A FIRST APPROACH TO THE LEXICAL PROFILE OF TELECOMMUNICATION ENGLISH: FREQUENCY, DISTRIBUTION, RESTRICTION AND KEYNESS.

Camino Rea Rizzo
Universidad de Murcia

***Abstract**: This corpus-based study is conducted to gain an insight into the lexis of Telecommunication English, with the aim of characterizing the lexical profile of this specialized language. The applied methodology integrates quantitative techniques and qualitative interpretations to perform an analysis from two different perspectives, and according to two parameters: restriction and keyness. The first approach is focused on the lexical behaviour and the extent that a word is restricted to the constituent areas of a domain, whereas the second approach is directed towards the extent that a word is significant in the domain, regardless of lexical category. The set of empirical and statistical data obtained contribute to map the lexical profile and will serve as a baseline for future studies.*

***Key words:** Corpus Linguistics, vocabulary classification, frequency, distribution, restriction, keyness.*

1. INTRODUCTION

Research in the field of English for Specific Purposes from a Corpus Linguistic approach has given rise to a series of studies on the detection and classification of the different types of vocabulary in specialized texts (Yang, 1986; Farrell, 1990; Coxhead, 2000; Nation, 2001; etc). The availability of a linguistic corpus offers the great advantage of quantifying language by associating a frequency index to every single word and performing statistical analysis of empirical data.

Vocabulary in specialized languages has been traditionally classified in three major groups: technical vocabulary or terminology, semi-technical or subtechnical vocabulary and general vocabulary. The definition of each category from a qualitative perspective refers to a general description but does not provide an automatic procedure for classification. Although qualitative criteria are enriched and coordinated with quantitative criteria, there exist several possibilities in the combination of the variables which determine the inclusion of a word in a category, and a definite method has not been clearly stated yet.

In this study, the vocabulary of Telecommunication Engineering English is explored in relation to the different combinations of the variables involved in term detection, in the light of the literature available dealing with non-tagged corpora and specialized languages.

The corpus specialized in Telecommunication Engineering English (TEC) has been compiled for the purpose of the research. TEC is a sample of 5.5 million words of academic and professional written English extracted from a wide range of sources (magazines, books, web pages, journals, brochures, advertisements and technology news), originating in native and non-native parts of the world and covering 18 subject areas subsumed under seven major areas of knowledge (Electronics; Computing Architecture and Technology; Telematic Engineering; Communication and Signal Theory; Materials Science; Business Management; and System Engineering) and two specializations in Telecommunication Engineering (Communication Networks and Systems; and Communication Planning and Management).

Due to the characteristics of the samples gathered, academic vocabulary is also included as a constituent of the lexical level in the specialized language, since the corpus encompasses the language used both in the academic environment and the corresponding professional career. Therefore, in a specialized text, it is possible to find four types of vocabulary: general, academic, technical and semi-technical (Sager, 1980; Cabré, 1993; Alcaraz, 2000; Nation, 2001).

This study is conducted to gain an insight into the lexis of Telecommunication English, with the aim of characterizing the lexical profile of this specialized language. The characterization is carried out from two different perspectives and according to two parameters: restriction and keyness. The first approach is focused on the lexical behaviour and the extent that a word is restricted to the constituent areas of a domain, while the second approach is directed towards the extent that a word is significant in the domain, regardless of lexical category.

2. COMBINATION OF QUALITATIVE AND QUATITATIVE CRITERIA IN VOCABULARY CLASSIFICATION

The qualitative and quantitative criteria proposed to define and recognize the different categories of vocabulary are based on the combination of the variables of frequency, distribution and restriction to a domain. Furthermore, there exist two well-known word lists which are usually taken as a reference to detect the academic and the most frequent general vocabulary: The Academic Word List (Coxhead, 2000) and The General Service List of English Words (West, 1953).

The group of general vocabulary is made of functional words and the content words registered in the General Service List. They generally coincide with the most frequent 2,000 words in a one-million-word general corpus and their statistical behaviour is characterized by high frequency and high distribution (Barber, 1962; Nation, 2001).

The academic vocabulary corresponds with the 570 word families registered on the Academic Word List. Those words are typically frequent in a great deal of academic disciplines, relatively infrequent in other kind of text (novels or oral colloquial language), and not related to any subject in particular (Wang and Nation, 2004).

Technical vocabulary consists of content words whose meaning is restricted to the specific subject, characterizes the specific language as an individual area of the global language and constitutes the terminology of the domain. Within technical vocabulary, there are also words from the general language which acquire a specialized meaning in the domain, but their general meaning is not applicable to their corresponding meaning in the technical context.

Quantitatively, technical words or terms show wide distribution in a specialized corpus, and their frequency is high in comparison to a general corpus (Yang, 1986; Farell, 1990; Nation, 2001; Chung, 2003). On the other side, a set of words qualitatively classified as technical terms exhibits either high frequency and distribution, or low frequency and distribution in a specialized corpus. The latter behaviour reveals that the term belongs to a narrower domain in the specialized field.

Once technical vocabulary has been identified, there comes a point where the classification of lexical units poses a great difficulty, as both qualitative criteria and statistical behaviour are quite controversial. This leads to classify the same words as semi-technical or subtechnical vocabulary, less specialized technical vocabulary or even as academic vocabulary.

Subtechnical vocabulary comprises general content words whose meaning becomes specialized in a domain but it is understandable from its meaning in a general context. They are called re-designated general language items by Sager et al. (1980). This category is also characterized by including non-technical formal words, independent of context, and occurring more frequently in varied technical and academic texts than in one specific domain. Such features induce to confuse the boundaries between semi-technical and academic vocabulary, being the reason why this group is named procedural vocabulary (Widowson, 1993) or specific common core (Robinson, 1991; Farrell, 1990; Dudley-Evans and St. John, 1998). Sometimes, a word may belong to two groups at the same time and categories overlap. Subtechnical vocabulary may even be part of technical vocabulary when the former is distinguished as lexical units less specialized or less restricted to the subject.

However, specialized vocabulary can be conceived from a broader standpoint, taking the position that it covers technical vocabulary or terminology and semi-technical vocabulary (Alcaraz, 2000 and Nation, 2001); that is, specialized vocabulary, as a whole, is made of lexical units of different degrees of specialization: both words whose use is restricted to a domain, and those used in other fields or in general language and acquire a specialized meaning in the domain.

3. CORPUS ANALYSIS IN TERMS OF RESTRICTION

3.1 Frequency and stop lists

The processing of TEC is carried out with the aid of WordSmith (Scott, 1996). The tools available in the computing program allow to retrieve frequency lists previously filtered by a stop list containing the words to exclude from the analysis. The applied stop list comprises functional, academic and general words. Those groups include the most frequent 2,000 words from the General Service List of English Words (West, 1953) and the words registered on the Academic Word List (Coxhead, 2000).

Prior to filtering, it is crucial to check the lists in order to spot the general and academic words which may acquire a specialized meaning within Telecommunications. If those words stayed on the stop list, relevant information would be lost. For instance, network, signal or system belong to the title of a subject area in the corpus (Signal Processing) or a branch of specialization in Telecommunications (Communication Networks and Systems). Therefore, the following family words closely related to the domain are extracted from the stop list: access, assemble, bond, channel, code, communicate, component, compound, compute, concurrent, couple, convert, data, design, device, discrete, distribute, image, input, link, layer, logic, media, network, offset, output, overlap, process, protocol, route, simulate, signal, system, technology, transmit, transfer and transform. Finally, a total of 10,773 word forms remain on the stop list and are excluded from the analysis.

After the subtraction, the whole of 59,826 word forms in the corpus are reduced to 50,864. The academic and general words included on the stop list cover around a 15% out of the whole of forms in the corpus. Data show that a high proportion of forms (85%) does not correspond to the most frequent general vocabulary or academic vocabulary. Nevertheless, the resulting list needs a second filter which involves us in manual cleaning, in order to dispose of the following kind of words: names and surnames (Acharya, Martínez, Stuber, Sugymoto, Zare, Zinio, etc), words in languages different to English (asuntos, attaché, universidad, vivisimo, etc), toponyms and nationalities (Cartagena, Chinese, Galway, Italy, Portugal, etc), misprints (acheived, amplidude, therfore, trafic, utput, etc) and words mistakenly joined (todigital, actiontooutputs, topeer, actorsmay, aproblem, etc).

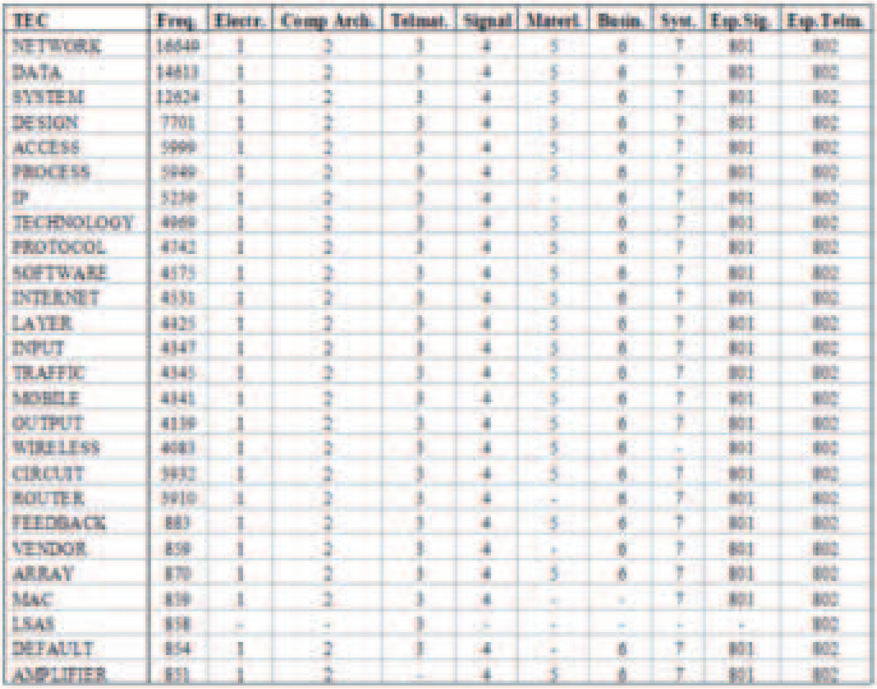
The outcome of the filtering should be a frequency list of the technical and semi-technical vocabulary. However, further criteria are required to sift data and refine results, since there remains a high figure of words which are not truly terms together with general words used in a great variety of subjects. Therefore, stop list application is not a method suitable for vocabulary classification, although it offers the advantage of reducing the original volume of words. After the two filterings, the frequency list contains 36,077 word forms and is taken as a base of reference for subsequent analysis.

3.2 Word distribution

The variable of distribution may be examined in several ways depending on the intended target. The occurrence of a lexical unit might be counted every time that it appears in a text or in a particular section, so that it is estimated how this lexical unit spreads across. In the present study, distribution is valued with respect to the areas of knowledge where a word occurs. Thus, that parameter allows to identify to what extent lexical units are restricted to each area of knowledge, which lexical units occur in several areas or in only one.

Stop lists are applied again, but the procedure is slightly different as it implies filtering the corpus by areas of knowledge, and contrasting the results with the reference list and the areas with each other. For this purpose, the corpus is divided into nine files (seven main areas and two specializations) which are filtered with the same stop list used for the whole corpus, generating nine independent frequency lists. Then, data are transferred to an Excel sheet and, after the required operations, the computer graphically displays which word from the reference list occurs in each section. If a word from the reference list also occurs in an area, the square is marked with the number corresponding to the area of knowledge, whereas a hyphen means that there is no coincidence. The distribution value ranges from 0 to 9. Table 1 illustrates this procedure.

As is noticeable from the table, high frequency words usually occur in the great majority of areas or in all the areas. The words occurring only in one section, which are restricted to one



TEC	Freq.	Electr.	Comp Arch.	Telmat.	Signal	Materl.	Busin.	Syst.	Esp.Sig.	Esp.Telm.
NETWORK	16640	1	2	3	4	5	6	7	801	802
DATA	14833	1	2	3	4	5	6	7	801	802
SYSTEM	12624	1	2	3	4	5	6	7	801	802
DESIGN	7701	1	2	3	4	5	6	7	801	802
ACCESS	5999	1	2	3	4	5	6	7	801	802
PROCESS	5949	1	2	3	4	5	6	7	801	802
IP	5239	1	2	3	4	-	6	7	801	802
TECHNOLOGY	4869	1	2	3	4	5	6	7	801	802
PROTOCOL	4742	1	2	3	4	5	6	7	801	802
SOFTWARE	4575	1	2	3	4	5	6	7	801	802
INTERNET	4531	1	2	3	4	5	6	7	801	802
LAYER	4425	1	2	3	4	5	6	7	801	802
INPUT	4347	1	2	3	4	5	6	7	801	802
TRAFFIC	4345	1	2	3	4	5	6	7	801	802
MOBILE	4341	1	2	3	4	5	6	7	801	802
OUTPUT	4139	1	2	3	4	5	6	7	801	802
WIRELESS	4083	1	2	3	4	5	6	-	801	802
CIRCUIT	3932	1	2	3	4	5	6	7	801	802
ROUTER	3910	1	2	3	4	-	6	7	801	802
FEEDBACK	883	1	2	3	4	5	6	7	801	802
VENDOR	859	1	2	3	4	-	6	7	801	802
ARRAY	870	1	2	3	4	5	6	7	801	802
MAC	839	1	2	3	4	-	-	7	801	802
LSAS	838	-	-	3	-	-	-	-	-	802
DEFAULT	854	1	2	3	4	-	6	7	801	802
AMPLIFIER	831	1	2	-	4	5	6	7	801	802

Table 1. Word distribution.

subject, are found in lower frequency levels. Furthermore, from the distribution of high frequency words we may infer that the areas of knowledge are closely related, excepting the area of Materials Science (05) that shares fewer words. After the filtering, a 1.8% of the remaining words occurs in all the sections. This percentage stands for 661 forms whose frequency ranges from 14 to 16,649. For example: network, data, system, design, process, access, technology, protocol, software, internet, layer, input, traffic, mobile, flux, rotation, chips, workshop, etc.

The words restricted to each area of knowledge are separated from the reference list and classified into 11 groups. There are two extra sections (03+0802 and 04+0801) which correspond to Telematics (03) plus its specialization (0802), and Signal Processing (04) plus its specialization (0801). Those sections are introduced as a result of the classification process, where we observed a tendency in some words to occur in two particular sections, coinciding with a specialization and its main area of knowledge. After the recount, a significant number of words are found in the two sections and are restricted to them. Therefore, those words are considered candidates for technical terms, as well as any word occurring in only one area: “one of the ways in which terminology can be automatically identified is in terms of its greater tendency to occur only in a limited range of texts” (Skehan, 1981 in Aston, 1996). In chart 1 is exhibited a sample of lexical units restricted to each area. Afterwards, table 2 displays relevant objective data about the quantitative behaviour of vocabulary: the figure of words restricted to each area, the proportion they cover, the highest frequency value and the number of words registered in the specified frequency ranges. This information helps get a broad idea of the potential number of terms and the level of abstraction in every area of knowledge.

AREAS	Distribution value = 1
01Electronics	photoresist, WLR, silicide, transimpedance, electronicast, optimiser, monolayers
02Comp.Arch	halfband, optocoupler, outfile, sequencer, diamondoid, vectorization, CSAS, ribosome
03Telematics	grouplet, NSSA, teletraffic, etherware, DSSAS, OGSI, flipper, virtualmedia, telecities
04Signal	biconical, SVM, dilation, PSPICE, cellview, gaussmeter, radiances, multiprocess
05Materials	foams, copolymers, nanofibres, amide, flexural, polycondensation, recoil, solubility
06Business	harassment, globalized, idealism, relativism, bystander, imobile, blogs, codevelopment
07Systems	DMCS, scalea, homeomorphic, divergences, neurofuzzy, multiprogrammed, ripper
801Esp.Signal	micropayment, monopulse, javacard, picocells, goniometer, beamforming
802Esp.Telmt	DLSW, multihomed, isoline, minislots, multivoip, agenthood, smartparther, infragard,
03+0802	LSAS, appletalk, boomers, collaborationware, desynchronization, permutable, gigafast
04+0801	radiometer, radionavigation, smartphone, undersampling, microcellular, microstrips

Chart 1. Samples of words restricted to an area.

AREAS	Restricted words	Coverage	Highest freq.	F=1	F=2	F=3	F from 4 to 8	F from 9 to 20	F from 21 to 65	F >65
01Electronics	2,456	12.8%	65	1,436	615	174	225	111	35	-
02Comp.Arch	729	6.4%	48	395	180	49	57	40	8	-
03Telematics	3,377	13.1%	217	1,963	548	205	448	146	55	12
04Signal	1,592	8.2%	107	1,060	237	105	97	75	177	1
05Materials	830	9.5%	51	545	137	44	73	24	7	-
06Business	1,527	9.3%	218	1,032	250	79	97	47	20	2
07Systems	816	6.2%	192	503	95	67	98	34	16	3
801Esp.Signal	2,219	10%	362	1,312	419	-	300	128	49	11
802Esp.Telmt	898	4.1%	150	535	-	42	140	120	55	6
03+0802	315	0.6%	858	-	158	78	72	-	-	7
04+0801	237	0.5%	74	-	56	50	66	42	22	1

Table 2. Words restricted to one area: Quantitative behaviour.

3.3 Application of criteria

Thus far, the quantitative analysis of the corpus has given access to the numerical values required for the application of the criteria which determine vocabulary classification. Once frequency and distribution values are available, the lexical units are subjected to analysis under the different combinations of the variables involved.

A simple test is performed by extracting at random lexical units from the corpus and placing them in a chart according to the parameters in a scale. As a matter of fact, the selection criteria initially proposed are laid aside, and all the possible combinations are proved, so that both high and low distribution may combine with high and low frequency as shown in chart 2.

		Technical words	Semi-technical words	General words
High distribution	High frequency	*ATM, bandwidth, latency, AL, java, wavelength, DC*	*software, switched, chip, network, array, interface, mapping*	*Communication, technology, system, storage, client, peak, image, link*
	Low frequency	*MO, gray, scalar, Mi, NE, calculus*	*Assembling, watts, Customize*	*Deficiency, obstacle*
Low distribution	High frequency	*Etherware, coons, JSC, nanomachinery, grouplet, DIOS*	*Salinity, sacks, enforcer, interferes, feeders, teletraffic*	*Pores, extractions, bumpers, foams, payout*
	Low frequency	*ACMOS, radioscope, alkyldithol, outband*	*Saline, teleoperators, abortion, antivirus, aerospatial*	*Basements, wrinkle, affirming, weary, monument, addressees*

Chart 2. Vocabulary classification according to frequency and distribution.

The test does not reveal illuminating results, since among high frequency and high distribution words there are both technical and semi-technical units as well as general vocabulary related to Telecommunications and other subjects. Even the same phenomenon is evidenced among low or high frequency words which are restricted to one area of knowledge, such as Etherware, coons, salinity, nanomachinery, pores, bumpers, foams, etc. From this moment on, the next stage should imply the qualitative study of every word in its context, since we lack further data to determine the lexical category.

Although the values of absolute frequency and distribution are not enough to identify efficiently and automatically the different types of vocabulary, they are basic criteria to take into account when deciding the lexical content for a course of technical English. Especially, high frequency content words which, at the same time, occur in all the areas (maximum distribution value) would be quite useful because those words have proved to be habitual and recurrent in the language of Telecommunications. On the other hand, high frequency words occurring in only one area reflect a feature typical of technical terms.

Nevertheless, if we aim at improving the results and attaining an automatic or semiautomatic classification of vocabulary, it is essential to compare the statistical behaviour of the lexical units in the specialized corpus with their behaviour in the general language and apply more sophisticated methods. Moreover, an accurate lexical description of the register requires a comparative approach as “systematic differences in the relative use of core linguistic features provide the primary distinguishing characteristics among registers” (Biber et al., 1998: 136).

4. CORPUS ANALYSIS IN TERMS OF KEYNESS

The fact that a word is a term or a specialized lexical unit whose use is restricted to a subject does not imply that it is also representative of the domain it belongs. In addition, the specialized nature of a lexical unit entails a higher relative frequency in the technical discourse than in the general one, but it does not impose a high probability of occurrence in specialized texts. The relation between specialization and representativiness is not directly proportional. It is easy to illustrate this argument by taking as example words like fastchip (Frequency 7) or bimos, polyspectra, securID and thinkpad (Frequency 4); or even words like bootable, vectorizable and axially whose frequency is 1 and each one occurs in a different area of knowledge. All these words are terms closely related to Telecommunications, but their incidence does not compare to wireless, voip, wan or routers whose frequencies are 4,082, 580, 452 and 1,892 respectively. The last examples, qualitatively deemed technical (voip and wan) or semi-technical units (wireless and routers), occur at a greater incidence and their use is more widespread than the first examples. Consequently, the next step of the analysis is to find out which words are more probable to occur in Telecommunications and then check if they are specialized and representative.

Mastering the technical terms typical of a domain is essential for successful communication, mainly in the most specialized situations that demand accuracy and precision: "El carácter monoreferencial de los términos desempeña un papel clave en la precisión y univocidad de la comunicación especializada" (Cabré, 1993:167). Indeed, a subject domain is not completely assimilated, if the speaker is not familiar with terminology. Hence, in a learning language situation, it would be very convenient for learners to study first the most probable specialized lexical units that they may encounter.

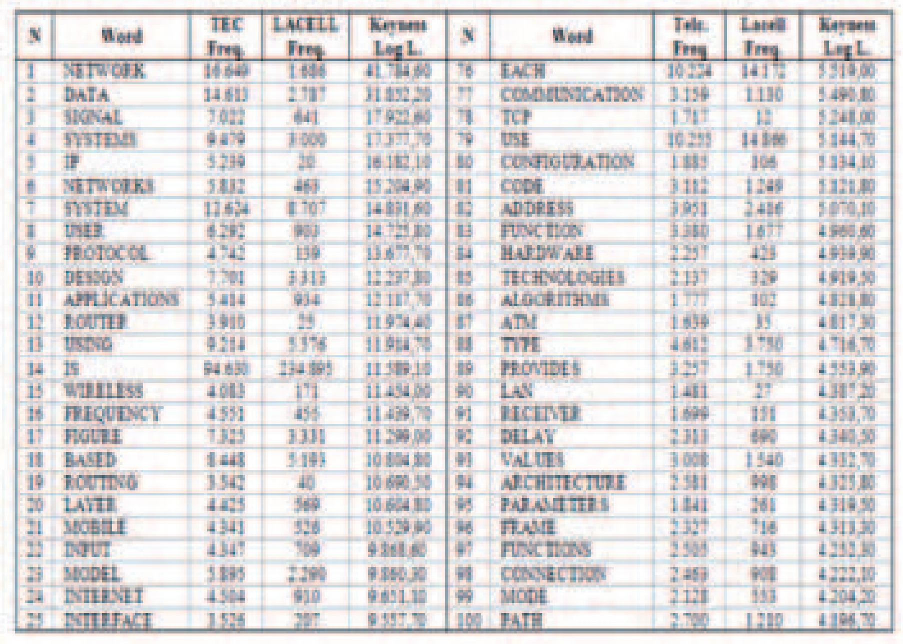
Next, the lexical level is analysed according to the degree of relevance or keyness of the words in TEC, with a double objective: first, to find out the vocabulary needed for effective communication and second, to get further data for a better description of lexical profile.

4.1 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 a mechanical basis by comparing patterns of frequency. A keyword is defined as "a word which occurs with unusual frequency in a given text" (Scott, 1997: 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 corpus. 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 and Scott, 1998). As a result, the test detects if the frequency of a word in the technical corpus is significantly higher or lower than its frequency in the general corpus. Then, the program generates a keywords list sorted by the keyness index associated to every keyword. 16,000 keywords are registered on the list, out of which 12,602 are positive and 3,398 negative.

Positive keywords have a significantly higher frequency in TEC than in the general corpus. The highest keyness value associated to a word is 41,784.6 (network) and the lowest one is 10.8 (broad). Likewise, each value is calculated within a margin of error from 0 to 0.000997, getting 7,815 positive keywords whose margin of error is equal to 0. (Table 3 provides a sample). As for negative keywords, their incidence is significantly lower in the specialized corpus. The highest keyness negative value is -10.9 (educating) and the lowest one is -29.8 (necklace). 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, particularly on those keywords with a margin of error 0. 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).



N	Word	TEC Freq.	LACELL Freq.	Keyness Log L.	N	Word	Tele. Freq.	Lacell Freq.	Keyness Log L.
1	NETWORK	16.649	1.688	41.784,60	76	EACH	10.224	14.172	5.519,00
2	DATA	14.633	2.787	31.832,20	77	COMMUNICATION	3.159	1.130	5.490,80
3	SIGNAL	7.022	641	17.922,60	78	TCP	1.717	12	5.248,00
4	SYSTEMS	9.479	3.000	17.377,70	79	USE	10.233	14.869	5.144,70
5	IP	5.239	20	16.182,10	80	CONFIGURATION	1.885	106	5.134,10
6	NETWORKS	5.832	463	15.206,90	81	CODE	3.112	1.249	5.121,80
7	SYSTEM	12.624	8.707	14.831,60	82	ADDRESS	3.951	2.486	5.070,10
8	USER	6.282	903	14.725,80	83	FUNCTION	5.380	1.677	4.980,60
9	PROTOCOL	4.742	139	13.677,70	84	HARDWARE	2.237	425	4.959,90
10	DESIGN	7.701	3.313	12.237,80	85	TECHNOLOGIES	2.137	329	4.919,50
11	APPLICATIONS	5.414	934	12.117,70	86	ALGORITHMS	1.777	102	4.828,80
12	ROUTER	3.910	25	11.974,40	87	ATM	1.639	35	4.817,30
13	USING	9.214	5.376	11.914,70	88	TYPE	4.612	1.750	4.716,70
14	IS	94.650	234.895	11.589,10	89	PROVIDES	3.257	1.750	4.553,90
15	WIRELESS	4.083	171	11.454,00	90	LAN	1.481	27	4.387,20
16	FREQUENCY	4.551	455	11.438,70	91	RECEIVER	1.699	151	4.353,70
17	FIGURE	7.325	3.331	11.299,00	92	DELAY	2.313	690	4.340,50
18	BASED	8.448	5.193	10.804,80	93	VALUES	3.008	1.540	4.332,70
19	ROUTING	3.542	40	10.690,50	94	ARCHITECTURE	2.581	998	4.325,80
20	LAYER	4.425	569	10.664,80	95	PARAMETERS	1.841	381	4.319,50
21	MOBILE	4.341	528	10.529,90	96	FRAME	2.327	718	4.313,30
22	INPUT	4.347	709	9.868,60	97	FUNCTIONS	2.595	845	4.252,30
23	MODEL	5.885	2.290	9.860,30	98	CONNECTION	2.483	908	4.222,10
24	INTERNET	4.504	910	9.851,10	99	MODE	2.128	553	4.204,20
25	INTERFACE	3.526	287	9.537,70	100	PATH	2.700	1.210	4.196,70

Table 3. Keywords.

There exist noticeable differences between the keywords list and the frequency list before its first filtering. When comparing the first 100 words, it is interesting to note the predominance of functional words on the frequency list over a small presence of words connected to Telecommunications (network/s, system/s, information, design, control, service/s, signal, user, access, process, performance, IP and applications); whereas almost all the keywords are content words related to the subject, apart from is, each and can. Among the first 100 keywords it is possible to detect qualitatively technical terms restricted to the domain like IP, bandwidth, Ethernet, TCP, ATM and LAN. Additionally, within the following 50 keywords, there is a rise in restricted terms which can be also distinguished quantitatively thanks to their statistical behaviour: high frequency in the specialized corpus and frequency 0 in the general corpus, for example OSPF, QOS, VHDL and MPLS.

4.2 Distribution of keywords

The distribution of keywords is examined employing a procedure analogous to that used in section 3.2. The method consists in contrasting a main keywords list with the nine independent keywords lists from the sections in the corpus. The main keywords list comes from the comparison of TEC with LACELL which establishes the norm. However, the keywords lists of each individual area have been generated taking TEC as the reference language in order to pinpoint the distinctive words in every area. With this action we are assuming that the statistical behaviour of the words common to the specialized language and one of its subject components will be similar, whereas the typical words of the subdomain will reveal significant keyness values. In this way, it could be possible to highlight terms and the general words which acquire a specialized meaning in the field.

It is worth mentioning that several experts in the areas of knowledge were consulted about the results. After retrieving two keywords lists for every section, one generated with TEC as reference and another with LACELL, experts were asked to choose which list better represents, in their opinion, the lexical content of their subject. Most experts opted for the keywords list originated with the specialized language as reference.

The final procedure developed for the nine areas follows the next steps:

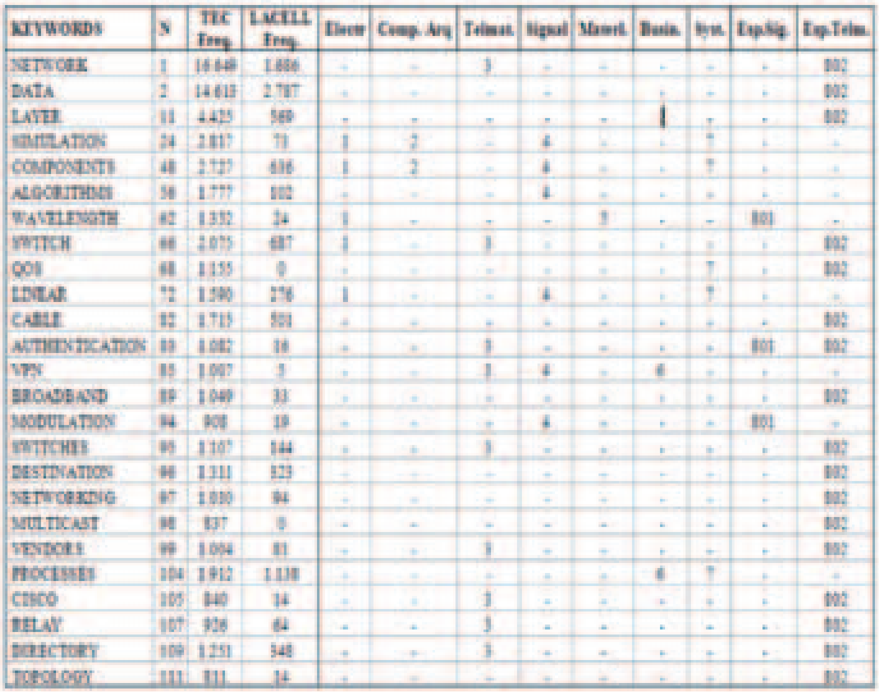
1. The corpus is divided into nine files according to the areas of knowledge.

2. A frequency list of a single area is retrieved, and a different frequency list is generated from the rest of areas altogether.

3. A keywords list is generated from the previous lists.

4. Only the keywords whose margin of error is equal to 0 are transferred to an Excel sheet. This selection criterion is adopted in order to concentrate the most significant words and reduce the volume of data which decreases towards the following figures: Electronics (1,054), Computing Arch. (501), Telematics (799), Signal (769), Materials (51), Business (636), Systems (550), Esp. Signal (936) and Esp. Telematics (903).

5. Finally, the individual keywords list is contrasted to the reference keywords list which also contains the words whose error of margin is 0 (5,834 keywords). Table 4 shows a sample selected at random which illustrates the distribution of keywords across the subject areas.



KEYWORDS	N	TEC Freq.	LACELL Freq.	Electr.	Comp. Arq	Telmat.	Signal	Materl.	Busin.	Syst.	Esp.Sig.	Esp.Telm.
NETWORK	1	16 649	1 886	-	-	3	-	-	-	-	-	802
DATA	2	14 615	2 787	-	-	-	-	-	-	-	-	802
LAYER	11	4 425	569	-	-	-	-	-	6	-	-	802
SIMULATION	24	2 817	73	1	2	-	4	-	-	7	-	-
COMPONENTS	48	2 727	636	1	2	-	4	-	-	7	-	-
ALGORITHMS	58	1 777	102	-	-	-	4	-	-	-	-	-
WAVELENGTH	62	1 332	24	1	-	-	-	5	-	-	801	-
SWITCH	66	2 075	687	1	-	3	-	-	-	-	-	802
QOS	68	1 135	0	-	-	-	-	-	-	7	-	802
LINEAR	72	1 590	276	1	-	-	4	-	-	7	-	-
CABLE	82	1 715	551	-	-	-	-	-	-	-	-	802
AUTHENTICATION	83	1 082	18	-	-	3	-	-	-	-	801	802
VPN	85	1 007	3	-	-	3	4	-	6	-	-	-
BROADBAND	89	1 049	33	-	-	-	-	-	-	-	-	802
MODULATION	94	808	18	-	-	-	4	-	-	-	801	-
SWITCHES	95	1 107	144	-	-	3	-	-	-	-	-	802
DESTINATION	96	1 311	523	-	-	-	-	-	-	-	-	802
NETWORKING	97	1 030	64	-	-	-	-	-	-	-	-	802
MULTICAST	98	837	0	-	-	-	-	-	-	-	-	802
VENDORS	99	1 094	81	-	-	3	-	-	-	-	-	802
PROCESSES	104	1 912	1 138	-	-	-	-	-	6	7	-	-
CISCO	105	840	14	-	-	3	-	-	-	-	-	802
RELAY	107	926	64	-	-	3	-	-	-	-	-	802
DIRECTORY	109	1 251	348	-	-	3	-	-	-	-	-	802
TOPOLOGY	111	811	14	-	-	-	-	-	-	-	-	802

Table 4. Keywords distribution.

The first remarkable observation from keywords distribution is the behaviour of network and data. The two words with the highest score in the likelihood test, that is, the most significant words in TEC, are keywords only in two and one area of knowledge respectively. In fact, no word becomes key in all the sections. The highest distribution value is reached only by four lexical units (simulation, components, graph, quantum), which are present in four areas. Most keywords from the reference list, (3,509 exactly) are restricted to only one section, and a high figure (1,767) does not appear among the most relevant keywords on the individual lists. As regards the rest of distribution values, 487 words are keywords in two areas and 67 in three. From the results we may conclude that words become key as a result of all their occurrences throughout the whole corpus, and then the restriction of keywords to a particular area might be so significant as to concentrate the specialized words of such area. Chart 3 reports on the number of keywords restricted to the areas and provides some examples.

Areas	Restricted Keywords	Examples
Electronics	570	electrodes, reflectivity, fabricate, photoconductor, transmittance, nanotechnology, subthreshold, polarity, gradient, wavefront
Computing Architecture	223	configurable, mapped, microcontroller, caches, Neumann, microprocessor, verilog, flops
Telematics	558	spanning, repository, portal, applets, IGP, directories, buffered, panellist, DSA, verizon, transparency, telecities, OGSI
Signal	357	infiniband, testability, embedding, multichannel, scintillation, wavelets, images, amplitudes, lumped, bandpass, nonlinearity, capstone
Materials	150	tantalum, anisotropic, nanofibres, covalent, foams, polarizer, metallic, annealed, erbium, cascades
Business	200	compact, internationalization, teleworkers, marketplace, roamabout, globals, cookie, virtualization, entrepreneurship
Systems	238	iterations, controllers, invariant, executable, scheduler, interpolation, portability, IDL, ionosphere, debugger, pipeline
Esp. Signal	577	WAP, Hispasat, multiplexed, navigation, authenticate, comint, terrestrial, laptop, smart, GIS, offline, Raleigh, convolutional, layered
Esp. Telematics	636	data, layers, nodes, ISDN, OSI, interoperability, voip, byte, encoding, modems, identifier, hackers, payload, session, firewalls, unicast

Chart 3. Restricted Keywords.

All the statistical data so far obtained are gathered together and displayed on a table in order to get a comprehensible picture of lexical behaviour. Table 5 shows a sample of keywords followed by several values: frequency in the two corpora, keyness, distribution across areas (the figures between brackets correspond to the area/s where the word does not occur) and the areas where the lexical unit is restricted as a keyword.

Keyword	TEC Freq.	LACELL Freq.	Keyness	Distribution	
				Areas	Keywords in areas
BANDWIDTH	3,119	20	9,551.10	9	081, 802
BLUETOOTH	488	2	1,505.50	7 (-5, 7)	4, 801, 802
CHIP	1,229	353	2,340.20	9	1, 2
FIREWALL	437	33	1,147.10	7 (-2, 5)	3, 082
LAYER	4,425	569	10,604.80	9	802
NETWORK	16,679	1,686	41,784.60	9	3, 802
PROTOCOL	4,742	139	13,677.70	9	3, 802
ROUTER	3,910	25	11,974.40	8 (-5)	802
SIGNAL	7,022	641	17,922.60	9	0
WIRELESS	4,083	171	11,454	8 (-7)	4, 801, 802

Table 5. Data summary.

5. CONCLUSION

The body of language samples stored in digital format has allowed to process language so as to implement a method of analysis based on the combination of quantitative techniques and qualitative interpretations, in keeping with Corpus Linguistics. This methodology has enabled an approach to the lexical level from different perspectives and in relation to the variables of frequency, distribution, restriction and keyness.

The results so far obtained from the analysis of TEC have provided significant information on the lexical behaviour in several respects. The variable of distribution has revealed both the words shared by all the areas of knowledge in Telecommunications and those restricted to each one. Nevertheless, the joining of distribution and absolute frequency to determine the specialized nature of a lexical unit has not been effective. The application of the different combinations of those parameters has retrieved specialized and general vocabulary, and has not identified the distinctive behaviour of the categories.

Leaving aside lexical categories, the second part of the analysis has offered a new perspective of vocabulary on the basis of representativiness or statistical relevance, by identifying keywords. The fact that keywords are used more frequently in Telecommunication English than in general English is a decisive factor: the more significant a word is, the higher the probability of encountering it in specialized texts and therefore, the more useful to know how to use it. In addition, the most significant keywords have revealed the thematic content of the corpus and, thanks to keywords distribution value, it has been possible to highlight the most representative specialized words in every area.

The combination of the results yielded by the different analysis has given a good but incomplete account of the lexical behaviour in Telecommunication Engineering English. However, the set of empirical and statistical data will serve as a base of future studies and make a sound contribution to map the lexical profile of this specialized language.

6. REFERENCES

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

Aston, G. (1996). "What corpora for ESP?" http://sslmit.unibo.it/gy/pavesi.htm

Barber, C.L. (1962). "Some Measurable Characteristics of Modern Scientific Prose", *Contributions to English Syntax and Philology. Gothenburg Studies in English* 14: 21-43. Stockholm: Almquist & Wiksell.

Biber, D., Conrad, S. and Reppen, A. (1998). Corpus Linguistics. Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Cabré, M.T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Antártida/Empúries.

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

Chung, T. (2003) "A corpus comparison approach for terminology extraction". *Terminology* 9-2.

Dudley-Evans, T. and St John, M. (1998). *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.

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

Farrell, P. (1990). "Vocabulary in ESP: a lexical analysis of the English of Electronics and a study of semitechnical vocabulary". *CLCS* occasional; 25. Dublin: Trinity College.

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

Rea, C. y Carrillo, M. (2002). "Corpus lingüístico del inglés de Ingeniería de Telecomunicaciones: diseño, recopilación y expectativas" *Actas del II Congreso de la Asociación de Centros de Lenguas en la Enseñanza Superior*. Cartagena: Universidad Politécnica de Cartagena.

Robinson, P. (1991). *ESP Today: A Practitioner's Guide*. Hertfordshire: Prentice may.

Sager, Dungwort and McDonald (1980). *English Special languages. Principles and practice in science and technology*. Wiesbaden: Brandstetter Verlag KG.

Scott, M. (1997). "PC analysis of key words – and key key words", System 25-2: 233–245.

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

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

Wang, K. and Nation, P. (2004). "Word Meaning in Academic English: Homography in the Academic Word List". *Applied Linguistics* 25-3: 291-314.

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

Wynne, M. (Eds.) (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. ASDS Literature, Languages and Linguistics: Oxford.

Yang Huizhong (1986). "A New Technique for identifying Scientific/Technical Terms and Describing Science Texts (An Interim Report)". *Literary and Linguistic Computing*, 1-2: 93-103.