GETTING ON WITH CORPUS COMPILATION: FROM THEORY TO PRACTICE.

Camino Rea Rizzo
Technical University of Cartagena
Spain

ABSTRACT

The aim of this paper is first to present some theoretical guidelines on the design and compilation of a specialized corpus in compliance with Corpus Linguistics standards; second, to provide a comprehensive description of the stages followed in the creation of the Telecommunication Engineering Corpus and its characteristics; and finally, to report the immediate results obtained from the basic analysis of the corpus: statistical information and frequency list.

This paper is particularly addressed to those ESP practitioners who have to deal with the compilation of a specialized corpus for the first time, without a working knowledge of the specific subject domain they are involved in[1].

1. Introduction.

Throughout the history of English for Specific Purposes discipline, numerous studies have been carried out within the framework of Corpus Linguistics as a result of the symbiotic relationship they come to establish. Both disciplines have been connected with different and varying objectives. In one of the earliest studies, Barber (1962) intended to obtain a core word list for scientific and technical English, in the light of Thorndike's *Teacher's Word Book* (1932) and West's *General Service List of English Words* (1953) in general language. The spread of computing made arouse an interest in some scholars like Skehan (1981), who got involved in writing programs for text processing in ESP (Aston, 1996). Later, McEnery and Wilson (1996) transferred corpora to the ESP classroom so that students could directly handle data and learn from authentic samples. In the late 90's, Dudley-Evans and St. John published *Developments in English for Specific Purposes* (1998), offering an overview of the latest activity performed in the field of ESP, where they acknowledged the utility of linguistic corpora for ESP practice and the productive liaison of both disciplines.

[1] This paper has been developed as a result of the program PMPDI-UPCT-2009.

Nowadays, corpus-based research on specialized languages is still strongly promoted, particularly by university ESP teachers, in an endeavour to be responsive to students' linguistic needs. Contrary to general English, distinguished publishing companies do not provide an assortment of ESP coursebooks for every university degree and branch of specialization. So, the inherent versatility that characterizes ESP practitioners and the industrious efforts they make to bring the discourse community's authentic language into class, have led them to develop and compile specialized corpora in an attempt to compensate for the serious lack of specialized coursebooks: "*corpora can be used to provide many kinds of domain-specific material for language learning, including quantitative accounts of vocabulary and usage which address the specific needs of students in a particular domain*" (McEnery and Wilson, 1996).

The aim of this paper is to present the process of design and compilation of the Telecommunication Engineering Corpus (TEC) according to Corpus Linguistics standards. First, the concept of linguistic corpus is reviewed. Next, the essentials of corpus compilation are examined, followed by a detailed description of TEC's design, compilation and distribution, to finish with the results obtained from the basic analysis of the corpus.

This paper is dedicated to all ESP practitioners who had to face with corpus compilation as newcomers to both ESP teaching and Corpus Linguistics, and who were totally alien to the subject domain they were involved in.

2. Definition of linguistic corpus.

It is not so unusual to find some academic papers whose authors assert to have conducted a corpus-based study, when they actually have a collection of texts arranged without observing almost any rule: any compilation of texts does not make a corpus. Basically, the set of texts gathered must fulfil some specific requirements, and must be selected and ordered according to some criteria previously established in order to represent something.

The implications that the concept of linguistic corpus involves, as the term is currently understood in corpus linguistics, have been made clearer and clearer in the course of its history. Linguistic research used to be regarded as the purpose which characterized this body of texts, as shown in the definition provided by Renouf for the Cobuild project: "*a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research*" (Sinclair, 1987).

The Expert Advisory Group on Language Engineering Standards (EAGLES, 1996) remarks on the selection process of the language samples and warns about the adequacy of computing storage for data retrieval:

a) *A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.*

b) *A computer corpus is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.*

In a more recent definition, Sinclair (2004) points out the representative dimension that a corpus should acquire: "*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*." A corpus must represent something and its merits will often be judged on how representative it is (O'keefe, A., McCarthey, M., and Carter, R., 2007).

Considering all the definitions of corpus consulted in this study (Johansson, 1991; Atkins, Clear and Ostler, 1992; Sánchez, 1995; McEnery and Wilson, 1996; Biber, Conrad and Reppen, 1998; etc), we can draw the conclusion that a linguistic corpus is 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.

2.1 Specialized corpus.

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 the aim of creating a dictionary, studying the development of children language, analyzing the language used in a specific subject domain, etc., depending on the research goals.

Nowadays, thanks to the so called third generation corpora, the approach to a limited subcorpus may be straightforward. Such vast corpora allow to select an amount of data enough to constitute a specific subcorpus, like the section covering scientific language in general. Nevertheless, if the target is a narrower and more technical domain, it is necessary to compile a specialized corpus with appropriate samples so that the results of the analysis are reliable and not merely illustrative.

Corpus-based analyses start from the selection of the operating corpus. If there is not any one available for the research purpose, the first stage is devoted to the design

and compilation of an appropriate corpus. There exist some corpora related to engineering such as JDEST (Yang, 1980), HKUST (Fang, 1992), SEEC (Moudraia, 2003) and ACIA project (1999), but they do not suit our needs in several respects. Some of them do not collect samples from professional English, do cover multiple specialized languages but telecommunication is underrepresented or only include one type of text. Basically, at the beginning of our research, there was not any representative corpus of the academic and professional language used in the field of telecommunication engineering.

3. Essentials of corpus compilation.

Since the 60's, compilation, structure and size of corpora have been a moot point for corpus linguists, given that those aspects have a direct effect on the validity and reliability of the research. Several authors have suggested some guidelines which they state to be crucial factors in the success and reliability of the study. Namely, Sinclair (1991) introduces some instructions, which later he will expand into ten fundamental principles to follow in the design of a general corpus and in the compilation of language samples (Wynne, 2005):

1. *The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.*
2. *Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.*
3. *Only those components of corpora which have been designed to be independently contrastive should be contrasted.*
4. *Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.*
5. *Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.*
6. *Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.*
7. *The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.*

8. *The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.*
9. *Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.*
10. *A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.*

As regards specific corpus, we agree with Pearson (1998) on the claim that in the literature available on the topic, authors simply report on the compilation process and do not provide thoughtful arguments about the criteria governing corpus design. Therefore, Pearson combines the principles stated in different sources for their application to the creation of a specific corpus. Additionally, we may resort to the standards established for general corpus and adapt them to a specific situation.

3.1 Representativiness.

The most important aspect according to all the authors is representativiness (Renouf et al., 1987; Biber et al., 1993; Sánchez, 1995; Sánchez y Cantos, 1997; EAGLES, 1996; McEnery and Wilson, 1996; Pearson, 1998; Sinclair, 1991 and 2004; etc). The aim of representing the general usage of a language variety through a set of linguistic samples is a controversial issue. The discussion stems from the complexity found in defining representativiness itself and achieving that a fraction contains the components which confer the representative nature of the whole. Biber defines representativiness as "*the extent to which a sample includes the full range of variability in a population*" (Biber 1993).

A corpus is intended to typify a specialized language whose constituents must be identified and captured. Therefore, the corpus design must be extremely careful and the fixed criteria faithfully followed, so that the compiled data stand for the characteristics of the language under study.

In spite of the efforts made, the representativiness of a corpus will only be approximate, since it is not possible that a limited compilation of texts fully represent the whole language of a discourse community. In this respect, Leech (1991) suggests that a corpus may be considered representative when the findings obtained from its analysis could be generalized to the rest of the language that it typifies.

3.2 Size.

The size of the corpus is tightly related to its representativiness. Although the exact amount of words needed in order to achieve a representative corpus has not been definitely determined yet, there exist several options.

According to Pearson (1998), a million words is the size usually recommended for a specialized corpus. The reasons range from just intuition to solid conclusions reached in major projects, where a million words has been found to represent a reasonably high proportion of a specialized register. On the other hand, Sinclair (1991) points out that a 10 or 20-million-word corpus may constitute a useful small corpus, but it will not be appropriate for a reliable description of the whole of the language. However large a corpus might be, it will always be scarcely a tiny sample of all the language used by all the speakers of a language. Sinclair states that a corpus must be as large as possible and keep growing in order to enable the study of word behaviour in texts, as words distribute unevenly and most of them occur only once.

Kennedy (1998) asserts that a big corpus does not have to represent a register or a language better than a smaller one and concedes: "*At this stage we simple do not know how big a corpus needs to be for general or particular purposes*". Nevertheless, he proposes calculating how many word occurrences are necessary for an accurate description of different linguistic items. In general, some 40% or 50% of the lemmas in a corpus occurs once, and more than one occurrence is required to establish a comparison. On the contrary, the preposition *at*, taken as an example of high frequency words, occurs around 5,500 times in a one-million corpus, what means a quantity more than enough for descriptive purposes. It should be also taken into account that low frequency phenomena, such as collocations, require a big-sized corpus. Nevertheless, there is no sense in enlarging more and more a corpus if the analyst cannot cope with the data. A counter-argument is found in Pearson (1998), who holds that there is no justification for setting a limit on corpus size. The limit is imposed by the amount of texts available or convertible into digital format and that, at the same time, fulfil the selection criteria established.

3.3 Variety.

Corpus-based analyses have proved that there are considerable differences in the use of lexis, grammar and discourse features among language varieties. For that reason, it is essential to introduce language samples from varieties of topic, author, register, geography, source, etc.

3.4 Chronology.

Time criterion defines the span of time where the samples were produced, that is to say, the period of time that the corpus covers. In terms of time, there exist synchronic

and diachronic corpora. A synchronic corpus is a static collection of texts, selected according to some principles aiming at representing the language within a particular period of time; whereas a diachronic corpus is dynamic and systematically embraces different periods in order to study language changes and development.

With respect to a specific corpus compiled for terminological studies, it is advisable to gather samples that should have been delivered in the last 10 years prior to the date of compilation (Pearson, 1998).

3.5 Types of text or genres.

A key procedure to succeed in representing linguistic variety is to introduce samples of all type of texts or genres. The characteristic texts of a language community can be identified in the different communicative situations generated by the speakers of the community.

4. Design of the Telecommunication Engineering Corpus.

The specialized corpus has been created intentionally to serve research purposes with a careful design, so that it might be considered as reasonably representative of the written use of the language in the telecommunication engineering community. In pursuance of this aim, we have attempted to adhere as far as possible to both the criteria proposed for text selection and the guidelines for the compilation of a general corpus. These instructions have been transferred to a specific register and completed with the recommendations from the literature on specialized corpora.

In accordance with the principles presented by Sinclair (Wynne, 2005), representativiness and balance must be a prime objective even though they are out of reach and not accurately definable. In spite of that, these notions guide the design of the corpus and the selection of samples. Additionally, it is important to acknowledge that all corpora have limitations regarding time, computing, financing, text availability, etc., but a well-designed corpus creates an excellent opportunity to look into language evidence and perform quantitative and qualitative analyses.

The samples of language collected in TEC originate from real communication acts, have been prepared for computer processing and have been systematized in relation to the following criteria: topic variety, chronology, origin, mode and size.

4.1 Topic variety.

As a guide in the search of topic representativeness within the wide field of telecommunication, the curricula of two university degrees have been taken as a

reference: Telecommunication Engineer (five-year degree) and Telematic Engineer (three-year degree), at the Technical University of Cartagena in Spain.

The curricula consist of manifold areas of knowledge, and every single area has meant a thematic line to look for samples of the language. Under no circumstances is there a bias towards academic language, the corpus is intended to collect texts at random both from academic and professional language.

4.2 Chronology.

TEC is a synchronic corpus that comprises samples dating mainly from the period of time which ranges from 1997 to 2005. The selection rule is to capture the most recent texts. However, there are some exceptions in the case of subjects where older texts are used for teaching because they are essential for engineers' training.

The compilation process has been done in two phases:

Phase I.

The first stage started in 2000 and finished in 2003. During that period, the areas of knowledge concerning the three first years of each degree guided the selection of samples. The texts collected in this stage were produced between 1997 and 2003.

Phase II.

The second phase started in 2004. The corpus was enlarged by adding texts related to the two last years of telecommunication engineering which correspond to the two branches of specialization.

This stage was also devoted to complete and balance the results from the first phase. From the beginning of the compilation, language samples have been aggregated up to reaching a considerable size. Then, the collecting process was stopped to perform the analysis. Nevertheless, it is envisaged adding updated texts to the corpus so as to avoid its becoming obsolete and achieve its transformation into a dynamic corpus. The information extracted from a specific corpus is valid while the stored language keeps being used.

4.3 Origin.

As far as geographical origin is concerned, TEC embraces mainly samples from American and British varieties. Canadian, Australian English and other geographical varieties have not been included although they were taken into account at the beginning of the design process. Unfortunately, they were disregarded due to the scarcity of texts available. British and particularly American English texts outnumber the other varieties. American publications and web pages are broadly spread and predominate in this

domain, one of the factors which has let them be the outstanding representatives in telecommunication.

A third section includes samples produced by non-native speakers of English. Owing to the international communication role that English plays nowadays, a great part of scientific and technological production is written in English by non-native speakers for its circulation around the world. This section of the corpus will enable to carry out contrastive analyses of the language.

4.4 Mode.

The corpus is focused on the written mode of the language. The texts are authentic naturally occurring samples of written language, which have been produced without teaching purposes but as a real communication act among people.

The written samples correspond to complete documents, that is, they are not excerpts from longer texts but the whole. Therefore, the samples do not share the same size. Books are introduced by chapters or complete sections, research papers are fully stored and only single abstracts are kept – abstracts found detached from their origin text. In all the cases, graphs and reference section are excluded.

However, sometimes, it is really hard to delimit the text, like on web pages, where two ways have been followed: either the text appearing just on the screen is selected, or the texts found by clicking on the different links redirected by the page are also included.

4.5 Size.

TEC is made of 5,533,705 words, that is, 5.5 millions of strings of characters between two blanks. The texts related to the three first courses of Telecommunication Engineering and the three courses of Telematic Engineering account for 3,652,548 words, while the remaining 1,881,157 words are connected to the two branches of specialization.

Although the design of the corpus was intended to hold a similar amount of samples in every section so as to achieve a balanced corpus, the final size has been conditioned by diverse factors. It has not been feasible to get the same volume of words for all the components due to the heterogeneous nature of the discipline itself and its constituting subdisciplines, in addition to the fact that information does not spread equally, through the same means or in exact proportions.

Even though the final volume does not reach hundreds of million words, TEC is five times the size recommended for a specialized corpus in early proposals – 1 million words (Kennedy, 1998; Pearson, 1998; Curado, 2001). For the objective of this

research, 5 millions of words are presumed to constitute a suitable size in relation to subject domain and text variety. Undoubtedly, the more data are gathered, the clearer and sharper picture of the language may be drawn, but the only sine qua non concerning size is that the corpus must reach an appropriate sampling size. In other words, it must hold enough data for statistical inferences of lexical behaviour.

5. Corpus distribution.

The corpus has been hierarchically structured in order to facilitate the storage of the samples and guarantee an accurate classification, hence a main folder includes three big subcorpora according to the three geographical varieties. Table 1 and figure 1 show the proportions of each variety in the corpus.

Geographical variety	**Number of tokens**
1. American English (US)	3,513,282
2. British English (BE)	874,624
3. Non-native English (NN)	1,145,799
Total	**5,533,705**

Table 1. Number of tokens according to geographical varieties.

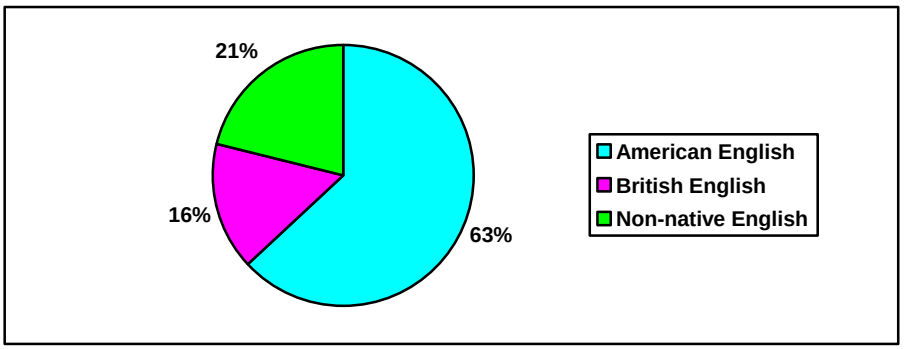

Figure 1. Corpus distribution according to geographical varieties.

Each geographical subcorpus displays seven files corresponding to the different areas of knowledge regarded in the syllabus of the degrees in Telecommunication Engineering and Telematic Engineering. On the same level as these areas, the two branches that define the specialization are included. The first one corresponds to *Telecommunication Networks and Systems* which entails an extension in signal analysis; and the second one, *Telecommunication Planning and Management*, goes deeper into Telematics. Every section is subdivided into a number of folders depending on their constituting subjects:

1. Electronics (01):

- Electronic components (01)
- Analogue Electronics (02)
- Digital Electronics (03)
- Photonics (04)

2. Computing Architecture and Technology (02):

- Computing Fundamentals (01)
- Digital Electronic Systems (02)

3. Telematic Engineering (03):

- Projects (01)
- Telematics (02)
- Distributed Information Systems (03)
- Communication Software (04)

4. Communication and Signal Theory (04):

- Signal Processing (01)
- Circuits and Systems (02)
- Electromagnetic Fields (03)
- Instrumentation (04)

5. Materials Science (08):

- Materials for Information Technology (01)

6. Business Management (06):

- Business (01)

7. System Engineering (07):

- Concurrent Systems (01)
- Control Engineering (02)

8. Supplement (08):

- Communication networks and systems (01)
- Communication planning and management (02)

In addition to these areas of knowledge, there are *Applied Physics, Applied Mathematics, Statistics* and *Operative Research* which have been disregarded because the language used is basic to understand and communicate in other subjects that integrate this type of language. Therefore, the proportion of samples is enough and

balanced with the language captured for other content subjects. Similarly, the subjects whose title refers to laboratory, supplement or extension of a main subject have also been ignored. The major subject often includes the relevant contents, so that it provides its characteristic use of the language, and the subsequent subject is devoted to broaden or put knowledge into practice, such as *Supplement to Telematics, Electronics Laboratory* and *Communication Software Laboratory.*

Within the area of Telematic Engineering, several subjects dealing with the operation of networks in their physical form have been grouped together under the label of Telematics: *Telematic Fundamentals, Telematics, Communication Services and Networks, Telecommunication Systems* and *Switching*. Likewise, in the area of Signal Processing and Communications, the label Signal Processing covers the subjects concerning the processing of signals: *Introduction to Telecommunications, Communication Theory, and Linear Systems, Digital Communications, Digital Signal Processing* and *Data transmission*. The criteria guiding the classification of subjects have been the result of the cooperation with expert teachers in the areas of Electronics, Computing Architecture, Computing Technology, Telematics and Signal Theory. Figure 2 illustrates the proportion of the corpus according to subject areas.

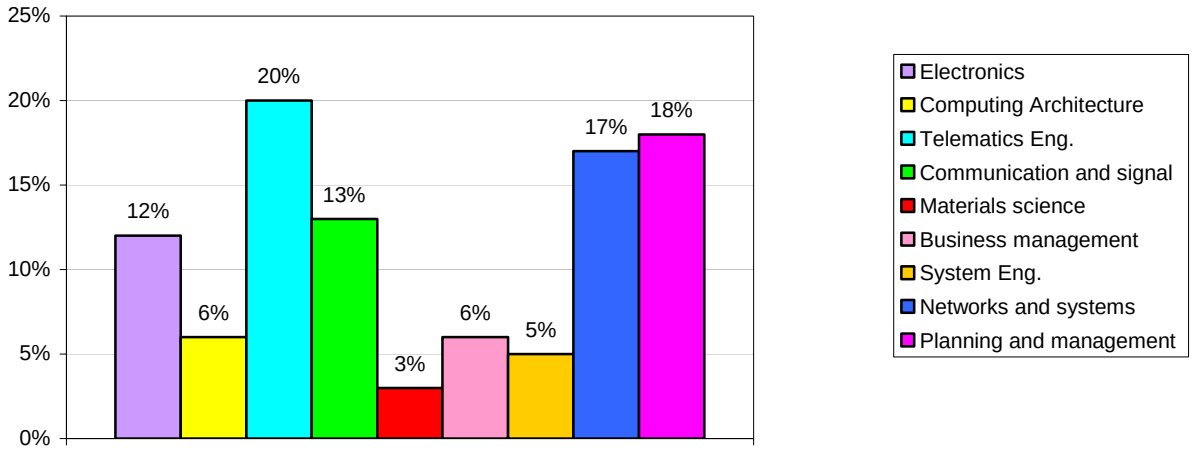

Figure 2. Corpus size according to subject areas.

Finally, every subject folder splits into eight subsections covering the variety of sources where the samples come from. If we aim at encompassing the types of text or genre typical of the professional and academic communication of the domain, different communicative situations should be taken as reference to collect language samples. Still, only the communicative situations arising when at least one user of the language is an expert or professional are regarded:

a) Communication among experts.

In this communication act, all language users are expert in the subject. Speakers, professionals and/or teachers, are assumed to share an extensive and specialized knowledge of the subject. Therefore, they are likewise expected to be proficient in the specialized language and accurately use the terminology and conventions of the discourse community they belong to.

It is highly likely that the texts produced within this communicative situation have a high density of specific terms. The characteristic types of text are research papers, datasheets, academic books, technical reports and legal documents.

b) Communication from expert to initiate.

The speaker has a wider knowledge of the subject than the initiate, so that the expert probably explains or clarifies some terms he is using in order to facilitate his audience's understanding. This communicative situation usually occurs when the reader or listener tries to increase their level of knowledge such as last year students, less experienced colleagues, experts in different aspects of the same area, etc. Textbooks illustrate the typical text in this situation.

c) Communication from expert to beginner.

When someone knows nothing about the subject is considered to be a beginner or a lay person. This kind of audience embraces mainly students who must acquire technical knowledge for training or professional purposes and also people just interested in the subject without a well-defined purpose.

Being aware of the beginner's lack of knowledge, the expert uses the appropriate terminology or a simplified technical language in his discourse, and introduces clarifications more frequently so that beginner learns and understands the specific meaning assigned to the terms in the domain. This type of language is found in articles published in popular magazines, elementary textbooks introducing a subject, instruction manuals, etc.

The last eight subsections previously mentioned which cover the sources where the samples come from have been arranged in the corpus as follows and as depicted in figure 3:

1. Magazines (01). Articles from popular magazines on the subject.
2. Books (02). Texts from scientific books, textbooks, technical reports, and the so called tutorials (lectures and explanations about a topic in formal style, spread by teachers, scientists, professionals and companies).
3. Web (03). Texts from web pages available on the internet and not be classifiable into the rest of sections.

4. Research papers (04). Papers addressed to experts in a subject with the aim to report the process, condition and/or results of a piece of research, which have been presented in conferences or published in scientific specialized journals.
5. Abstracts (05). Short texts which precede and summarize the content of a speech, article or book.
6. Brochures (06). This section contains brochures, booklets, leaflets and instruction manuals.
7. Advertising (07). Advertisements related to any aspect of telecommunications.
8. Technology News (08). News related to any aspect of telecommunications.

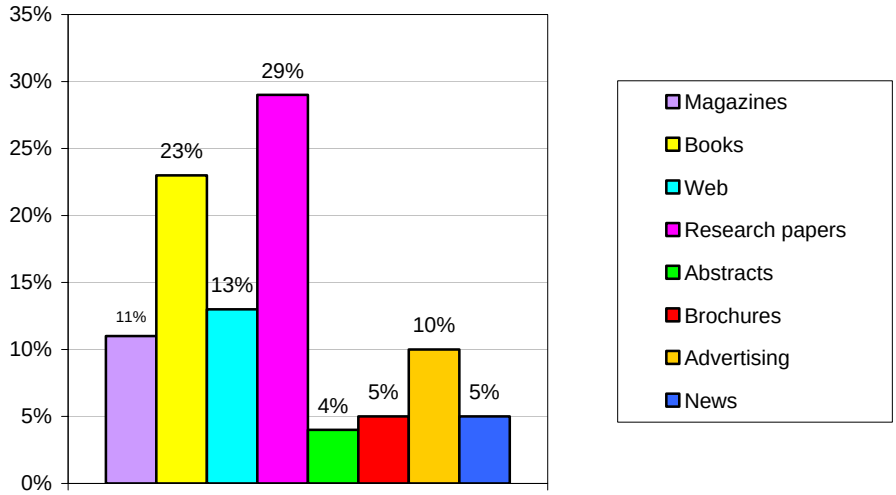

Figure 3. Corpus size according to text types.

The number and abbreviation appearing next to every section make part of the code labelling each text. All the language samples are stored under an embedded structure and headed by a code which allows a subsequent indexation of the corpus. This structuring code enables to trace the origin and all the information available on the samples when they are retrieved on the screen. Figure 4 displays the hierarchical structure applicable to the samples from British English in the area of *Computer Architecture and Technology* and their corresponding code:

Telecommunications Corpus
British English BE
Computing Architecture and Technology 02
Computing Fundamentals 01
Magazines 01
Books 02
Web 03
Research 04
Abstracts 05

Brochures 06
Advertising 07
Technology News 08
Digital Electronic Systems 02
Magazines 01
Books 02
Web 03
Research 04
Abstracts 05
Brochures 06
Advertising 07
Technology News 08

Figure 4. Coding for ATC in British English.

To sum up, 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).

6. First approach: basic statistical information and frequency list.

Once the appropriate corpus has been compiled, it is processed by using WordSmith (Scott, 1998). The immediate data obtained are those related to the basic statistical information (Table 2) and the frequency word list (Table 3).

6.1 Basic statistical information.

The program counts 5,533,705 tokens defined as sequences 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. 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. Those ratios provide information on the corpus lexical diversity from different perspectives.

Type/token ratio is obtained from the division of the whole number of different forms by the number of occurrences and later 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. The probability to find new forms decreases gradually as the length of the text increases. Likewise, the acquisition of new words substantially diminishes as the sample of language grows, since the rise in new forms and their frequency follows the curve of a hyperbola (Sánchez and Cantos, 1997) (Figure 5). However, a specialized corpus might be expected to have more forms than a general corpus, owing to the nature of the specialized discourse where speakers need technical terms to convey specific concepts accurately.

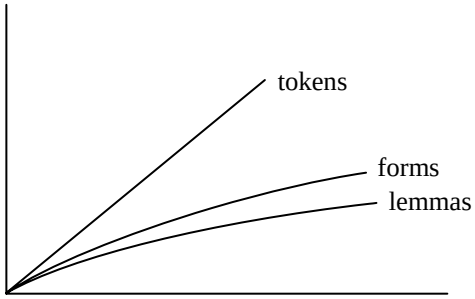

Figure 5. Relationship among tokens, forms and lemmas.

The program computes the standardized type/token ratio every *n* words, being *n*=1,000. The ratio is calculated for the first 1,000 tokens, then computed for the next 1,000 and successively until the end of the text, 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.

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 2). 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.

TEC	
Tokens	5,533,705
Types	59,826

Type/token Ratio	1.08
Standarised Type/token	38.26
Sentences	223,278
Sentences length	23.87
Paragraphs	30,472
Paragraphs length	102.95
Hapax legomena	21,755

Table 2. Basic statistical information.

6.2 Frequency list.

Due to the large amount of data available from the corpus, the frequency list is illustrated in table 3 which displays the most frequent 200 words in TEC.

TEC					
1	THE	374,598	101	NO	5,118
2	OF	170,493	102	MUST	5,087
3	AND	145,104	103	WOULD	5,072
4	TO	140,361	104	CURRENT	5,064
5	A	131,832	105	VALUE	5,063
6	IN	104,696	106	POWER	5,031
7	IS	94,630	107	HOW	5,002
8	FOR	63,834	108	TECHNOLOGY	4,969
9	THAT	52,034	109	SHOULD	4,941
10	ARE	41,713	110	MANY	4,935
11	BE	41,543	111	YOUR	4,916
12	AS	40,644	112	MANAGEMENT	4,890
13	THIS	37,780	113	ABOUT	4,879
14	WITH	36,388	114	BECAUSE	4,841
15	ON	33,072	115	HOWEVER	4,827
16	BY	31,017	116	PROTOCOL	4,742
17	IT	28,218	117	WHAT	4,616
18	AN	27,199	118	TYPE	4,612
19	CAN	25,288	119	DO	4,596
20	OR	24,488	120	WORK	4,594
21	FROM	21,405	121	SOFTWARE	4,575
22	AT	20,529	122	FREQUENCY	4,551
23	WE	17,733	123	INTERNET	4,504
24	WHICH	17,453	124	USERS	4,493
25	NOT	17,069	125	APPLICATION	4,478
26	NETWORK	16,649	126	STATE	4,458
27	WILL	16,128	127	LAYER	4,425
28	HAVE	16,114	128	CASE	4,375
29	DATA	14,613	129	SINGLE	4,368
30	HAS	13,444	130	INPUT	4,347
31	ONE	13,218	131	TRAFFIC	4,345
32	SYSTEM	12,624	132	MOBILE	4,341
33	TIME	12,391	133	POINT	4,236
34	USED	11,874	134	OUTPUT	4,139

35	IF	11,826	135	OUT	4,129
36	ALL	11,602	136	ORDER	4,120
37	THESE	11,577	137	PROVIDE	4,119
38	MORE	11,454	138	WELL	4,096
39	YOU	11,448	139	WIRELESS	4,083
40	ALSO	10,382	140	SUPPORT	4,077
41	OTHER	10,291	141	NEED	4,013
42	USE	10,255	142	WHILE	4,010
43	EACH	10,224	143	ADDRESS	3,951
44	SUCH	10,006	144	CIRCUIT	3,932
45	ITS	9,714	145	SECTION	3,912
46	WHEN	9,681	146	ROUTER	3,910
47	WAS	9,664	147	END	3,872
48	SYSTEMS	9,479	148	LINK	3,853
49	TWO	9,407	149	VERY	3,835
50	USING	9,214	150	C	3,819
51	INFORMATION	9,161	151	RESULTS	3,712
52	BUT	8,933	152	FIG	3,702
53	THEIR	8,930	153	AREA	3,680
54	THEY	8,734	154	WITHIN	3,677
55	BASED	8,448	155	RESEARCH	3,672
56	BETWEEN	8,403	156	AVAILABLE	3,627
57	NEW	8,347	157	B	3,609
58	THAN	8,052	158	WAY	3,599
59	I	8,004	159	DIGITAL	3,595
60	ONLY	7,939	160	SECURITY	3,578
61	MAY	7,808	161	PACKET	3,577
62	NUMBER	7,759	162	SERVER	3,574
63	DESIGN	7,701	163	SEE	3,552
64	THERE	7,494	164	THREE	3,545
65	FIGURE	7,325	165	ROUTING	3,542
66	INTO	7,308	166	THEM	3,531
67	BEEN	7,150	167	INTERFACE	3,526
68	CONTROL	7,124	168	SHOWN	3,507
69	SERVICE	7,085	169	BIT	3,479
70	SIGNAL	7,022	170	RATE	3,467
71	SOME	6,904	171	COULD	3,466
72	E	6,607	172	SINCE	3,452
73	EXAMPLE	6,379	173	DEVICES	3,430
74	HIGH	6,348	174	LIKE	3,415
75	S	6,315	175	N	3,408
76	FIRST	6,314	176	LOW	3,404
77	USER	6,292	177	FUNCTION	3,380
78	DIFFERENT	6,198	178	GIVEN	3,368
79	SERVICES	6,161	179	FOLLOWING	3,361
80	ANY	6,052	180	DOES	3,338
81	ACCESS	5,999	181	PROBLEM	3,273
82	PROCESS	5,949	182	PROVIDES	3,257
83	OVER	5,925	183	POSSIBLE	3,221
84	SAME	5,904	184	COST	3,214
85	THEN	5,897	185	CHANNEL	3,212
86	MODEL	5,895	186	FIELD	3,205
87	SO	5,861	187	MAKE	3,200
88	NETWORKS	5,832	188	NOW	3,164

89	SET	5,813	189	BEING	3,162
90	PERFORMANCE	5,686	190	COMMUNICATION	3,159
91	UP	5,652	191	COMPUTER	3,156
92	MOST	5,602	192	ANALYSIS	3,153
93	WHERE	5,463	193	COMMUNICATIONS	3,144
94	APPLICATIONS	5,414	194	SOURCE	3,120
95	WERE	5,378	195	BANDWIDTH	3,119
96	BOTH	5,355	196	CODE	3,112
97	LEVEL	5,309	197	STANDARD	3,081
98	IP	5,239	198	MULTIPLE	3,046
99	THROUGH	5,202	199	APPROACH	3,043
100	OUR	5,152	200	DEVELOPMENT	3,032

Table 3. The most frequent 200 words in TEC.

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 4, *the* is the most frequent word in the corpus and stands for 6.77% of the total tokens. In general language, the 3 most frequent words commonly reach an 11% of the whole, the 10 most frequent ones a 22%, the 50 most frequent ones a 37%, the 100 most frequent ones a 44% and the 2,000 most frequent words cover around the 80% (Schmitt, 2000). Those figures agree with the results obtained from TEC with only some slight differences:

Most frequent words	**Coverage in General language**	**Coverage in TEC**
3	11%	12%
10	22%	23%
50	37%	36%
100	44%	42%
2.000	80%	79%

Table 4. Coverage of the most frequent words.

From the 5.5 million-word sample, 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. In brief, more than half of the text is 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.

The distinction between notional and functional words permits to measure lexical complexity in a text by the lexical density index, that is, the proportion of notional words given in a percentage. This is obtained from dividing the number of content words by the number of tokens and then multiplied by 100. The lexical density index is higher in those texts which contain a greater proportion of notional words. According to Ure (Ure, 1971 in Stubbs, 2001), lexical density in written texts tends to be 40%, ranging from 36% to 57%. As shown in table 5, the resultant value for TEC gets close to the superior level of lexical density.

TEC	
Content words	3.076.453
Tokens	5.533.705
Lexical density	55,59%

Table 5. Lexical density.

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 3 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 greater number of notional words found in the high frequency levels may be another indicator of TEC's lexical density. Besides, the fact that many of the most frequent 50 notional words are of specialized character (Table 6), might lead to expect a high presence of technical terms in TEC.

TEC			
1	NETWORK	26	PROCESS
2	DATA	27	MODEL
3	SYSTEM	28	NETWORKS
4	TIME	29	SET
5	USED	30	PERFORMANCE

6	USE	31	WHERE
7	SYSTEMS	32	APPLICATIONS
8	USING	33	BOTH
9	INFORMATION	34	LEVEL
10	BASED	35	IP
11	NEW	36	CURRENT
12	ONLY	37	VALUE
13	NUMBER	38	POWER
14	DESIGN	39	TECHNOLOGY
15	FIGURE	40	MANAGEMENT
16	CONTROL	41	PROTOCOL
17	SERVICE	42	TYPE
18	SIGNAL	43	WORK
19	EXAMPLE	44	SOFTWARE
20	HIGH	45	FREQUENCY
21	FIRST	46	INTERNET
22	USER	47	USERS
23	DIFFERENT	48	APPLICATION
24	SERVICES	49	STATE
25	ACCESS	50	LAYER

Table 6. The most frequent 50 content words.

7. Conclusion.

This paper has meant to be a constructive suggestion to overcome the difficulties arising when it comes to compiling a specialized corpus. Although the central principles underpinning Corpus Linguistics are firmly established and their application in the field of ESP practice is extremely beneficial, there exist some gaps in the literature with respect to the guidelines applying in specialized corpus compilation. Therefore, a review of both the concept of linguistic corpus and the criteria established for gathering general samples has been undertaken, to follow up on their practical application in case of specialized corpora. The development of TEC has been fully described to set an example and offer a detailed account of the successive stages and decision-takings.

The results obtained from the basic analysis of TEC have pinpointed just a minimal part of the data potentially available, both within the scope of ESP teaching and linguistic research. A further stage in the analysis proceeds to adopt a comparative approach to assess to what extent the specialized variety differs from the general language. For this purpose, a large general corpus is required to establish the reference norm which the specific corpus is contrasted to.

8. References.

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

Atkins, B. and Zampolli, A. (1994) *Computational approaches to the lexicon*. Oxford: OUP.

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

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

Biber, D. and Conrad, S. (2001) “Quantitative Corpus-based Research: Much More Than Bean Counting”. *TESOL Quarterly* 35, 2: 331-336.

Curado Fuentes, A. (2001) *A Lexical Common Core in English for Information Science and Technology*. Cáceres: Universidad de Extremadura, Servicio de Publicaciones.

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

EAGLES (1996) “Preliminary Recommendations on Corpus Typology”. Expert Advisory Group on Language Engineering. EAG-TCWG-CTYP/P.

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.

Leech, G. (1991) “The state of the art in corpus linguistics.” In Aijmer, K. and Altenberg, B. (Eds.) *English corpus linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

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

Moudraia, O. (2004) “The Student Engineering English Corpus”. *ICAME Journal* 28: 139-143.

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.

Sarmiento, R. y Sánchez, A. (2005) Corpus para fines lexicográficos y de análisis gramatical: el corpus Cumbre. *Oralia*, 8: 57-79.

Sánchez, A., Cantos, P., Sarmiento R., Simón, J. (1995) *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Sánchez, A., Cantos, P. (1998) El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas. *Atlantis* 1, 2: 205-227.

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

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

Sinclair, J. (1987) *Looking up. An account of the Cobuild Project in lexical computing*. London: Collins.

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

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.

Stubbs, M. (2001) *Word and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.

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.