	TILES	11	3	46	6,7654595
ABRASION	TEST	3	1	7	4,9023805
ABRASION	HARDNESS	1	1	5	5,7552519
ABRASION	GLAZED	11	2	25	8,0471678
ABRASION	CHEMICAL	3	3	10	5,6813598
ABRASION	FROST	1	3	7	8,4501324
ABRASION	UNGLAZED	9	2	14	8,2981291
ABRASION	RESISTANCE	16	1	60	8,20784
ABRASION	ABRASION	6	2	20	8,5800419
ABRASION	CLASS	10	1	12	8,5647745
ABRASION	REMOVED	5	4	5	6,7683082
ABRASION	RESISTANT	5	1	12	7,9630365

Table 5: Adaptation of the English MI list with a focus on the part devoted to the term abrasion.

In the display, the column *Word 1* refers to the first word in a pair; *Word 2* makes reference to the other word in that pair. If you have computed "to right only" (as in this case), then Word 1 precedes Word 2. The column *Texts* indicates the number of texts this pair was found in, whereas *Gap* specifies the most typical distance between Word 1 and Word 2. Finally, *Joint* provides their joint frequency over the entire span (not just the joint frequency at the typical gap distance).

In the specific case shown here, the minimum number which the MI must come up with to be reported was set to be 3.0 because "below this, the linkage between node and collocate is likely to be rather tenuous" (http://www.lexically.net/downloads/version5/HTML/?wordlist_overview.htm). In the same way, the minimum frequency for any item to be considered for the mutual information calculation was the one established by default, that is, 5. Finally, the span – or the number of intervening words between collocate and node – was set 5.

Thanks to MI lists, the terminographer may "discover", for instance, that *abrasion test, abrasion hardness, abrasion resistance, abrasion class* and *abrasion(-)resistant* – just to mention the most striking ones at first sight from Table 7 – are relevant combinations in the field. Afterwards, further/complementary analysis of these collocations in Concord (section 3.4) showed, for instance, that there is a specific abrasion test called *Capon abrasion test*, that *abrasion resistance* tends to be qualified by adjectives such as *high* or *deep*, that *abrasion resistance* tends to concern *glazed tiles/ floor tiles / refractory bricks/ a surface* or/and

concrete aggregate, that *abrasion class* in our corpus is always part of the syntagm *abrasion class for glazed tile* or, finally, that *abrasion-resistant* is a common collocate in the domain for the nodes *ceramics / parts / materials / components* and *ceramic products*.

Thus, from MI lists, the terminographer may start to determine whether the co-occurrence of a node and a collocate (s), that is to say, a potential collocation, is purely by chance or statistically significant. However, MI unduly privileges low frequency words (Heid 2003) so it may be misleading in the case of words with a very low frequency: if a low-frequency word occurs only with another word, it will receive a high MI score in spite of its low frequency. Concord analysis will help to overcome this problem by offering a complementary contextual analysis of possible collocations.

3.4. Concord

Concord is the pure concordance application of WST and the one in charge of generating, among other data, clusters (in this case 3+ polylexical lists), collocates, lists of concordance lines (also known as *Key Word in Context* – KWIC – lists) and patterns of the prospective terms obtained in the previous WordList stage. It also shows the source text and identifies the names of the files where terms appear.

The analysis generated by Concord provides evidence on the way terms are used in real communication amongst specialists. Concord allows the terminographer to penetrate deep into the collocational behaviour of terms since, as Gilquin (2002: 210) states in the case of grammatical patterns, "such phenomena are much more difficult to extract from a corpus than simple words or tags". Their identification and retrieval is, however, absolutely fundamental in order to accurately account for the communicative, real aspect of the terminology of a domain.

It is in this light that the behaviour of terms in context (i.e. the way they collocate and are used in real discourse) started to be analysed. At the same time, work also began on their segmentation (in the case of multi-word terms), the adequacy of including them as dictionary entries or the selection of examples of use from the corpus to be included in the entries. This observation and analysis of terms in context thus constituted the final step in order to understand whether the frequency of the prospective TUs retrieved with WordList and the key character of the units signalled by KeyWords were real indicators of their terminological nature.

Concord applications are "Concordance", "Collocates", "Plot", "Patterns", "Clusters", "Filenames", "Source text" and "Notes". Following on with the research outlined above, the work done by Concord in this TEP continues to be illustrated here with the study of the English TU *abrasion*, which occupies position 339 on the monolexical frequency list generated by WordList after preselection and has a frequency of 227 and a keyness value of 2.040 in the KeyWord list.

3.4.1. Clusters analysis: polylexical lists (stage 5) [8]

In the fourth stage of the method, two-word combinations obtained on the basis of MI scores have been obtained. In the fifth one, 3/4/5-word lists were generated. These multi-word groupings are called clusters (by WST) or bundles . As YouJin (2009) states lexical bundles are defined as the most common recurrent sequences of words in a register and as such they should be regarded as a basic linguistic construct with important functions for the construction of discourse for different languages.

[8] Sections 3.4.1, 3.4.2 and 3.4.3 have been labelled according to the different applications from WST- Concord used in this study.

Clusters represent a tighter relationship than collocates, more like groups or phrases, and constitute a good starting point for the analysis of the phraseology of a domain. WST gives the terminographer two opportunities for identifying word clusters, in WordList and Concord. Both use the same method but Concord only processes concordance lines, while WordList processes whole texts.

In this TEP, clusters lists generated with Concord and containing lexical bundles were employed mainly in conjunction with the collocates list. This allowed frequent multi-word fixed patterns to be more easily identified and verified whereas the MI lists were mainly used to tell us about connections between 2 words.

"Clusters" displays a list of fixed sequences of recurrent words in the concordance (understood as the entire list of concordance lines). However, one must be careful with this option because Clusters are based on searching within the collocational horizons established by the terminographer (the default horizons are 5L to 5R), which may be made smaller if required. This last point is especially important because Cluster looks for the items repeated in the concordance without limiting itself to the parts containing the search word so, if the Horizons are too big, it may be the case that the clusters retrieved contain the search word or not. In this case, the terminographer can "force" the clusters to contain the search words if the horizons are made small enough. For instance, for the specific case of the term *abrasion*, the most significant clusters obtained with WST with different horizons have been summarised in table 6.

CLUSTER SETTINGS	SELECTION OF CLUSTERS OBTAINED
Words in cluster: 6 to 6 Minimum frequency: 5 Horizons: 5L, 5R	*Determination of resistance to deep abrasion /surface abrasion*
Words in cluster: 5 to 5 Minimum frequency: 5 Horizons: 5L, 5R	*Resistance to surface abrasion of…*
Words in cluster: 3 to 3 Minimum frequency: 5 Horizons: 3L, 3R	*Resistance to abrasión // Abrasion of glazed… // Deep abrasion of…// Tiles abrasion resistance // Abrasion class for…// (Un)Glazed tiles abrasion* *Surface abrasion of…*

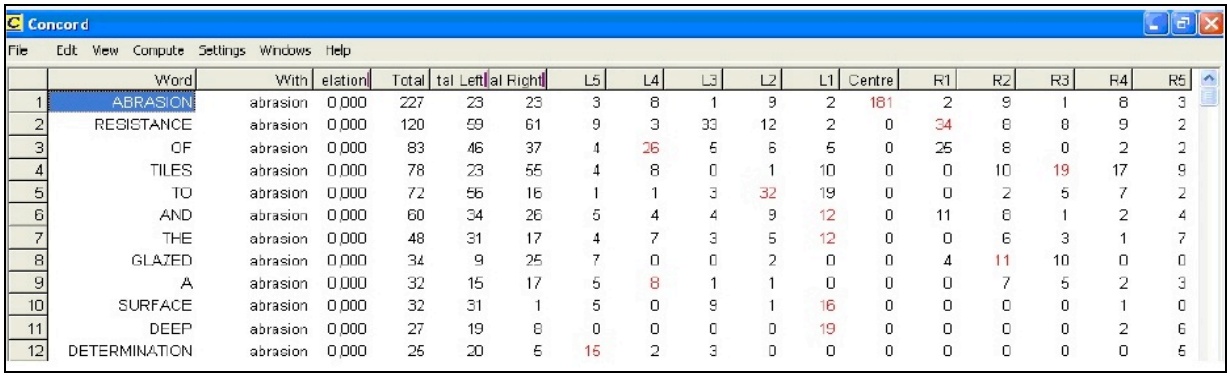
Table 6: Selection of meaningful clusters obtained for abrasion with different cluster settings.

From these results, the terminographer may start to corroborate that clusters with the node *abrasion* and collocates such as *resistance, deep, tile* and *surface* are key in the domain. Further stages in the TEP will corroborate or dismiss this information.

3.4.2. Collocates analysis (stage 6)

The next of the Concord tools to be used in this TEP was "Collocates". This application provides a list of the prospective collocates that appear in the immediate context of the search word (node), arranged according to the frequency with which they appear and showing the frequency with which they occur in each position. As may be observed in Figure 9, the collocates that appear, for instance, first on the left with respect to the search word – occupying position *centre* – are represented in column L1 by the frequency with which they appear. The same happens with column R1 (first on the right) and so forth. "Collocates" also highlights in red the highest frequency – and thus the commonest position – of each collocate retrieved with respect to the node, while also displaying the whole list of collocates according to their overall position. Corroborating the data in stage 5, Figure 3 shows how the search word *abrasion* has *resistance* as its most frequent collocate (with a total of 120 instances) and that the most frequent position of this collocate with respect to the node is R1 (first word on

the right). By double clicking number 34, highlighted in red and representing position R1 of *resistance*, the concordance lines from the study corpus showing *abrasion resistance* in context are displayed, thereby corroborating the need to consider it a relevant multi-word term and a subentry of the main term *abrasion.*



Concord

File Edit View Compute Settings Windows Help

	Word	With	elation	Total	tal Left	al Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	ABRASION	abrasion	0.000	227	23	23	3	8	1	9	2	181	2	9	1	8	3
2	RESISTANCE	abrasion	0.000	120	59	61	9	3	33	12	2	0	34	6	8	9	2
3	OF	abrasion	0.000	83	46	37	4	26	5	6	5	0	25	8	0	2	2
4	TILES	abrasion	0.000	78	23	55	4	8	0	1	10	0	0	10	19	17	9
5	TO	abrasion	0.000	72	56	16	1	1	3	32	19	0	0	2	5	7	2
6	AND	abrasion	0.000	60	34	26	5	4	4	9	12	0	11	6	1	2	4
7	THE	abrasion	0.000	48	31	17	4	7	3	5	12	0	0	6	3	1	7
8	GLAZED	abrasion	0.000	34	9	25	7	0	0	2	0	0	4	11	10	0	0
9	A	abrasion	0.000	32	15	17	5	8	1	1	0	0	0	7	5	2	3
10	SURFACE	abrasion	0.000	32	31	1	5	0	9	1	16	0	0	0	0	1	0
11	DEEP	abrasion	0.000	27	19	8	0	0	0	0	19	0	0	0	0	2	6
12	DETERMINATION	abrasion	0.000	25	20	5	15	2	3	0	0	0	0	0	0	0	5

Figure 3: Top 12 collocates of the term abrasion provided by "Collocates" in Concord

An important issue at this point is the notion of collocation and whether words that tend to collocate should be considered as multi-word terms that deserve a dictionary entry in their own right or as collocations in the broadest sense of the word. The term *collocation* was first introduced by Firth (1957: 14) and was defined as "actual words in habitual company", but it cannot simply be assumed that all collocations showing up in concordance lines or on the Collocates tab are multi-word terms. In Taljard and de Schryver's (2002: 59) words:

> "The terminological status of the *term: collocate(s)* combination depends on whether the combination of a term and its collocate(s) can be seen as the denomination of a new concept in its own right. If this is the case, such a collocational combination will qualify as a multi-word term; if not, it would be described as a false positive. In some cases, false positives in a concordance are quite obvious and easily identifiable, whereas others seem to be on the borderline between multi-word terms on the one hand, and simple collocations on the other".

It may be stated that a lexical combination should display a series of signs which, although they are not equally productive, shed some light on the segmentation of dubious-to-delimit units. These signs act as proof of the presence of a multi-word term (single conceptual unit) and include: the fact that a whole is lexically organised around a single basis; the impossibility of inserting other linguistic elements within the terminological syntagma; the impossibility of separately complementing any of the parts of the whole; the fact of being able to substitute the whole by a synonym; the fact of having an antonym in the same speciality; the frequency with which a terminological syntagma appears in the text of a given speciality; and the fact that in other languages the syntagma is a unique lexematic unit (Cabré: 1993).

From the data obtained with collocates, important multi-word terms (prospective subentries of the term *abrasion*) were preliminarily identified, namely: *abrasion finish, abrasion by sandblast, abrasion resistance classification, abrasion resistance test, abrasion hardness* and *abrasion hardness test* amongst others.

Likewise, the application "Patterns" performs, in general terms, a similar function to that of the tool "Collocates". However, the kind of data and the way of arranging and displaying them are different. "Patterns" will show the words adjacent to the search word, organised in terms of frequency within each column. That is, the top word in each column is the word most frequently found in that position, the second word is the second most frequent and so on. The result is to make the most frequent items in the neighbourhood of the search

word “float up” to the top. Like “Collocates”, this helps you to detect lexical patterns in the concordance. In the specific case of the TEP here presented, “Collocates” has been the preferred option for considering it more visual and complete than other Concord applications performing similar functions.

3.4.3. Concordance analysis (stage 7)

In order to proceed with the TEP, “Concordance” provides concordance lines containing the prospective TU highlighted together with its co-text or surrounding text. These are KWIC lists, which can be sorted. Sorting is another interesting possibility offered by WST in order to display the results in different ways, depending on the kind of research/analysis to be conducted. Sorting is especially interesting with the tool Concordance because by means of sorting the program may, for instance, display the data retrieved in different ways, either highlighting different locations such as L1 or R2. The point of sorting is also to find characteristic lexical patterns. It can be hard to see overall trends in concordance lines, especially if there are lots of them. In the TEP here proposed, you can sort the concordance lines retrieved, separate out multiple search words and examine the immediate context to left and right. Sorting by the words in the immediate co-text of the search word (especially L2, L1 and R1, R2) is a way to start detecting collocations and multi-word units. Hence, single-word mother terms or prospective multi-word terms already detected in previous stages are used in KWIC searches (with or without sorting) to show the way these terms collocate and thus corroborate their term status.

By simply observing, for instance, the LU *abrasion* in a concordance like the one in Figure 10, potential multi-word terms and collocations may be identified at first sight and corroborated, as in this case: *abrasion by sandblast, abrasion finish, abrasion resistance, abrasion hardness, abrasion resistance classification, abrasion resistance test, abrasion test,* and so forth. It may also be observed that the term *abrasion* frequently collocates with adjectives such as *severe, deep, heavy* or *serious*, with verbs such as *resist, prevent* or *assess* and with other nouns such as the ones previously mentioned (forming multi-word terms) or others such as *surface*.

Additionally, when a search word is introduced in Concord, it is normally with the aim of observing and analysing its behaviour in context as well as the way it combines with other lexical units. In this sense, the option called “Context word(s) and context search horizons” (shortened here to “Search Horizons”) allows the terminographer to undertake specific searches regarding how words that are likely to collocate appear in the corpus.

For instance, figure 4 presents a display of the Concordance-Search Horizons type. In this case, the “Search Horizons” option has been activated for *abrasion* and *resistance* and results seem to corroborate that *abrasion* and *resistance* tend to collocate. In the same way, frequent collocations such as *resistance to deep/surface abrasion of glazed/unglazed tiles* (already detected in previous stages and also appearing here) show the kind of habitual company that the multi-word term *abrasion resistance* carries with it.

89	- Determination of resistance to surface abrasion - Gla zed tiles" "Ceramic tiles	213	378%	016%
90	- Determination of resistance to deep abrasion - Unglazed tiles" "Ceramic	154	339%	011%
91	the appearance of a major area of tiles Abrasion resistance (a) Resistance to.	1,031	2737%	076%
92	volume in mm3 (b) _Resistance to abrasion of glazed tiles. Class I - IV	1,048	2795%	077%
93	resistance (a) Resistance to. deep abrasion of unglazed tiles: removed	1,037	2777%	076%

Figure 4: Results retrieved by “Concordance” with the option “Search Horizons” activated for abrasion and resistance.

To sum up what we have up to the moment: from the data obtained with “Collocates”, it may be concluded that *resistance* is the most frequent collocate of the term *abrasion*, as “Clusters”

and “Concordance” also corroborate. As such, they appear either together or in a close position a total of 120 times. If we combine the three kinds of data retrieved from the program so far, it can be observed that with “Collocates” the collocate *resistance* appears with a frequency of 33 in position L3 (third on the left), and *abrasion* (centre) and *resistance* (R1) also collocate in positions other than *abrasion resistance*. When further research in this respect is carried out, we can observe that the collocate *resistance* occupying position L3 with respect to the node *abrasion* gives rise to concordance lines such as the ones shown in figure 5:

Figure 5: Further concordance lines generated with “Search Horizons” for abrasion and resistance (in position L3 with respect to the node).

From these concordance lines obtained with “Search Horizons”, the existence of the collocations *resistance to surface abrasion* and *resistance to deep abrasion* (see table 7) is revealed. At the same time, the tendency of *resistance* to collocate in position L3 with respect to the node is also corroborated:

Resistance	*to*	*surface*	*abrasion*
Resistance	*to*	*deep*	*abrasion*
L3	L2	L1	CENTRE

Table 7: Immediate “Search Horizons” for abrasion and resistance.

However, the preferred position of the collocate *resistance* with respect to the node *abrasion* in the English corpus is the aforementioned R1 (first on the right), with a total of 34 instances, as concordance evidence shows.

Additionally, the option “Search Horizons” also offers the possibility of looking for and retrieving inflected forms of the collocations detected. Among the TUs detected in this way in our corpus we find *abrasion-resistant ceramics*, *abrasion-resistant materials*, *abrasion-resistant components*, *abrasion-resistant ceramic products*, and so on. The same happened when the form *abras** was introduced, since *abrasive* compounds were also detected (see section 4 “Results”).

In the same way, the terminographer should also make the most of existing resources. Lists of already identified terms obtained from previous terminographical works are also very useful for term extraction since they help to make the work easier and corroborate the information that is obtained.

With these stages, current resources and the common sense of the terminographer, he/she will be able to identify, retrieve from the corpus and characterise the terms that shape the speciality field under study.

4. Results

Although the TEP proposed here has already retrieved the thousands of terms to be included in the prospective dictionary on industrial ceramics (more specifically 24,000), this method has been illustrated here mainly by the TU *abrasion* in order to depict the “path” that terms have followed in this research before actually becoming dictionary entries. As such, this term

has proved to be the base of many other multi-word terms that the TEP has allowed us to retrieve and discover, and which also deserve to have a dictionary subentry. In order to present this process as a methodological sequencing of stages leading to a final specific result, Figure 6 shows the final dictionary entry of the term *abrasion*, in which a semasiological organisation based on form has been employed (as well as in the rest of dictionary entries). It can be noticed how, in accordance with the theoretical principles underlying our work, this entry includes: a main entry term and subentry terms resulting from the analysis of the collocational nature of terms; real contexts illustrating the way terms are used; the semantic field for each entry and subentry; equivalents, which are so necessary for a bilingual translation tool; definitions including notes on the usage; and the part of speech. As well as for term extraction, most of the data regarding the combinatorial aspect of terms has also been used for the elaboration of this or any other entry in the dictionary.

abrasion *n:* PROPQUIM-FIS abrasión, desgaste por fricción/rozamiento; desgaste *-wear-* o erosión *-erosion-* causado/a en una superficie por una acción continua y producido por fricción *-friction-*, por impacto *-impact-* o por agentes erosivos *-erosive agents-* como el viento, la lluvia etc. durante largos periodos de tiempo ◊ *Among the advantages of ceramics tile are an ability to withstand damage from heat, and resistance to abrasion*; V. *corrosion; wear; erosion.* [Exp: **abrasion by sandblast/sand-blast** (PROPQUIM-FIS abrasión mediante chorro de arena ◊ *Physical properties were determined on specimens prepared under laboratory conditions using applicable ASTM procedures and showed, for instance, an excellent resistance to abrasion by sandblast*; V. *sand-blast*), **abrasion finish** (ELABPROC acabado con abrasivos; proceso de eliminación de las rebabas de los objetos moldeados y/o deslustrado de sus superficies, sometiéndolas al impacto de materias como huesos de albaricoque machacados, cáscaras de nuez o gránulos de plástico, con suficiente fuerza como para fracturar la rebaba V. *surface finish; sand blasted finish*), **abrasion hardness** (PROPQUIM-FIS V. *abrasion resistance*), **abrasion resistance** (PROPQUIM-FIS resistencia a la abrasión, dureza a la abrasión; resistencia al desgaste [por rozamiento]; propiedad que presenta una superficie a la hora de resistir el desgaste *-wearing-* producido por frotamiento *-rubbing-* con un material extraño que puede producir la erosión de dicha superficie ◊ *Abrasion resistance is determined by abrasion tests, and tiles are grouped accordingly*), **abrasion resistance classification** (ENSAYO/CALIDAD clasificación de resistencia a la abrasión; clasificación en la que se determina la resistencia a la abrasión *-abrasion resistance-* de un producto/material; la clasificación de baldosas esmaltadas para piso según su resistencia a la abrasión es: CLASE 1 (PEI I): uso individual ligero como cuartos de baño doméstico y dormitorios sin acceso directo desde el exterior; CLASE 2 (PEI II): uso individual normal como cualquier zona de vivienda particular a excepción de cocinas y entradas; CLASE 3 (PEI III): uso individual elevado o uso colectivo moderado como todas las zonas de una vivienda privada; CLASE 4 (PEI IV): uso colectivo normal como cocinas, restaurantes, exposiciones, boutiques, peluquerías; CLASE 5 (PEI V): uso colectivo elevado como centros comerciales, bares, tiendas con mucho tránsito, zonas peatonales y aplicaciones industriales V. *PEI rating; resistance to surface abrasion; abrasion resistance test*), **abrasion resistance test** (ENSAYO ensayo de resistencia a la abrasión, ensayo de abrasión; test de resistencia al desgaste; ensayo consistente en someter a la loseta cerámica sobre la parte vidriada a una acción abrasiva compuesta de una mezcla de esferas de acero, arena de Corindón y agua destilada; a este tipo de prueba se le conoce también como prueba de P.E.I. y da lugar a una clasificación de los azulejos en 5 grupos V. *abrasion resistance classification; PEI rating; resistance to surface abrasion*), **abrasion test** (ENSAYO ensayo de abrasión V. *abrasion resistance test*), **abrasion/abrasive hardness, HA** (PROPQUIM-FIS dureza a la abrasión, resistencia al desgaste [por rozamiento]; propiedad o grado de resistencia que presenta un material al desgaste por abrasión ◊ *Granite is a well known building stone and has high abrasion hardness, with very high resistance to weathering and extreme resistance to chemical attack*)].

Figure 6: Final dictionary entry of the term "*abrasion*" *with its corresponding subentries.*

In a similar way, for the entry term a*brasive* many subentries in the form of multi-word terms have been retrieved following the same TEP, namely *abrasive action, abrasive agent, abrasive bead, abrasive belt/band, abrasive blade, abrasive blasting, abrasive charge, abrasive cleaning, abrasive cloth, abrasive collector, abrasive disk, abrasive finish, abrasive flow, abrasive grain, abrasive grinder, abrasive machining, abrasive rock, abrasive sand, abrasive slurry, abrasive strength, abrasive substance, abrasive tool* and *abrasive wear.*

It is also true that terminographical and lexicographical works are deeply dependent on space and economic limitations and that some multi-word terms have been left out because of being too self-revealing. However, above anything else, this kind of TEP and the final entries generated through it try to remain faithful to Firth's (1935: 37) words when he said that "the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously".

5. Conclusion

In a long-term terminographical project like the one this research is part of, the TEP is conceived as the stage dealing with the recognition, delimitation and retrieval from the corpus of all the segments of language that can be considered to be terms belonging to the speciality field in question. The raw material for the TEP is the corpus itself. Hence, this TEP aims to identify characters or chains of characters that could be potential terms so that subsequently they can be analysed in context in order to confirm or reject their "term status" in real use together with their collocational behaviour.

This paper, then, has attempted to illustrate how the recognition/detection, delimitation and retrieval of prospective terms that are worthy of becoming dictionary entries or significant collocations (to be reflected as, for instance, examples of use in such dictionary) can be systematised by means of a corpus-based process of semi-automatic term extraction, in which both human and technological means (WST) are necessarily employed.

The use of data generated by corpus-query tools for term extraction allows the terminographer to undertake the most time-consuming operations of the TEP quickly. Furthermore, these data are also the key to correctly including the linguistic information about each term in the terminographical database created in the sixth stage of the overall dictionary-making process (data processing). If, thanks to sequential and systematic terminographical work, terms are correctly identified and characterised with respect to their function in the text, the semantic field to which they belong, the way they behave in context and the kind of words that normally appear in their company, then the resulting dictionary entries will be (at least presumably) what the prospective user of the dictionary needs and expects to find in it.

However, human beings will always remain the final judges in any terminological activity. The terms retrieved by the software will always need to be scrutinised by a terminologist and, even though *terminotic* applications such as WST have meant an enormous advance in corpora exploitation and terminological analysis, human means are "still" the key with which to analyse the data obtained and to confirm or reject their validity for the aims posed at the outset.

6. References

Ahmad, K. and M. Rogers. 2001. Corpus Linguistics and Terminology Extraction. In S. Wright and G. Budin (eds.), *Handbook of Terminology Management. Vol*. 2. 725-760. John Benjamins Publishing Company.

Auger, P. and L.J. Rousseau. 1987. *Metodologia de la recerca terminològica*. Sant Adrià de Besòs: Departament de Cultura de la Generalitat de Catalunya.

Baker, P. 2006. "The question is, how cruel is it?" Keywords, in debates on Foxhunting in the House of Commons. *Paper delivered in the Methods Network Expert Seminar on Linguistics*. Lancaster University, UK.

Berber Sardinha, T. 1999. Using Key Words in Text Analysis: practical aspects. *DIRECT Papers* 42: 1-9.

Berber Sardinha, T. 2000 Comparing corpora with WordSmith Tools: How large must the reference corpus be? In A. Kilgarriff and T. Berber Sardinha (eds.), *Proceedings of The Workshop on Comparing Corpora* 9: 7-13. Stroudsburg, PA: Association of Computational Linguistics.

Bergenholtz, H. and S. Tarp. 1995 *Manual of Specialised Lexicograpy: The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.

Cabré Castellví, M.T. 1993. *La Terminología : Teoría, metodología, aplicaciones*. Empúries: Editorial Antártida.

Cabré Castellví, M.T. 1999. *La terminología, representación y comunicación: elementos para una teoría de base comunicativa y otros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada (IULA), Universitat Pompeu Fabra.

Cabré Castellví, M.T. and R. Estopà. 2002. El conocimiento especializado y sus unidades de representación: diversidad cognitiva. *Sendebar: Revista de la Facultad de Traducción e Interpretación*, 13: 141-153

Čermák, F. 2002. Today's Corpus Linguistics: Some Open Questions. *International Journal of Corpus Linguistics* 7/2: 265-282.

Chung, T. M. 2003. A corpus comparison approach for terminology extraction. *Terminology* 9/2: 221-245.

De Schryver, G.M. and D.J. Prinsloo. 2000. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The *macrostructure*. *South African Journal of African Languages* 20/4: 291-309.

Firth, J.R. 1935. The technique of Semantics. *Transactions of the Philological Society* 34: 36-72.

Firth, J.R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, *Special Volume*: 1-32.

Gilquin G. 2002. Automatic retrieval of syntactic structures. The quest for the Holy Grail. *International Journal of Corpus Linguistics* 7/2: 183-214.

Gómez González-Jover, A. and Ch. Vargas Sierra. 2002. Córpora comparables y paralelos para la detección de terminología bilingüe: su explotación y uso con herramientas informáticas. *Proceedings of the VIII Simposio Iberoamericano de Terminología* 8.

Gómez González-Jover, A. and Ch. Vargas Sierra. Ch. 2003a. Metodología para alimentar una base de datos terminológica desde las necesidades del traductor. *Proceedings of the I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación.* 629-648.

Gómez González-Jover, A. and Ch. Vargas Sierra. 2003 b. Utilización de herramientas informáticas para la elaboración de diccionarios bilingües. *Interlingüística* 13: 269-289.

Heid, U. 2003. The Handling of Collocations and Idiomatic Multiword Expressions: From Corpora to Dictionaries. Paper delivered at the 8th International Conference of the *African Association for Lexicography, Afrilex*.

Lardilleux, A. and Y. Lepage. 2007. The contribution of the notion of *hapax legomena* to word alignement. *3rd Language and Technology Conference*. 458-462.

Lerat, P. 1995. *Las Lenguas especializadas*. Barcelona: Ariel.

Sager, J., *et al*. 1980. *English Special Languages: Principles and Practice in Science and Technology.* Wiesbaden: Oscar Brandstetter.

Sager, J. 1990. *A practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Scott, M. 1997. PC analysis of Key Words and key key words. *System* 25/2: 233-245.

Scott, M. and Oxford University Press. 1998. *WordSmith Tools Manual version 5.0*. [Available at: http://www.lexically.net/downloads/version5/HTML]

Taljard, E. and G.M. de Schryver. 2002. Semi-automatic Term Extraction for the African Languages, with Special Reference to Northern Sotho. *Lexikos* 12: 44-74.

Vargas Sierra, Ch. 2005. *Aproximación terminográfica al lenguaje de la piedra natural. Propuesta de sistematización para la elaboración de un diccionario traductológico*, PhD dissertation. Universidad de Alicante.

YouJin, K. 2009. Korean lexical bundles in conversation and academic texts. *Corpora* 4: 135-165.