(C-207)

SPECIALIZED CORPORA ON THE BASE OF TEACHING INNOVATION IN ESP.

Camino Rea Rizzo

Mª José Marín Pérez

(C-207) SPECIALIZED CORPORA ON THE BASE OF TEACHING INNOVATION IN ESP.

Camino Rea Rizzo y Mª José Marín Pérez

Afiliación Institucional: Universidad Politécnica de Cartagena y Universidad de Murcia

Indique uno o varios de los siete Temas de Interés Didáctico:

[] Metodologías didácticas, elaboraciones de guías, planificaciones y materiales adaptados al EEES.

[ ] Actividades para el desarrollo de trabajo en grupos, seguimiento del aprendizaje colaborativo y experiencias en tutorías.

[ ] Desarrollo de contenidos multimedia, espacios virtuales de enseñanza- aprendizaje y redes sociales.

[X] Planificación e implantación de docencia en otros idiomas.

[ ] Sistemas de coordinación y estrategias de enseñanza-aprendizaje.

[ ] Desarrollo de las competencias profesionales mediante la experiencia en el aula y la investigación científica.

[ ] Evaluación de competencias.

Resumen.

La implementación del Plan Bolonia en las universidades españolas ha provocado cambios estructurales entre los cuales el uso de lenguas extranjeras se ha convertido en un factor esencial dentro del concepto de Espacio de Educación Superior. Es más, la idea de internacionalización es uno de los objetivos principales del "Campus Mare Nostrum", integrado por las Universidades de Murcia y Cartagena. Así pues, se hace necesario un idioma de comunicación para desarrollar ese concepto en las áreas de la investigación y la docencia superior y el inglés es la primera lengua vehicular a este respecto. Este artículo explora las posibilidades que ofrecen los corpus especializados para la guía y elaboración de materiales didácticos en la enseñanza del inglés de Telecomunicaciones y Derecho. Se estudiarán dos corpus específicos: TEC y BLaRC, diseñados, compilados y analizados para fines de investigación. Se prestará especial atención a aspectos esenciales en el análisis de los corpus lingüísticos como los índices de frecuencia y relevancia, entre otros, por tratarse de factores determinantes a la hora de establecer el vocabulario clave de ambos lenguajes especializados, siendo éste un punto de partida para la creación de nuevos materiales didácticos.

Keywords: *lingua franca*, Corpus specializados, inglés con fines específicos: telecomunicaciones y derecho, vocabulario básico

Abstract.

The implementation of the Bologna Reform at Spanish universities has brought about major changes amongst which the use of foreign languages has become a key issue within the concept of a European area of higher education. Moreover, the idea of internationalisation is one of the fundamental goals of the "Campus Mare Nostrum" integrated by the Universities of Murcia and Cartagena. Henceforth, a language of communication is required to develop such concept in

the fields of research and teaching and English appears to be the major vehicular language or *lingua franca*. This paper explores the possibilities offered by specialised language corpora for the planning and elaboration of didactic materials to teach English as a specialised language within the fields of Telecommunications Engineering and Law. Two specific corpora will be studied: TEC (Telecommunication Engineering English Corpus) and BLaRC (British Law Report Corpus) designed, compiled and analysed for research purposes. Special attention will be paid to such questions as frequency rates or keyness, amongst others, as determining factors to identify the core vocabulary of both specialised languages, a point of departure for the creation of new didactic materials.

1. INTRODUCTION

The arrival of the Bologna reform at Higher Education in Spain has marked a radical turn in the way of managing university teaching and learning. The whole university community has been impelled to shift its attitudes towards a European concept of 'university' according to the present society's needs and demands. Some results of the convergence to European standards are already plain to see at several levels, whereas some other actions seem to be still in progress, such as the integration of teaching through the medium of a foreign language and the adequate language training for students.

The Bologna Declaration of 19 June 1999, signed jointly by the European Ministers of Education, stresses the importance of a European area of higher education "*as a key to promote citizens' mobility and employability and the Continent's overall development*". Hence, a common language for international communication becomes essential so that speakers with different first languages could communicate with each other through a vehicular language, that is, a lingua franca.

In addition, the Campus Mare Nostrum project sets internationalization as the primary immediate goal that the university shall strive for. The programme for innovation, quality teaching and language training is one of the sub-targets constituting the CMN3.1 objective: teaching excellence (http://www.campusmarenostrum.es). Moreover, such objective is further developed as follows: "*New resources that encourage innovation and teaching quality for European Higher Education Area will be developed, (...) Additionally, education in an international context will be bolstered, and research will be promoted as a structural part of the educational system. A specific formation system will be established for Professional Development technicians and an International Postgraduate School and a Personnel Education Centre will be created in coordination with other international institutions, thus promoting continuous learning and student mobility.*" The projection of the institution on an international scale requires a language of communication likewise capable of reaching education, research, training and mobility scopes on an international scale as well. Nowadays, there is no doubt that such lingua franca role is certainly played by English, which is being used as the working language of twelve major international domains (Graddol, 1997:8): Working language of international organizations and conferences; Scientific publication; International banking, economic affairs and trade; Advertising for global brands; Audio-visual cultural products, e.g. tv, popular music; International tourism; Tertiary education; International safety; International law; Interpretation and translation as a relay language; Technology transfer; and, Internet communication.

The domains of science and technology and international law are particularly addressed in this paper, as far as the teaching of English for specific purposes is concerned within the area of Tertiary Education that connects both institutions covered under the umbrella of the Campus Mare Nostrum project.

After the Bologna reform, English has been integrated in the current university programmes in two possible ways. On the one hand, English has been adopted as the language of instruction in a considerable part of some compulsory subjects, or, on the other hand, English is offered for specific purposes but as a separate subject independently of content subjects. The former is the case of some teaching innovation projects at the Technical University of Cartagena, and the latter is the situation of both English for Telecommunication Engineering and Legal English incorporated into the new degrees at the Technical University and the University of Murcia respectively. Those ESP subjects are compulsory and run only for one term during the whole programme. Legal English is placed in the first year's second term and Telecommunication English in the third year's first term. Concerning the proficiency level that students must reach, those subjects are devised according to B2 or Vantage level on the scale of the Common European Framework

for languages (CEF) as a logical progression from the B1 or Threshold level that students are assumed to have attained on completion of Post-Compulsory Secondary Education. On the CEF scale, B2 corresponds to an intermediate level by which the independent user of the language "*can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation; can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party; can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options*." However, the skills that students are expected to develop and the targets set at this level are extremely difficult to achieve taking into account that English is scarcely visible during their training period. In this situation, corpus-driven studies come into play by facilitating a straightforward approach to the essential vocabulary actually used by a discourse community, narrowing the gap between the existing lacks and the aimed target within the short stretch of time assigned to English. Furthermore, corpus-driven findings may also guide the content-subject teaching in English as the teachers are provided with the language resources appropriate for lecturing in that language. The benefits that accrue from corpus research place specialized corpora on the base of teaching innovation.

We present two cases of specialized corpora, the Telecommunication Engineering English Corpus (TEC) and the British Law Report Corpus (BLaRC), which have been devised for linguistic research and teaching innovation, because of their great potential for improving ESP teaching and learning. TEC and BLaRC are also a result of the cooperation and synergy between both universities conforming our Excellence Campus.

Next, the concept of linguistic corpus will be defined together with a brief definition of TEC and BLaRC. Then, we will focus on vocabulary in relation to word frequency and text coverage, we will move on to corpora analysis in terms of frequency and keyness, to finish with some final remarks.

2. DESCRIPTION OF THE CORPORA

A brief explanation of the concept of linguistic corpus may be considered suitable in the present setting. According to Sinclair (2004), who has been one of the fathers of corpus linguistics in the world, "*a corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research*." The implications that the concept of linguistic corpus involve, as the term is currently understood in corpus linguistics, have been made clearer and clearer in the course of its history. All in all, a linguistic corpus is defined as a collection of written and/or oral naturally occurring texts; it is selected under specific criteria in order to characterize a variety of the language or the whole language; it is computer processed and used for linguistic research (Johansson, 1991; Atkins, Clear & Ostler, 1992; Sánchez, 1995; McEnery & Wilson, 1996; Biber, Conrad & Reppen, 1998; O'keefe, McCarthey, & Carter, 2007).

In keeping with the concept of linguistic corpus, the samples of the language may be gathered to serve different purposes. A general corpus is normally compiled to be used as a reference for contrastive analysis or to provide a description of the general language. Thus, the compilation usually comprises texts from a wide range of genres and topic areas with the intent to reflect the typical usage of the general language. Alternatively, specialized corpora are designed to collect samples of a particular variety or register of the language with manifold objectives, depending on the research goals. The primary aim of the compilation of the two corpora presented herein is to identify the most relevant vocabulary in the language produced by the speakers of either telecommunication or law discourse communities so that a reliable vocabulary list could stand as a solid base to guide language learning and teaching.

Both corpora have been created intentionally to serve the research purposes with a careful design, so that they might be considered as reasonably representative of the written use of the language they capture. All the samples originate from real communication acts, have been prepared for computer processing and systematized in relation to the following criteria: topic variety, chronology, origin, mode and size. A well-designed corpus creates an excellent opportunity to look into language evidence and perform quantitative and qualitative analyses, since linguistic behaviour can be

quantified by attaching a frequency index to individual forms being therefore possible to make statistical inferences of the language. In addition, it is important to highlight that the suggested size for specialized corpora is one million words (Pearson, 1998) and our corpora far exceed it, as the larger the corpus is, the clearer the description of the language may be.

On the one hand, the Telecommunication Engineering Corpus comprises a sample of 5.5 million words of the professional and academic written language, which covers the main divisions of the realm (Electronics; Computing Architecture and Technology; Telematic Engineering; Communication and Signal Theory; Materials Science; Business Management; and System Engineering) and two branches of expertise (Communication Networks and Systems; and Communication Planning and Management). The language samples were produced by native and non-native speakers of English and extracted from a wide gamut of sources (magazines, books, web pages, research papers, abstracts, brochures, advertisements and technology news).

On the other hand, the British Law Report Corpus is projected to reach 6 million words – its present size is 1,609,330 words and it keeps on growing. Law reports, that is, collections of judicial decisions or judgements, are the legal genre singled out to constitute the corpus, due to the pivotal role they play in the UK judicial system as well as in any other common law countries. The United Kingdom belongs to the realm of common law, as opposed to civil or continental law which is the judicial system working in most Western European countries. Case law is at the basis of common law systems which rely on the principle of binding precedent to work, that is to say, a case judged at a higher court must be cited and applied whenever it is similar to the one being heard in its essence (the *ratio dicendi*). Another fact that makes law reports an outstanding genre in common law legal systems is that they not only cover all the branches of law, but also touch upon and cite other public and private law genres, proving certainly useful as far as lexical coverage is concerned. Regarding the time span covered by BLaRC, it ranges from 2008 to 2010. It follows the structure of UK courts and tribunals and responds to it basically because of two questions. Firstly, the relevance of the hierarchy of courts and tribunals in the UK legal system. The principle of binding precedent establishes that any decision made at a higher court or tribunal will set binding precedent as long as the case is similar to the one under examination. Secondly, if this structure is maintained, the texts will be grouped according to the field of law they belong to, so they may be similar in lexical terms. The texts are full authentic transcriptions of judicial decisions.

Finally, BLaRC is structured into five main categories depending on the jurisdictions of their judicial systems, that is, the geographical scope of their courts and tribunals: Commonwealth countries; United Kingdom, England and Wales, Northern Ireland, and Scotland.

3. FOCUS ON VOCABULARY.

Research has proved that different vocabulary sizes are needed to understand different types of text depending on several factors such as genre, register, text length, topic, and number of authors. According to Laufer (1989) and Nation (1990), it is necessary to understand 95% of the words in a text in order to get a fair comprehension. This rate means that the reader would encounter one unknown word in every 20 running words, that is, around one word per two lines. Regarding academic discourse, a 4000 family word size vocabulary could reach 95%. This vocabulary should be made of 2000 high-frequency general service words, about 570 general academic words and 1000 or more specialized words, proper nouns and low-frequency words (Nation, 2001).

Students usually access Higher Education with a B1 level of English which is characterized by the ability to maintain interaction and deal flexibly with topics focusing on daily activities and events (Council of Europe, 2001). The type of texts they can understand comprises mainly high frequency everyday language whose lexical content is consistent with the most frequent words in the general language. Therefore, B1 students have presumably gained an overall command of the 2000 most frequent words registered in the *General Service List of English Words* (West, 1953) or in a general list based on corpus frequency data like Nation's frequency lists drawn from the British National Corpus (Nation, 2004). Then, the successive vocabulary goal for learners moving on to special purposes study is set on academic and

specialized vocabulary (Nation & Hwang, 1995). On the one hand, academic vocabulary will facilitate the acquisition and understanding of the subject content as conveyed by the university community. On the other hand, it is arguable that terminology learning is a process running in parallel to subject learning. Still, content subject lectures are delivered in Spanish so that explicit teaching of technical vocabulary is worthwhile, especially how terms operate and are idiomatically used in the register.

The availability of the specialized corpora allows extracting the most significant vocabulary in the realms in comparison to general English. Our objective fits in the tradition within lexical studies of developing word lists (Thorndike & Lorge 1944; West 1953; Nation 1990; Coxhead 2000) for teaching and learning English as a second language, and even so, it comes to satisfy the existing demand for discipline-based lexical repertoires (Nation, 2001; Hyland & Tse, 2007; Read, 2007; Rea, 2008) in order to guide materials writers and exams developers, assist instructors' teaching, and meet students' specific needs.

4. WORDS IN TEC AND BLARC.

Once the appropriate corpora have been compiled, they are processed by using WordSmith program (Scott, 1998). The immediate data obtained are those related to the basic statistical information (Table 1) and the frequency word list (Table 2). Next, a corpus-comparison approach will be adopted so as to identify the key words in both specialized languages, since word frequency within one corpus is not a direct index of relevance or keyness. Such index is established by the differences in word frequency in the specialized corpus as compared to a general corpus.

4.1 Basic statistical information.

The program counts 5,533,705 tokens in TEC and 1,609,330 in BLaRC. A token is defined as a sequence of characters divided by blank spaces or punctuation marks. Many of the tokens are the repetition of same words, so the number of types or wordforms indicates the number of different words in the corpus, including each form derived from a main lemma or headword (59,826 types in TEC and 20,398 in BLaRC). The set of types constitutes the vocabulary of the text. In the following string of four tokens: *controls*, *controlled*, *controller*, *controls*, there are three different forms from only one lemma: *control*. The concept of lemma corresponds to the lexical entry in a dictionary, that is, a lemma is the canonical form or citation form of a set of forms.

The relationship existing between the total number of types and tokens is given by type/token ratio and standardised type/token ratio, which provide information on the corpus lexical diversity from different perspectives. First, the type/token ratio is obtained from the division of the whole number of different forms by the number of occurrences and multiplied by 100. The higher the result is, the greater the lexical diversity of the sample. On the contrary, a lower ratio means a lower lexical burden in the text due to the repetition of the same forms. Next, the program computes the standardized type/token ratio every *n* words, being *n*=1,000 and yielding the average of the obtained values. In TEC, there is an average of 38.26 different forms per each text sequence of 1,000 tokens, whereas in BLaRC, the figure is a bit lower: 33.03. Specialized corpora might be expected to contain more forms than a general corpus, owing to the nature of the specialized discourse where speakers need technical terms to convey specific concepts accurately.

The basic statistical information gives an account of the number of sentences and paragraphs; the average length of words, sentences and paragraphs; the figure of words according to the number of letters, etc. Nevertheless, it does not report on the *hapax legomena* phenomenon, that is, the words occurring only once in the corpus (table 1). Although they can be easily identified in the frequency list, it is interesting to show the corresponding figure so as to improve the overall view of the corpus composition, and because this is also an indicator of lexical variety.

Basic statistics	**TEC**	**BLaRC**
Tokens	5,533,705	1,609,330
Types	59,826	20,398

Type/token Ratio	1.08	1.32
Standarised Type/token	38.26	33.03
Sentences	223,278	47,530
Sentences length	23.87	29.89
Paragraphs	30,472	12,392
Paragraphs length	102.95	124.79
Hapax legomena	21,755	6,019

Table 1. Basic statistical information.

4.2 Frequency lists.

Due to the large amount of data available from the corpora, the frequency lists are illustrated in table 3 which displays the most frequent 100 words in TEC and BLaRC.

	TEC	**Frequency**	**BLaRC**	**Frequency**
1	THE	374598	THE	137253
2	OF	170493	OF	59698
3	AND	145104	TO	52675
4	TO	140361	IN	37844
5	A	131832	THAT	37563
6	IN	104696	AND	31128
7	IS	94630	A	30275
8	FOR	63834	WAS	20414
9	THAT	52034	IS	18501
10	ARE	41713	FOR	15880
11	BE	41543	ON	15291
12	AS	40644	IT	15066
13	THIS	37780	NOT	14685
14	WITH	36388	AS	13649
15	ON	33072	BE	12991
16	BY	31017	BY	12401
17	IT	28218	OR	9795
18	AN	27199	WHICH	9311
19	CAN	25288	HAD	9140
20	OR	24488	S	9104
21	FROM	21405	WITH	8592
22	AT	20529	THIS	8399
23	WE	17733	TRIBUNAL	8030
24	WHICH	17453	HAVE	7604
25	NOT	17069	HE	7584

26	NETWORK	16649	AN	7583
27	WILL	16128	AT	7549
28	HAVE	16114	MR	7271
29	DATA	14613	FROM	6521
30	HAS	13444	BEEN	6403
31	ONE	13218	WOULD	5605
32	SYSTEM	12624	WERE	5532
33	TIME	12391	THERE	5444
34	USED	11874	NO	5040
35	IF	11826	I	5024
36	ALL	11602	HIS	4996
37	THESE	11577	HAS	4796
38	MORE	11454	ANY	4756
39	YOU	11448	ARE	4591
40	ALSO	10382	WE	4533
41	OTHER	10291	APPELLANT	4530
42	USE	10255	CASE	4522
43	EACH	10224	DECISION	4489
44	SUCH	10006	CLAIMANT	4422
45	ITS	9714	APPEAL	4420
46	WHEN	9681	EVIDENCE	3786
47	WAS	9664	BUT	3763
48	SYSTEMS	9479	IF	3706
49	TWO	9407	UNDER	3639
50	USING	9214	RESPONDENT	3586
51	INFORMATION	9161	SHE	3551
52	BUT	8933	MADE	3527
53	THEIR	8930	HER	3478
54	THEY	8734	ITS	3348
55	BASED	8448	THEY	3342
56	BETWEEN	8403	MAY	2872
57	NEW	8347	OUT	2865
58	THAN	8052	EMPLOYMENT	2840
59	I	8004	OTHER	2834
60	ONLY	7939	DID	2830
61	MAY	7808	WHETHER	2751
62	NUMBER	7759	TIME	2708
63	DESIGN	7701	SUCH	2642
64	THERE	7494	PARAGRAPH	2582
65	FIGURE	7325	SO	2501
66	INTO	7308	DOMAIN	2478
67	BEEN	7150	NAME	2477

68	CONTROL	7124	SHOULD	2470
69	SERVICE	7085	ALSO	2428
70	SIGNAL	7022	ALL	2335
71	SOME	6904	SECTION	2318
72	E	6607	ONE	2279
73	EXAMPLE	6379	COMPLAINANT	2218
74	HIGH	6348	COULD	2216
75	S	6315	BEFORE	2177
76	FIRST	6314	THOSE	2168
77	USER	6292	ACT	2101
78	DIFFERENT	6198	ONLY	2052
79	SERVICES	6161	THEIR	2029
80	ANY	6052	SAID	2018
81	ACCESS	5999	WHEN	2001
82	PROCESS	5949	WHAT	1995
83	OVER	5925	FIRST	1978
84	SAME	5904	INFORMATION	1964
85	THEN	5897	DO	1921
86	MODEL	5895	WHO	1894
87	SO	5861	WILL	1887
88	NETWORKS	5832	BEING	1859
89	SET	5813	PART	1764
90	PERFORMANCE	5686	WHERE	1742
91	UP	5652	LAW	1739
92	MOST	5602	RELEVANT	1689
93	WHERE	5463	THESE	1674
94	APPLICATIONS	5414	WORK	1674
95	WERE	5378	CLAIM	1671
96	BOTH	5355	V	1658
97	LEVEL	5309	THAN	1654
98	IP	5239	HOWEVER	1650
99	THROUGH	5202	SOME	1647
100	OUR	5152	ABOUT	1614

Table 2. The most frequent 100 words in TEC and BLaRC.

One of the key findings discovered from the examination of frequency lists reveals that the most frequent words cover a high percentage of occurrences in a language (Sinclair, 1991; Schmitt, 2000). As noticeable in table 2, *the* is the most frequent word in the corpora and stands for 6.77% and 8.52% of the total tokens in TEC and BLaRC respectively. In general language, the 3 most frequent words commonly reach 11% of the whole, the 10 most frequent ones 22%, the 50 most frequent ones 37%, the 100 most frequent ones a 44% and the 2,000 most frequent words cover around 80% (Schmitt, 2000). Those figures agree with the results obtained from our corpora with some variations (Table 3):

Most frequent words	Coverage in General language	Coverage in TEC	Coverage in BLaRC
3	11%	12%	15%
10	22%	23%	27%
50	37%	36%	40%
100	44%	42%	52%
2.000	80%	79%	85%

Table 3. Coverage of the most frequent words.

From the 5.5 million-word sample in TEC, only 59,826 words are different forms, and 21,755 of them occur only once in the corpus, which correspond to just 0.39% of the whole sample. More than 34,000 forms occur from 1 to 3 times and there are around 30 words whose frequency is 100, whereas around 750 words are used more than 1,000 times in the corpus. BLaRC, in turn, has 20,398 different forms of which 6,019 occur just once in the 1.6 million (0.37% of the whole). There are 182 words whose frequency is higher than 1,000 and around 700 words fall within a frequency range from 100 to 200. In brief, more than half of the texts are made on the basis of repetition.

Frequency list analyses have also shown that the most recurrent words are functional words. Auxiliary and modal verbs, pronouns, articles, prepositions and conjunctions help to construct the grammatical structure of the language, do not convey lexical meaning and their behaviour does not change. On the other side, notional words convey the bulk of lexical content. Contrary to functional words, content words depend on the language variety registered in the corpus. In addition, the most frequent words are inclined to keep a steady distribution, so that any outstanding change in the ranking may be significant (Sinclair, 1991). In a general corpus, around the most frequent 100 words are functional. Therefore, the intrusion of notional words into that range points out a remarkable behaviour.

The more specialized a corpus is, the more content words reach high frequency levels, whereas in general corpora, notional words start predominating from the most frequent 150 words onwards (Kennedy, 1998). In table 2 it is noteworthy that *network*, the first content word in TEC, is found in the 26th position. Henceforth, functional and notional words alternate until *control* in the 68th position, where there is a decreasing presence of functional words and content words are more recurrent. The first content word in BLaRC is found earlier, *tribunal* reaches the 24th position followed by *appellant* in the 41st from which content words start to be more noticeable.

The greater number of notional words found in the high frequency levels may be an indicator of lexical density, which translates as lexical burden for teaching and learning. Besides, the fact that some of the most frequent 50 notional words are of specialized character might lead to expect a high presence of technical terms, even so, there are also general content words within this high frequency range (Table 4). Therefore, a further analysis is required to check to what extent such frequency is significant. This leads to the next step of the analysis which focuses on the comparison of the statistical behaviour of the specialized corpora with a general language corpus so as to identify which words are particularly relevant.

TEC
network, process, data, model, system, networks, time, set, used, performance, use, systems, applications, using, both, information, level, based, IP, new, current, only, value number, power, design, technology, figure, management, control, protocol, service, type, signal, work, example, software, high, frequency, first, internet, user, users, different, application, services, state, access, layer.

BLaRC
tribunal, appellant, case, decision, claimant, appeal, evidence, respondent, employment, paragraph, domain, name, section, complainant, information, part, law, relevant, claim, fact, date, period, hearing, application, question, state, letter, rights, judge, person, registration, issue, circumstances, order, set, reasons, view, regulation, income, consider, judgment, VAT, use, account, company, respect, basis, right, business, court.

Table 4. The most frequent 50 content words in TEC and BLaRC.

4.3 Keywords.

The degree of relevance or keyness is obtained by running the KeyWords tool available in the pack of utilities in WordSmith. This tool identifies keywords on statistical basis by comparing patterns of frequency. A keyword is defined as "*a word which occurs with unusual frequency in a given text*" (Scott, 1998: 237), that is to say, a word whose frequency is unusually high or low in comparison to a general norm. A large general corpus establishes the reference norm which is contrasted to the specific corpora. In this case, the general corpus Lacell (21 million words compiled by Lacell research group) is used to perform the analysis.

According to the characteristics of the samples, the Log Likelihood statistical test is applied to generate keywords list: "*Log Likelihood test, gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus*" (Dunning, 1993; Scott, 1998). As a result, the test detects if the frequency of a word in the specialized corpora is significantly higher (positive keywords) or lower (negative keywords) than its frequency in the general corpus. Then, the program generates a keywords list sorted by the keyness index associated to every keyword (Table 5). 12,602 keywords have a significantly higher frequency in TEC, where the highest keyness value associated to a word is 41,784 (*network*) and the lowest one is 10 (*broad*). BLaRC, in turn, gains 2,860 positive keywords whose indexes spread from 40,227 (*tribunal*) to 24 (*contending*).

	TEC	**Keyness**	**BLaRC**	**Keyness**
1	NETWORK	16.649	TRIBUNAL	8.030
2	DATA	14.613	APPELLANT	4.530
3	SYSTEMS	9.479	CLAIMANT	4.422
4	IP	5.239	RESPONDENT	3.586
5	NETWORKS	5.832	APPEAL	4.420
6	SYSTEM	12.624	THE	137.253
7	PROTOCOL	4.742	MR	7.271
8	DESIGN	7.701	THAT	37.563
9	ROUTER	3.910	DECISION	4.489
10	WIRELESS	4.083	COMPLAINANT	2.219
11	LAYER	4.425	DOMAIN	2.478
12	MOBILE	4.341	PARAGRAPH	2.582
13	INPUT	4.347	EVIDENCE	3.786
14	INTERNET	4.504	EMPLOYMENT	2.840
15	INTERFACE	3.526	CASE	4.522
16	BANDWIDTH	3.119	NOT	14.685
17	PACKET	3.577	REGISTRATION	1.462
18	CIRCUIT	3.932	HMRC	1.030

19	ACCESS	5.999	WAS	20.414
20	OUTPUT	4.139	RELEVANT	1.689
21	SERVER	3.574	SECTION	2.318
22	DIGITAL	3.595	VAT	1.226
23	SOFTWARE	4.575	REGULATION	1.259
24	SIMULATION	2.817	JUDGMENT	1.230
25	DEVICES	3.430	HEARING	1.592
26	VOLTAGE	2.945	COMMISSIONERS	992
27	OPTICAL	2.822	JUDGE	1.467
28	TRAFFIC	4.345	V	1.658
29	ALGORITHM	2.799	UNDER	3.639
30	NODE	2.822	CLAIM	1.671
31	LINK	3.853	NAME	2.477
32	TECHNOLOGY	4.969	APPLICATION	1.524
33	FILTER	2.627	WHETHER	2.751
34	COMMUNICATIONS	3.144	ACT	2.101
35	CHANNEL	3.212	DISMISSAL	793
36	FIG	3.702	COMMISSIONER	838
37	TRANSMISSION	2.544	CIRCUMSTANCES	1.422
38	PACKETS	2.308	DATED	855
39	NODES	2.361	REGULATIONS	1.075
40	PROCESS	5.949	RESPONDENTS	686
41	ROUTERS	1.891	PAYMENT	1.101
42	WEB	2.978	DATE	1.598
43	PROTOCOLS	1.996	REGISTERED	895
44	CIRCUITS	2.122	RELATION	1.109
45	DEVICE	2.801	SUBMISSIONS	659
46	FIBER	1.869	HAD	9.140
47	ETHERNET	1.897	APPELLANTS	536
48	COMPONENTS	2.727	COMPLAINT	794
49	PROCESSING	2.770	OF	59.698
50	COMMUNICATION	3.159	ABUSIVE	609
51	TCP	1.717	REFERRED	989
52	CONFIGURATION	1.885	RIGHTS	1.472
53	CODE	3.112	UK	1.318
54	HARDWARE	2.257	EMPLOYER	794
55	TECHNOLOGIES	2.137	COMPLAINANT'S	511
56	ALGORITHMS	1.777	WEBSITE	596
57	ATM	1.639	MS	921
58	LAN	1.481	PROCEEDINGS	768
59	ARCHITECTURE	2.581	SUBMISSION	681
60	OSPF	1.284	SUBMITTED	737

61	LOOP	1.766	REASONS	1.328
62	WAVELENGTH	1.352	DISCLOSURE	662
63	LOGIC	1.920	LETTER	1.494
64	JAVA	1.344	NOTICE	1.112
65	COMPONENT	1.981	RESPECT	1.204
66	SWITCH	2.075	CONTRACT	1.087
67	ANALOG	1.264	BY	12.401
68	QOS	1.155	CONSIDER	1.237
69	VHDL	1.150	REASONABLE	928
70	ANTENNA	1.242	PURPOSES	854
71	DISTRIBUTED	1.824	PERIOD	1.594
72	LINEAR	1.590	WHICH	9.311
73	MPLS	1.112	ENTITLED	819
74	TELECOMMUNICATIONS	1.105	ANY	4.756
75	GSM	1.109	CONCLUSION	777
76	SPECTRUM	1.499	GOODS	938
77	LINUX	1.128	MADE	3.527
78	INTERFACES	1.187	REGENT	453
79	CHANNELS	1.569	BASIS	1.198
80	REMOTE	1.770	DISCRIMINATION	647
81	CELL	2.089	INCOME	1.242
82	CABLE	1.715	UNFAIR	554
83	AUTHENTICATION	1.082	LTD	640
84	SERVERS	1.208	PARTIES	1.059
85	VPN	1.007	SATISFIED	660
86	FILTERS	1.207	FACTS	796
87	SWITCHING	1.315	TAX	1.440
88	IEEE	1.002	TO	52.675
89	BROADBAND	1.049	LAW	1.739
90	X	3.002	BENEFIT	1.126
91	SATELLITE	1.401	ISSUE	1.429
92	DATABASE	1.451	PARA	413
93	LSAS	858	NOMINET	323
94	MODULATION	908	LJ	328
95	SWITCHES	1.107	CONSIDERED	1.054
96	DESTINATION	1.311	EMPLOYEE	635
97	NETWORKING	1.030	PROVISIONS	598
98	MULTICAST	837	ACCOUNT	1.219
99	VENDORS	1.004	PARAGRAPHS	447
100	IMPLEMENTATIONS	826	TRADER	403

Table 5. Keywords.

There exist noticeable differences between the keywords lists and the frequency lists. When comparing the first 100 words, it is interesting to note the predominance of functional words on the frequency lists over a small presence of words connected to the domains; whereas almost all the keywords are content words related to the subjects. Among the first 100 keywords it is possible to detect qualitatively technical terms restricted to the domain like *IP, bandwidth, Ethernet, wireless, LAN,* etc., in TEC, and *appellant, claimant, complainant, hearing, nominet* etc., in BLaRC.

The set of statistical features of the samples defines the specialized language against general language depending on the variation in the lexical choice, so that the meaning of lexical items is interpreted in discourse both by what they express and what they exclude. However, the current study focuses on the words that, statistically, are more probable to occur in telecommunications or in law reports. Moreover, positive keywords usually provide a good account of the subject content: "*positive keywords give a good indication of the text's aboutness*" (Scott, 1998).

The keyword lists driven from the corpora must be subjected to deeper analysis where additional statistical parameters are applied with several intends, such as sifting the different types of vocabulary. Nevertheless, those keywords mean a basic starting point for teaching and learning as they disclose the most representative, significant and relevant vocabulary of the corresponding discourse communities.

5. FINAL REMARKS.

The need of being a skilful user of English is an undeniable fact that our present European society and the whole university community are well aware of. Consequently, being capable of communicating in English is a realistic demand imposed on university students which does not seem to be in agreement with the offer in English training. Content-subject teachers who are actively involved in teaching innovation also share this pressure as they are asked to integrate English in their subjects and communicate through the medium of English independently of their expertise in the language.

A corpus-based approach to both English teaching, and teaching through English, would bring considerable benefits to the present situation not only in terms of vocabulary but also in many other respects. First, a corpus-driven vocabulary list would provide teachers and learners with the most relevant lexical repertoire of the corresponding discourse community, which would enable them to operate with the language according to its community lexical standards. Second, corpus-driven results are intended to guide ESP teaching within the time span allotted in the degrees. Traditional coursebooks, if they are available for every specific area of knowledge or domain, are designed according to the different levels that learners should go through, so they take into account the time estimated to learning, responding to a logical step-by-step learning process. In this sense, corpus-driven word lists might be used as a 'first aid kit' in an attempt to bridge the existing gap between the urgent demand and the range of courses on offer, to the maximum benefit of learning in such circumstances. Third, corpus-driven word lists are the starting point for further linguistic analysis. Words are not deployed or studied in isolation but in context, and they usually combine and cluster giving rise to higher order lexical units typical of the domain. Likewise, corpus software allows to retrieve at the same time all the contexts of a particular keyword so that its pattern of behaviour becomes easier to observe.

Finally, TEC and BLaRC are expected to be available on line as a database and reference source for teachers, students, practitioners and anybody interested in those specialized languages. The main thrust of this action is to contribute to a language-friendly environment and encourage life-long language so that teaching innovation could reach far and beyond university settings.

Bibliografía y Referencias.

Atkins, Clear and Ostler. (1992). "Corpus Design Criteria". *Literary and Linguistic Computing*, 7, 1: 1-16.

Biber, Conrad and Reppen. (1998). *Corpus Linguistics. Investigating Language Structure and Use.* C.U.P.

Bologna Declaration. (1999). *Joint declaration of the European Ministers of Education*.

Campus Mare Nostrum. (2011). *Plan director*. http://www.campusmarenostrum.es

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. URL: http://www.coe.int/T/DG4/ Linguistic/CADRE_EN.asp

Coxhead, A. (2000). "A New Academic Word List". *TESOL Quarterly* 34, 2: 213-238.

Dunning, T. (1993). "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics*. 19, 1: 61-74.

European Union. (2003). Promoting Language Learning and Linguistic Diversity: An Action Plan 2004-2006. Brussels: http//europa.eu.int/comm/education/doc/oficial/keydoc/actlang/ act_lang_en.pdf

Graddol, D. (1997). *The future of English*. London: British council.

Hyland, K. & P. Tse. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly* vol. 41:2, 235-253.

Johansson, S. (1991). "Computer corpora in English Language Research", in Johansson, S. & Stenström, A. (Eds.) *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyer.

Kennedy, G. (1998). *An introduction to corpus linguistics*. New York: Longman.

Laufer, B. (1989). 'What percentage of text-lexis is essential for comprehension?' in C. Lauren and M. Normand (Eds), *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.

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

Nation, P. (1990). *Teaching and Learning Vocabulary*. Newbury House.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–13). Amsterdam: Benjamins.

Nation, I.S.P. & K. Hwang, (1995). Where would general service vocabulary stop and special purposes vocabulary begin? in: *System, 23 (1)*, pp. 35-41, Pergamon Press, Oxford.

O'keefe, A., McCarthey, M., and Carter, R. (2007) *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Pearson, J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing Company.

Rea, C. 2008. *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico.* Tesis doctoral. URL: http://www.tesisenred.net/TDR-0611109-134048/index_cs.html>

Read, J. (2007). "Second Language Vocabulary Assessment: Current Practices and New Directions." *International Journal of English Studies* 7,2: 105-125.

Scott, M. (1998). *WordSmith Tools Manual version 3.0.*Oxford University Press.

Sinclair, J. (1991). *Corpus, Concordance and Collocation*. Oxford: OUP.

Sinclair, J. (2004) "Intuition and annotation - the discussion continues". *Advances in corpus linguistics.* Papers from the 23rd *International Conference on English Language Research on Computerized Corpora* (ICAME 23). Aijmer, K and Altenberg, B. (Eds.) Amsterdam/New York: Rodopi.

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Thorndike, E.L. and Lorge, I. (1944). *The teacher's Word Book of 30,000 Words*. Teachers College, Columbia University, New York.

West, M (1953). A General Service List of English Words. London: Longman.