ojs.uv.es/index.php/qfilologia/index

Lingüístics

Assessing *EPAP* lexical features: A corpus-based study

Análisis de los rasgos léxicos de IFE: un estudio de corpus

María José Marín[a] & Camino Rea Rizzo[b]

[a] Universidad de Murcia. mariajose.marin1@um.es
[b] Universidad Politécnica de Cartagena. camino.rea@upct.es

Received: 10/02/2017. Accepted: 12/09/2017

Resumen: Las características de los lenguajes de especialidad se han descrito profusamente en la literatura especializada. El trabajo de Enrique Alcaraz destaca entre otros por su exhaustiva y minuciosa descripción del IFE a todos los niveles: léxico, sintáctico, semántico y pragmático. Este estudio tiene como finalidad la constatación de dicha descripción desde una perspectiva basada en análisis de dos corpus de inglés jurídico y de telecomunicaciones. Los resultados obtenidos corroboran lo ya observado por Alcaraz en lo que se refiere al uso de los términos especializados, la relevancia del vocabulario subtécnico, las peculiaridades de los términos latinos en el inglés jurídico y la significativa presencia de las abreviaturas en el inglés de telecomunicaciones.

Palabras clave: IFE; inglés jurídico; inglés de telecomunicaciones; lingüística del corpus.

Abstract: The features of specialised languages have been extensively described by scholars in the literature. Amongst them, Enrique Alcaraz's work stands out as an exhaustive and comprehensive description of EPAP at all linguistic levels: lexical, syntactic, semantic and pragmatic. This research aims to provide a bottom-up assessment of his description on a lexical level through the implementation of corpus-based techniques on two specialised corpora of legal and telecommunications English. The results support Alcaraz's portrayal as regards term usage, the relevance of sub-technical vocabulary, the peculiarities of Latin single and multi-word terms in legal English and the significant presence and usage of abbreviations in telecommunications English.

Keywords: EPAP; ESP; Corpus Linguistics; Legal English; Telecommunications English.

» Marín, María José & Rea Rizzo, Camino. 2017. "Assessing *EPAP* lexical features: A corpus-based study". *Quaderns de Filologia: Estudis Lingüístics* 22: 165-186. doi: 10.7203/qf.22.11307

1. Introduction

Specialised languages have been traditionally deemed functional varieties or registers (Biber, 1988; Halliday, 1988) defined in terms of the variation of the recurrence of particular linguistic features in comparison to general language or other registers. Cabré (1993) considers special languages a set of sub-codes from general language which are characterised by their own special features and pragmatically determined by the variables of topic, user and communication act.

Focusing on the definition of the language of science and technology, Sager *et al.* (1980) provide a comprehensive description of specialised languages. Their definition, use and function are synthesised as follows: "Special languages are semi-autonomous, complex systems based on and derived from general languages; their use presupposes special education and is restricted to communication among specialists in the same or closely related fields (1980: 69)." Similarly, Tiersma (1999) asserts that law practitioners depend upon language in their profession. According to this author, the special features of their jargon undeniably reveal their membership of the same community.

Alcaraz's (2000) definition is in line with all of the above, as he states that so-called special languages refer to the specific language that professionals and specialists use in order to transmit information and negotiate terms, concepts and knowledge in a particular field of knowledge. In *El inglés profesional y académico* (2000)*,* Alcaraz describes the most relevant features of English for Professional and Academic Purposes (EPAP), a term that he coins to refer to the specialised language which professionals and specialists employ to communicate. EPAP embraces many different branches or varieties associated with different professional or scientific fields such as medicine, law, engineering or business, amongst many others.

This research was conceived as an appraisal of Alcaraz's fundamental work through the analysis of the lexicon of two specialised corpora, *TC* (*Telematics Corpus*: 1.2 million words) (Rea, 2008) and *UKSCC* (*United Kingdom Supreme Court Corpus:* 2.6 million words) (Marín, 2014; Marín & Rea, 2012a), in search of linguistic evidence supporting some of the most relevant characteristics which scholars (Mellinkoff, 1963; Tiersma, 1999; Sager *et al.*, 1980; Alcaraz, 2000, 2002) have portrayed in the literature. The reasons to single out such differing EPAP varieties as legal and telecommunications English were related to the major objective of this research, that is, attempting to provide a bottom-up characterisation of specialised lexicons based on the general portrayals provided by scholars in the literature, specifically, Enrique Alcaraz's (2000; 2002). In principle, one would expect legal and telecommunications English terminology to differ considerably owing to their very nature and origins, the former belonging to the field of humanities and social sciences and having Latin and French influence (often being archaic and redundant) (Mellinkoff, 1963; Tiersma, 1999; Alcaraz, 2000, 2002), the latter coming from the realm of engineering and science and being highly specific and accurate. However, with regard to the statistical data associated with these lexical units, our main hypothesis was that both technical and subtechnical terms would behave similarly in both EPAP varieties, confirming the general descriptions made by scholars.

Owing to the size of both corpora and, above all, to our wish to carry out a fully automatic analysis with the aim of processing as much data as possible, only some of the features described by Alcaraz in his work were considered in this appraisal, namely, the ratio and distribution of highly specialised terms in both corpora; the relevance of subtechnical vocabulary; the use of Latin words and phrases in the legal corpus and the presence and significance of abbreviations and acronyms in the telecommunications corpus.

2. Literature review

Following from the above, this research concentrates on four major lexical features of EPAP which have been assessed applying a bottom-up corpus-based methodology, that is, by observing the statistical behaviour of the lexicon found in two specialised corpora, *TC* and *UKSCC*.

The literature devoted to the study of such features highlights the usage of specialised terminology as one of the most noteworthy aspects of EPAP as regards both its frequency of use and its distribution across text collections. Specialised terms could be defined as conceptual vehicles which are employed to transmit specialised knowledge amongst scientists, researchers, or professionals in all specialised areas, hence their relevance in EPAP. As Cabré (2000: 62) puts it, terms are "form and content units which, used in different discursive conditions, acquire a specialised value". According to Alcaraz (2000), terms tend to be univocal and their understanding is key to a proper comprehension of specialised texts, both oral and written. In other words, terms encapsulate specific concepts and must be understood and mastered by specialists, otherwise communication will fail.

Still within the lexical level, Alcaraz (2000) underlines the significance of semitechnical or subtechnical vocabulary as another relevant feature of EPAP. Subtechnical vocabulary is defined as those lexical units present in general language which acquire one or several specific meanings within a field of knowledge (Alcaraz, 2000: 43). In addition, subtechnical vocabulary is also understood as a collection of general words which are shared both by the general and the specialised fields without changing their meaning. Numerous authors have approached this question and defined sub-technical terms from different angles (Cowan, 1974; Baker, 1988; Flowerdew, 2001; Chung & Nation, 2003; Wang & Nation, 2004), agreeing on their ambiguous character and the difficulties that they cause to EPAP learners due to such obscurity. For the concept to be clearly delimited, Marín (2016) attempts to define it taking into consideration both qualitative and quantitative criteria.

Another relevant feature of EPAP, specifically of legal English, is the strong influence of Latin on its terminology, something that does not happen in telecommunications English. Although common law bears almost no resemblance with Roman law (which civil/continental law systems are based on), the presence of Latin in its terminology is more than merely anecdotal. Alcaraz (2000: 78) distinguishes between purely Latin borrowings like *obiter dictum* or *ratio decidendi,* which were imported directly from Latin without being adapted into English, and cognates such as *exonerate* or *presumption,* which reflect the English orthography although their meaning and form remain closely linked to their etymological origin. In the present research we will concentrate on the former and attempt to support these observations with evidence obtained from *UKSCC*, our legal corpus.

Within EPAP, the area of science, technology and computing is also characterised by the constant creation of new lexical units by using the linguistic resources of the corresponding language (Alcaraz, 2000: 50). The creation of new words responds to the need for the unique naming of concepts. According to Sager *et al.* (1980) and Alcaraz (2000), the principal method of designation in general and even more so in special reference is the modification of existing resources by means of concatenative processes, which follow the principle of adding some morphological material to a given form, namely, derivation and compounding. Nevertheless, there are also word formation processes that do not follow the principle of concatenation so new items are formed by deleting linguistic material instead of adding it. Amongst them, abbreviation refers to any kind of word which has undertaken a shortening process, that is, any compressed form in general. Abbreviation is an umbrella term which covers initials (also called initialism), acronyms (also called letter words) and clippings (Sager *et al.*, 1980; Alcaraz, 2000; Plag *et al.*, 2007).

Despite the relevance of the features depicted above, to the best of our knowledge, there are no corpus-based studies which can contribute to a bottom-up characterisation of specialised lexicons, apart from the ones carried out by Marín (2014; 2016), Rea (2008) and Marín & Rea (2012a; 2012b; 2014), hence the need to develop further research along these lines.

3. Methodology, results and discussion

Alcaraz's (2000) work presents a comprehensive portrayal of the major features of EPAP, which comprises the lexical, semantic, syntactic and pragmatic levels of the language and focuses on the use of specialised terminology, the features of major phrase types, the presence of polysemic words and metaphors in specialised texts or the communicative dimension of these texts, amongst many others.

For practical reasons and given the fact that this study was intended to be carried out automatically, a selection of these features was made so as to concentrate on the lexicon of legal and telecommunications English applying corpus linguistics techniques, that is, adopting a bottom-up perspective for the analysis of the two corpora. The selected features are: the use of specialised terminology in EPAP; the relevance of subtechnical terms; the significance of Latin terms and phrases in legal English; and the use of compressed forms or abbreviations as a result of word formation processes in telecommunications English.

3.1. *Specialised terminology in TC and UKSCC*

As regards the identification of specialised terms in large text collections like *UKSCC* or *TC*, they can be mined automatically using Automatic

Term Recognition (ATR) Methods. There is a whole plethora of them, some of which were validated on both corpora (Marín & Rea, 2014). The methods selected for evaluation were: *TF-IDF* (term frequency-inverse document frequency) (Sparck Jones, 1972); *TermoStat* (Drouin, 2003); *C-Value* (Frantzi & Anniadou, 1999) and *Terminus* (Nazar & Cabré, 2012). Their assessment was deployed through a comparison between the output lists of candidate terms produced by each method and two specialised glossaries of legal and telecommunications terms[1]. The overlap percentage between both vocabulary inventories showed the precision levels achieved by each of the methods and therefore led to a selection of the most efficient one.

Out of the four methods tested by Marín & Rea (2014), *Terminus* (Nazar & Cabré, 2012) excelled in comparison with the other three, managing to extract 71.5% true terms (terms which coincided with the ones in the glossaries used as gold standard) from *UKSCC* and 60% from *TC* on average. Precision was even higher for the top 200 candidate terms, reaching 84.5% for the former corpus and 69.5% for the latter.

As regards the legal corpus, implementing *Terminus* as the selected ATR method, a list of 1,787 terms was obtained, which represented 6.6% of the total 27,060 types identified by *Wordsmith 5.0* (Scott, 2008)[2]. These terms displayed an average frequency of 1,037 (each of them repeats itself throughout the corpus on 1,037 occasions) and appeared in 27 texts on average (out of 193). If compared with the average distribution of all the word types in *UKSCC* (19.8), excluding *hapax* and *dis legomena*[3], the distribution of the specialised terms extracted by *Terminus* could be deemed considerably high, actually, almost twice as high as all the types in the corpus. Not only were legal terms well distributed, but their frequency was also much higher than the average frequency of all the word types, occurring on 1,037 occasions as opposed to the average value of such types, 169.45, 6 times lower than the former (again, *hapax* and *dis* legomena were excluded from this count).

[1] The automatic validation of the lists was performed by resorting to two specialised electronic glossaries of legal English (of 10,054 terms) and telecommunications English (of 5,102).

[2] The term *type* refers to each of the words present in a corpus without counting their repetitions. Each of these repetitions would be labeled as *tokens*.

[3] Those types which occur once or twice respectively in the corpus.

The number of terms identified in the telematics corpus by *Terminus* was smaller, 888 terms out of 25,774 types, which represented 3.44% of the whole list. Their frequency counts were also lower than the same value in the legal corpus since the terms identified in *TC* occurred on 38.62 occasions on average, whereas the mean frequency of the whole type list was almost three times as high, that is, 89.93. Nevertheless, they were well distributed in the corpus being present in 30.14 texts out of 272 (the whole text collection) as opposed to the same average value for the whole of the type list, 14.59.

Judging by these figures, although term frequency counts were not so high in the telematics corpus as they were in the legal one, it could be affirmed that Alcaraz's observation about the significance of the use of terminology in specialised texts was confirmed from a bottom-up perspective with regard to both the frequency and distribution of legal and telematics terms.

3.2. *Quantifying the relevance of subtechnical terms in TC and UKSCC*

Concerning *UKSCC* and *TC*, the presence of subtechnical vocabulary was measured using Heatley & Nation's (2002) software *Range.* This software allows the user to obtain the percentage of running words in a text or text collection covered by a given word list which is included in the software package. Both the term lists obtained from our corpora were processed using the British National Corpus (BNC) list of the most frequent 3,000 words of English as the base list to compare them with. The resulting percentage would reflect the proportion of specialised terms from our lists which could be found amongst the most frequent 3,000 words of English, comprising words like *father, bank, the* or *water,* amongst many others. Such overlap would signal the percentage of subtechnical words present in both corpora given the fact that they were identified as specialised terms by *Terminus*, validated as such against a specialised glossary and also found as general vocabulary amongst the 3,000 most frequent words of English.

The overlap percentages varied in both cases, legal English being the variety which presented a higher amount of terms which coincided with the general English vocabulary from the BNC. 47.35% of the terms identified by *Terminus* were also present in the list of the most frequent 3,000 words of English. Such frequent words as *action, claim, decision* or *criminal* were included in our term inventory. Apart from their high frequency counts in the general field, they could be labelled as subtechnical owing to the fact that they acquire a technical meaning when in contact with the legal context.

About one third (35.55%) of the terms mined from *TC*, our telematics corpus, could also be found in the BNC list. The words *processor* or *controller*, which have a specialised meaning both in the general and the telematics fields, were found amongst that third. Other terms like *backbone,* also in the list of subtechnical telematics terms, specialise in the technical environment referring not to the human spine but rather to a local computer network. Table 1 below illustrates the top 25 subtechnical terms obtained from both corpora.

SUB-TECHNICAL LEGAL TERMS	*SUBTECHNICAL TELEMATIC TERMS*
Ability	Application
Absolute	Backbone
Acceptable	Bit
Action	Box
Admit	Call
Complaint	Client
Consistent	Controller
Creditor	Depend
Criminal	Entry
Damages	File
Debt	Logic
Employ	Mapping
Evidence	Model
Excuse	Neighbor
Exercise	Noise
Expense	Object

SUB-TECHNICAL LEGAL TERMS	*SUBTECHNICAL TELEMATIC TERMS*
Fact	Operate
Form	Packet
Privilege	Path
Proof	Programme
Reveal	Refer
Signature	Resource
Suicide	Route
Suspend	Server
Terminate	Site

Table 1. Top 25 subtechnical terms obtained from UKSCC and TC

Once more, having adopted a bottom-up perspective, Alcaraz's observation about the relevance of subtechnical vocabulary in specialised English has been corroborated by corpus evidence.

3.3. *Latin terms in UKSCC: a corpus-based assessment*

This section presents the study of legal terms which are employed in legal English without being adapted to the English orthographic or phonetic system, that is, they are pure Latin borrowings, as defined by Alcaraz (2000: 78). These must be distinguished from cognates, which are adapted to the English language system although their meaning and form still remain close to their etymological origin. The data and discussion offered below revisit and upgrade the study by Marín & Rea (2012b).

As a preliminary step, a list of Latin terms was obtained from text and academic books[4] which acted as reference for the identification of these lexical units in *UKSCC*. Such identification was carried out using

[4] See Mellinkoff, 1963; Alcaraz, 1994; Borja, 2000 and Orts, 2006, for academic references on Latin vocabulary in legal English and Fernández, 1994; Rice, 2007; Krois-Linder & Firth, 2008; Frost, 2009; Callanan, 2010 and Orts, 2010 for textbook references.

an excel spreadsheet to compare the type list produced by *Wordsmith* (Scott, 2008) with the Latin term list obtained from the books cited below automatically. Once single word Latin units were extracted (187 in total), it was attested that the top 10 most frequent ones were mostly function words, as is the case in general English, namely: *versus (v), per, de, inter* or *re.* There were other forms which, owing to their similarity with English, were excluded from these considerations (i.e. *in, sub* or *ex*), since they might produce misleading results. However, if compared with the whole *UKSCC* type list, their frequency was considerably low standing between the 400th and 1800th positions of the frequency rank. As a matter of fact, only 17 of these single word Latin terms fell within the top 2,000 word types identified by *Wordsmith.* Other Latin terms within this frequency range were *affidavit, quantum, jure,* or *incapax*.

Text range was also considered in this study as an indicator of a term's representativeness. Nation (2001) affirms that the higher this value for a given word is in a corpus, the greater its relevance within that corpus. The concept *text range* points at the percentage of running words in a text covered by that term or word list. For the sake of comparison, a sample list of 35 crime nouns (also regarded as specialised terms) was mined from the list of word types confirming the low frequency counts associated to Latin terms. Nevertheless, as regards text range, the figures varied showing that the 187 Latin term list covered 0.0059% of the words in UKSCC, whereas crime nouns covered only 0.00095%, almost six times less. Therefore, it could be stated that Latin terms, although not excessively frequent, present higher text coverage values than other specialised terms like *murder*, *abduction*, *threats* or *battery,* always bearing in mind that Latin terms only represent 0.69% of the total types identified in the corpus.

In a similar fashion, keyness was computed with the aim of determining the level of representativeness of Latin single-word units within the legal text collection. According to Scott (2008: 184), "a word is considered key if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlists". Keyness can be calculated automatically by comparison with a general English corpus using *Wordsmith*. Resorting again to the list of crime nouns used as reference for comparison with our Latin word inventory, the results showed that, in spite of the lower frequency of Latin terms, they could be considered as relevant as crime nouns standing at only three points below the latter and displaying 94.3 keyness. This value is also considerably high if compared with the average keyness of the whole list produced by wordsmith, namely, 116.08.

Finally, the level of specialisation of these terminological units was also measured in an attempt to substantiate Alcaraz's observation on their relevance in legal English. In this case, Chung's (2003) ratio ATR method was applied to rank the Latin terms according to their degree of specificity. Chung's method is based on corpus comparison, classifying a word type as a term only "if it occurs 50 times more often in the technical text than in the comparison corpus, or if it only occurs in the comparison corpus" (2003: 53). This termhood ratio can be easily calculated by first dividing a word's frequency of occurrence in both corpora by the number of tokens in each corpus, and then dividing the result obtained using the data from the specialised corpus by the same data obtained from a general one[5]. The value obtained should be above 50 for a word type to be regarded as a specialised term. The Latin terms in our list were therefore arranged and filtered according to Chung's method, which resulted in an inventory which included terms such as *affidavit, caveat, proviso, extempore, quantum, lex* or *subpoena.*

Nevertheless, most of these forms are either part of general or academic vocabulary and could therefore not be regarded as legal terms proper, for instance *plus, nil, persona, memorandum, caveat* or *alibi,* or they simply do not occur in isolation but rather as part of phrases. This is why the study on their specificity level was extended to Latin phrases, displayed in table 2.

TYPE	*FREQUENCY UKSCC*	*DISTRIBUTION*	*RATIO*
Ex turpi causa	129	3	∞
Doli incapax	36	1	∞
Quantum meruit	27	5	∞
Mutatis mutandis	24	18	∞
Alter ego	21	5	∞

[5] The general English corpus used in this case was LACELL, a 20 million-word corpus of general English texts compiled and owned by the LACELL research group from the English Department at the University of Murcia, Spain.

TYPE	*FREQUENCY UKSCC*	*DISTRIBUTION*	*RATIO*
Forum non conveniens	13	3	∞
Actus reus	10	5	∞
Ad litem	10	3	∞
Usque ad coelum	8	1	∞
Pari delicto	7	1	∞
Ratione personae	6	3	∞
Doli capax	5	1	∞
Debet ese	4	1	∞
Ad factum	4	1	∞
Res iudicata	4	2	∞
De novo	4	3	∞
Praesumptio juris	3	1	∞
Jus cogens	3	1	∞
In par material	3	2	∞
De jure	52	5	145,6
Pari passu	28	4	117,6
Ex parte	115	26	96,6
Ultra vires	79	16	82,95
Et seq	29	17	81,2
A fortiori	32	28	67,2

Table 2. Top 25 Latin phrases and their level of specialisation

As shown in table 2, like single-word Latin terms, the average frequency of these phrases is far from the mean value of the whole corpus, the former being 27.66 whereas the latter is 7 times higher. This data clearly points at their high level of specialisation, which is reinforced by the ratio values. 22 out of the 53 phrases mined from *UKSCC* do not occur in the general English corpus, being therefore assigned an infinity ratio value and standing at the top of the specificity rank, namely, *mutatis mutandis, quantum meruit* or *actus reus,* amongst other.

In spite of their low frequency, their distribution across the corpus is quite high. Phrases like *de facto, inter alia, prima facie* or *pro rata* occur in approximately a fourth of the texts in the corpus. Furthermore, while the average text distribution of all the word types in the corpus (excluding *hapax* and *dis legomena*) is 25.82, an eighth of the texts in it, Latin terms appear in 14.97 texts on average (under the same conditions), quite a high value given their degree of specialisation.

Summing up, term distribution together with their specificity may be considered as two key factors in determining the relevance and representativeness of a word or group of words within a corpus, whereas frequency simply indicates how many times a word repeats itself. Thus, the low frequency rates associated with Latin terms in *UKSCC* should not be deemed indicative of their little significance within the corpus. On the contrary, their level of specialisation coupled with their considerably high text distribution clearly signals their keyness within the variety supporting, once again, Alcaraz's (2000) observations as well as other scholars' like Mellinkoff (1963), Tiersma (1999) or Borja (2000).

3.4. *Abbreviations in TC: major findings and discussion*

As stated in the section devoted to the literature review, the term abbreviation is an umbrella term which covers initials (also called initialism), acronyms (also called letter words) and clippings (Sager *et al.*, 1980; Alcaraz, 2000; Plag *et al.*, 2007). First, initialisms are formed by combining only the initial letter of multi-word combinations giving rise to a sequence of letters which are pronounced individually, in the way in which the letters are spelt in the alphabet, e.g. TNT, DVD, IP, GPS, etc. However, when the combination of initial letters is pronounced as regular words following the regular reading rules of English, it becomes an acronym, e.g. NASA, LASER, NATO, etc. Clippings, in turn, result from usually monosyllabic or disyllabic words where the first part of the word base is kept, e.g. doc from doctor, sec from second, etc. Sometimes, an initial or middle element of the word can be also omitted like gbyte from gigabyte (Sager *et al.*, 1980; Jackson, 1988; Alcaraz, 2000; Plag *et al.*, 2007).

Corpus analysis corroborates Alcaraz's description of EPAP, precisely in telecommunications English, where compressed forms play a crucial role, since they stand for 16% of the terms included in the Telecommunications Engineering Word List (TEWL) (Rea, 2008). This lexical repertoire includes the most salient, central and typical specialised lexical units in the domain. They are all found within the range of the 1000 most statistically significant word families in the domain, as drawn by the comparison of the general language corpus LACELL. Their specialty index is obtained by applying Chung's method (2003) and the keyness index is given by the likelihood test in *WordSmith* (Scott, 2008) mirroring the procedure applied to the study of Latin terms in legal English.

Rank	*TEWL*	*F.Tec*	*F.Lacell*	*Ratio*	*Keyness*
1	IP	5,239	20	994,85	16,182
2	TCP	1,717	12	543,41	5,248
3	ATM	1,639	35	177,85	4,817
4	LAN	1,481	27	208,32	4,387
5	OSPF	1,284	0	∞	4,027
6	QOS	1,155	0	∞	3,622
7	VHDL	1,150	0	∞	3,607
8	MPLS	1,112	0	∞	3,487
9	GSM	1,109	4	1052,96	3,427
10	VPN	1,007	5	764,89	3,097
11	IEEE	1,002	9	422,83	3,044
12	LSAS	858	1	3258,58	2,676
13	DSP	906	41	83,92	2,523
14	LSA	804	0	∞	2,521
15	CDMA	805	1	3057,29	2,510
16	CISCO	840	14	227,87	2,498
17	MHZ	792	18	167,11	2,319
18	GHZ	734	2	1393,82	2,275

Rank	*TEWL*	*F.Tec*	*F.Lacell*	*Ratio*	*Keyness*
19	FPGA	713	0	∞	2,236
20	SCTP	703	0	∞	2,205
21	RF	716	8	339,91	2,161
22	DB	774	36	81,65	2,149
23	WLAN	677	0	∞	2,123
24	ISDN	699	14	189,62	2,061
25	HTTP	801	96	31,69	1,946

Table 3. Top 25 abbreviations in TEWL

As table 3 illustrates, the relevance of abbreviations is evidenced by their quantitative behaviour both within *TEC*, the main telecommunications corpus, and *TC*, the subcorpus of telematics. As already stated, a whole of 443 abbreviations comprise 16% of the word forms of the specialised repertoire.

Considering the total number of abbreviations appearing in the term inventory extracted from *TC* and the ratio yielded by Chung's method, there are 237 forms (53%) which are not found in the general corpus, hence they are assumed not to be typical of general language but characterised by a high degree of specialisation. Such highly technical terms display a keyness index which ranges from 4,027 (*OSPF*) to 12.5 (*VDMS*). The higher their frequency in the specialised corpus, the higher their keyness index. The next group comprises the abbreviations whose ratio is > 50, which amount to 119 forms (27%), and also occur in the general English corpus LACELL. They are characterised by their high frequency in TC and their low frequency in the general corpus, their keyness being also dependent on their frequency in the former corpus. The most significant abbreviation, *TCP,* belongs to this group, being 543 times more frequent in the telecommunications domain than in general language, and scoring 5,248 in keyness. Finally, the remaining 87 abbreviations (20%) are also used to a greater or lesser extent in LACELL so that their ratio is < 50. This does not mean that they are not specialised terms but their use has been extended to general language, thus being subtechnical. Therefore, their frequencies in both corpora do not differ that much, although their keyness might vary considerably.

The most significant unit in this group is *HTTP* (1,946) and the lowest score is yielded by *GUIS* (11). Other forms in this category are the following: *RADAR, PC, ID, MAC, WAN, WWW,* etc.

A final perspective is gained when approaching the quantitative behaviour of abbreviations in connection with the whole telecommunications word list (TEWL). When the different values which define the lexical behaviour of the terms in the list are taken as reference, the particular performance of abbreviations may be contrasted so that it is evidenced to what extent they approach the top and bottom scores. The most relevant term in TEWL is *network* (F. TEC: 16,649; F. LACELL 1,686; R: 37,50; K: 41,784) and *microchips* gets the lowest score in keyness (F.TEC: 9; F. LACELL: 6; R: 5.69; K: 10). Such references highlight and clarify the terminological character of abbreviations and their relevance in the specific domain, particularly of those which rank the highest like *IP* (the fifth most relevant term in TEWL), *TCP, ATM, LAN, OSPF,* etc. Moreover, amongst the top 100 words of the specific list, there are 14 abbreviations of which *OSPF, QOS,VHDL, MPLS* and *LSA* cannot be found in the general language corpus, and *IP, TCP, ATM, LAN, GSM, VPN, IEEE, LSAS* and *DSP* give a ratio > 50 whereas their keyness is considerably high ranging from 16,182 (*IP*) to 2,521 (*LSA*).

With respect to the different shortening process that abbreviations undergo, initialisms (360) remarkably stand out from the rest since they represent 81% of the total. The majority of the abbreviations found in the specific corpus come from the combination of the initial letter of multi-word units which is pronounced as a sequence of letters such as *IP, TCP, ATM, GPRS, SNMP, BGP, DCE, GPS, IGRP, PBX* or *BS*. Concerning acronyms, there are 74 in the list covering 17% of the abbreviations. In that case, the combination of the letters is pronounced as regular words like *RADARS, FIFO, VOIP, RIP, IPSEC, PAC, QOS, MAC, CISCO, OSI, LABVIEW, LDAP, SPICE,* etc. Finally, there are only 11 clippings, where the first or last part of the word base has been kept.

Some abbreviations, particularly acronyms, have been lexicalised and accepted as full words capable of undergoing compounding, derivation and conversion processes. Clear evidence of this behaviour is observed directly from the list of abbreviations where pairs of singular and plural forms are found, for example *LAN/S, VLAN/S, RAM/S, COMSAT/S, RADAR/S, FIFO/S, PAC/S,* etc. The metal-oxide semiconductor (*MOS*) family neatly illustrates compounding and how it forms multi-word units which again undergo a shortening process and become a longer acronym: *CMOS* (complementary metal-oxide semiconductor), *NMOS* (n-channel metal-oxide semiconductor), PMOS (p-channel metal-oxide semiconductor), *BICMOS* (bipolar complementary metal-oxide semiconductor) and *MOSFETs* (metal-oxide semiconductor field-effect transistors).

In short, it follows from the above that both the quantitative behaviour and the lexicalisation of abbreviations demonstrate their terminological character and typicality in the subject field as pointed out by Alcaraz (2000). In addition, all those compressed forms are linguistic labels which stand for definitions, being characterised by special reference within telecommunications, even those which have been integrated into the general language. Therefore, standardised abbreviations are also terms which achieve complete and effective communication in the specialised language singling it out from general language.

4. Conclusion

Corpus Linguistic techniques can detect automatically what is usual or unusual in a sublanguage with respect to general language, which establishes a reference norm, or in comparison to other sublanguages. In this research, the adoption of a corpus-based approach has allowed to identify the typical behaviour of the lexicons of legal and telecommunications English, providing a bottom-up depiction of some of their most relevant characteristics and corroborating the portrayal carried out by authors such as Alcaraz (2000, 2002). The application of ATR methods and the quantitative parameters intended to measure how vocabulary performs in *UKSCC*, the legal corpus, and *TC*, the telematics Corpus, have permitted to depict the use of specialised terminology (including subtechnical terms) and the outstanding use of Latin terms and phrases in legal English and abbreviations in telecommunications English.

Our initial hypothesis departed from the assumption that specialised terminology would behave similarly across EPAP varieties, following Alcaraz's (2000; 2002) portrayal. Such hypothesis was confirmed although certain differences were also observed between the two varieties selected for this research, namely, legal and telecommunications English.

Concerning the use of specialised terms in both varieties, the results vary slightly particularly concerning the frequency of these lexical items in the field of telecommunications. While terms tended to occur 6 times as much (1,037) as the whole list of types (169.45) identified in the legal corpus on average, this value was three times lower (38.62) than the average for the whole type list (89.93) in *TC*, the telecommunications corpus. Nevertheless, they were well distributed throughout both corpora appearing in 13.98% legal texts and 11.08% telecommunications ones and representing 6.6% and 3.44% of the whole list of types identified in both text collections respectively.

The literature also signals the significance of subtechnical terms in specialised languages, that is to say, of those terms which can be found in both specialised and general language contexts either retaining their technical meaning or activating it when in contact with the specialised environment. Testing showed that a large proportion of legal and telecommunications terms overlapped with the list of the 3,000 most frequent words of English found in the BNC. In fact, almost half of the terms in the legal corpus (47.35%) and about one third (35.55%) of the telecommunications terms could be found amongst these general words.

Within the field of legal English, Alcaraz (2000) particularly underlines the relevance of Latin words and phrases, which was also tested from a bottom-up perspective. The results evidenced that their frequency was not as high as expected, that is, if compared with the whole type list, they stood between positions 400th and 1800th in the frequency rank. However, when considering only Latin phrases, it appeared that both their level of specialisation and their distribution throughout the text collection was much higher, standing at the top of the specificity rank and appearing in 14.97 of the texts in the corpus (on average) in spite of their low frequency.

Finally, the use of abbreviations was also assessed within the field of telecommunications English. It was attested that 16% of the terms in TC were abbreviations (almost one fifth of the whole list), displaying really high levels of specialisation since 53% of them were not even found in the general context. In fact, when processing the telecommunications corpus with Keywords (Scott, 2008), abbreviations were assigned an average keyness value of 1,634 as opposed to the same value for the whole term list, that is, 237.26, which clearly points at their specificity and relevance in the corpus.

5. References

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

Alcaraz Varó, Enrique. 2002. *El inglés jurídico: textos y documentos.* Madrid: Ariel Derecho.

Baker, Mona. 1988. Subtechnical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language* 4(2): 91-105.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Borja Albí, Anabel. 2000. *El texto jurídico en inglés y su traducción.* Barcelona: Ariel.

Cabré, María Teresa. 1993. *La teminología. Teoría, metodología aplicaciones*. Barcelona: Antártida/Empúries.

Cabré, MaríaTeresa 2000. Terminologie et linguistique: la théorie des portes. *Terminologies nouvelles. Terminologie et diversité culturelle* 21: 10-15.

Callanan, Helen & Edwards, Linda. 2010. *Absolute Legal English*. London: Delta.

Cowan, Ronayne. 1974. Lexical and syntactic research for the design of EFL. *TESOL Quarterly* 8: 389-399.

Chung, Teresa M. & Nation, Paul. 2003. Technical Vocabulary in Specialised Texts. *Reading in a Foreign Language* 15(2): 103-116.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1): 99-117.

Flowerdew, John. 2001. Concordancing as tool in course design. In Ghadessy, Mohsen; Henry, Alex & Roseberry, Robert (eds.) *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins.

Frantzi, Katerina T. & Ananiadou, Sophia. 1999. The c/nc value domain independent method formulti-word term extraction. *Journal of Natural Language Processing* 3(2): 115-127.

Halliday, Michael. 1988. On the language of physical science. In Ghadessy, Mohsen (ed.) *Registers of Written English: Situational Factors and Linguistic Features*. London: Pinter.

Heatley, Andrew & Nation, Paul. 2002. *Range,* computer software. Wellington, New Zealand: Victoria University of Wellington.

Krois-Linder, Amy & Firth, Matt. 2008. *Introduction to International Legal English: A course for Classroom or Self-study Use*. Cambridge: Cambridge University Press.

Jackson, Howard. 1988. *Words and their Meaning*. London: Longman.

Marín Pérez, María José. 2014. Evaluation of five single-word term recognition methods on a legal corpus. *Corpora* 9(1): 83-107.

Marín Pérez, María José. 2016. Measuring the Degree of Specialisation of Sub-Technical Legal Terms through Corpus Comparison: a Domain-Independent Method. *Terminology* 22(1): 80-102.

Marín Pérez, María José & Rea Rizzo, Camino. 2012a. Structure and design of the BLRC: a legal corpus of judicial decisions from the UK. *Journal of English Studies* 10: 131-145.

Marín Pérez, María José & Rea Rizzo, Camino. 2012b. How relevant are Latin wordforms and clusters in legal English? A corpus-based study on the representativeness and specificity of such elements in UKSCC: an ad hoc legal corpus. *ES. Revista de Filología Inglesa* 33: 161-182.

Marín Pérez, María José & Rea Rizzo, Camino. 2014. Assessing four automatic term recognition methods: Are they domain-dependent? *English for Specific Purposes World* 42: 1-27.

Mellinkoff, David. 1963. *The Language of the Law.* Boston: Little, Brown & Co.

Nation, Paul. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nazar, Rogelio & Cabré, María Teresa. 2012. Supervised Learning Algorithms Applied to Terminology Extraction. In Aguado de Cea, Guadalupe; Suárez-Figueroa, Mari Carmen; García-Castro, Raul & Montiel-Ponsoda, Elena (eds.) *Proceedings of the 10th Terminology and Knowledge Engineering Conference* (TKE 2012). Madrid: Ontology Engineering Group, Association for Terminology and Knowledge Transfer, 209-217.

Orts, María Ángeles. 2006. *Aproximación al discurso jurídico en inglés*. Madrid: Edisofer Libros Juridicos S.L.

Plag, Ingo; Arndt-Lappe, Sabine; Braun, Maria & Schramm, Maria. 2007. *Introduction to English Linguistics*. Berlin: Mouton de Gruyter.

Rea Rizzo, Camino. 2008. *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico* (Tesis doctoral). Universidad de Murcia.

Rice, Sally. 2007. *Professional English in Use: Law.* Cambridge: Cambridge University Press.

Sager, Juan; Dungworth, David & McDonald, Peter F. 1980. *English Special Languages*. *Principles and Practice in Science and Technology.* Wiesbaden: Brandstetter Verlag KG.

Scott, Mike. 2008. *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Sparck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28: 11-21.

Tiersma, Peter. 1999. *Legal Language.* Chicago: The University of Chicago Press.

Wang, Karen & Nation, Paul. 2004. Word Meaning in Academic English: Homography in the Academic Word List. *Applied Linguistics* 25(3): 291-314.