the ESP, São Paulo, vol. 24, nº 1 57-78

THE COMPILATION OF AN ENGLISH CORPUS OF BIOLOGY: SOME REMARKS ON SCIENTIFIC VOCABULARY

A Compilação de um Corpus Formado por Textos de Biologia em Inglês: algumas observações sobre vocabulário científico

Purificación SÁNCHEZ HERNÁNDEZ (Departament of English Phylology,University of Murcia, Spain)

Abstract

This paper deals with the elaboration of a corpus of Biology texts designed as a whole but integrated by different sub-disciplines. The compilation of the corpus was carried out taking into account the credits that the different sub-areas are given in the University studies of Biology and also the scientific and social impact of the subjects. We have gathered a corpus of 2,500,000 words with a percentage of 84% of the total texts devoted to scientific journals and 16% to books. In the same way 70% of the texts compiled were from American sources and 30% from British ones. Initially we focused on the lexical aspects of the corpus. Firstly, we have shown the utility of our specialised Corpus as compared with a general English one. After compiling the texts we extracted the 150 most frequent words of each file. As expected, the highest frequencies are those corresponding to grammar forms. A selection of the technical and sub-technical terms in these files revealed that the highest lexical density is found in books. In the same way, we have proved that the design we have proposed serves its purposes, as far as the lexical terms are concerned.

Key-words: *corpus; biology; scientific vocabulary.*

Resumo

Este artigo trata da elaboração de um corpus de textos de Biologia organizado como um todo mas composto por diferentes sub-disciplinas. A compilação do corpus foi realizada levando-se em consideração os créditos atribuídos às diferentes sub-áreas nos estudos de Biologia *na Universidade e também o impacto científico e social dos assuntos. Nosso corpus possui 2.500.00 palavras, sendo que uma porcentagem de 84% do total de textos é devotada a periódicos científicos e 16% a livros. Da mesma forma, 70% dos textos compilados são provenientes de fontes norte-americanas e 30% de fontes britânicas. Inicialmente, focalizamos os aspectos lexicais do corpus. Em primeiro lugar, demonstramos a utilidade do nosso corpus especializado em comparação com um corpus de inglês geral. Após compilar os textos, extraímos as 150 palavras mais freqüentes de cada arquivo. Conforme era esperado, as freqüências mais altas são aquelas correspondentes a formas gramaticais. Uma seleção dos termos técnicos e sub-técnicos presentes nesses arquivos revelou que a densidade lexical mais alta é encontrada nos livros. Da mesma forma, provamos que a organização que propusemos serve aos nossos propósitos, no que se refere aos termos lexicais.*

Palavras-chave: *corpus; biologia; vocabulário científico.*

1. The compilation of an English corpus in biology

The Department of English Philology has the task of teaching Language for Specific Purposes (LSP) in the different Degrees offered by the University of Murcia, Spain. This work originated in the course of 1999-2000 when I was responsible for teaching a course of “Biomedical English” to students in their third year of the Degree in Biology. The aim of the course was to develop the students’ communicative competence in English within their area of studies. I could not find a text-book appropriate to accomplish this goal. On this occasion I developed a program based on the exploitation, both lexical and syntactical, of texts from *English in Basic Medical Science* (1996) and *A Course in Intermediate Scientific English* (1989) for the students to get used to the actual language they are exposed to in their study of science. At the same time, and considering that authentic materials ensure an accurate representation of real usage, I designed a large corpus of biological English that could be used both for research and pedagogical applications.

The purpose of this project is to provide a machine-readable corpus to serve as the empirical basis for a number of specific languages and amenable to perform contrastive or comparative analysis of LSP texts. The corpus was created having in mind a wide range of different research interests: discourse, syntax, semantics and lexis. The terminological applications of this corpus will be probably considered in the future. The corpus contains 2,494,772 words and was compiled from March 2000 till May 2001.

1.1. Textual universe and bibliographies

The textual universe chosen was the language of Biology. This is a field central to the activities of the university community in general and thus also to the activities of the scientific community.

Many studies in scientific English have been carried out on corpora (Biber et al., 1998; Flowerdew, 1993; Johanson, 1975). In all of them Biology has been considered as a whole. Our corpus differs from others in that we designed it considering the different sub-disciplines present in Biology. As the texts are intended to reflect contemporary Biology language usage, the bibliographies cover texts published or written within the 6-year period going from 1994 to 2000. The textual universe was established on the basis of available bibliographical sources.

Since I am not an expert in Biology, one of the initial problems to be faced in this corpus was the distribution of the texts within the different sub-areas of Biology[1] . Finally, the structure of the corpus was designed considering the impact of the different sub-areas both on science and on society and also the credits they are given to complete the Degree in Biology at the University of Murcia.

---

[1] In this context, I was kindly advised by Dr. M. Gacto, Professor of Microbiology and an internationally recognised researcher in his field.

1.2. Criteria for the elaboration of the corpus

For the selection of texts two sets of criteria have been applied. First, texts were selected on the basis of the credits given to each subject in the Degree in Biology, and secondly on the basis of the scientific and social impact of the subject area. In this way, the assembled bodies of texts may be claimed to be reasonably representative of the textual universe and of the scientific language usage in the field of Biology.

The following distribution and percentage were used:

Biochemistry	15%
Genetics	15%
Microbiology	15%
Animal Physiology	10%
Cytology	10%
Ecology	10%
Plant Physiology	10%
Botany	7.5%
Zoology	7.5%

Table 1: Distribution and percentage of the texts selected

Indeed, this classification could have been made differently with other category divisions of the subject area, but, on the whole, this thematic division served its operational purpose.

The nine categories of the thematic classification were weighted according to their relative importance to ensure that the texts which form part of the corpus offer a representative picture of the textual universe of the bibliographies. The themes which have been given the highest percentages of textual coverage are those which are most central to Biology and which must therefore be assumed to result in the largest production of texts. As can be seen from the above classification, Biochemistry, Genetics and Microbiology were given priority because they form the core of the currently expanding Molecular Biology.

1.3. Selection of texts

A maximum limit of textual length of 250,000 running words per subject was taken as a norm. With this length we considered that the texts provide empirical data for research at different linguistic levels, including that of discourse.

All the texts collected have been published in journals and books that are copyright registered by major indexing services. Other "published" texts that are not copyright registered include government reports and documents (Biber, 1993).

We took into account the fact that the language of science is continuously evolving and that new words are being introduced. Hence, 84% of the words were taken from scientific journals of wide diffusion in the scientific community reporting novel findings and using new experimental techniques. All of these journals were included among those showing high impact according to the *Journal Citation Reports* (1998). In some cases, as for instance in Ecology, the texts were also taken from official reports. The remaining 16% of the words have been gathered from recent books. We understand that the different percentages given to journals and books correspond to the proportion of usage of these two sources by the scientific community. We also decided to include texts both from American English (70%) and British English (30%), some journals published in The Netherlands being considered as BE. Other varieties of English have not been measured, since no scientific journals of wide circulation are being published outside USA or UK.

The texts of the journals were mainly downloaded from the Internet. For each research article samples were taken from the few major sections: Introduction, Methods, Results and Discussion when possible. In some cases we have only had access to abstracts. As for the university-level textbooks, the samples were converted to computer-readable forms with the help of a scanner and incorporated as text files. Each text in the corpus was assigned to a separate file according to a thesaurus.

The actual figures for the distribution of texts in the English corpus are shown in Table 2.

		Journals (84%)		Books (16%)	
Subject	**N. of words**	**American English (70%)**	**British English (30%)**	**American English (70%)**	**British English (30%)**
Biochemistry	375,000	220,500	94,500	42,000	18,000
Microbiology	375,000	220,500	94,500	42,000	18,000
Genetics	375,000	220,500	94,500	42,000	18,000
Animal physiology	250,000	147,000	63,000	28,000	12,000
Plant physiology	250,000	147,000	63,000	28,000	12,000
Ecology	250,000	147,000	63,000	28,000	12,000
Cytology	250,000	147,000	63,000	28,000	12,000
Botany	187,500	110,250	47,250	21,000	9,000
Zoology	187,500	110,250	47,250	21,000	9,000

Table 2: Distribution of words according to sub-area, English register and source

It was sometimes very difficult to get the exact number of words for each subject, so that in the final count we had a few words more of a sub-area and a few words less of another one. This situation, however, did not significantly change the final balance.

After collecting the texts they were indexed according to a thesaurus previously designed in order to have an almost instantaneous retrieval of data.

2. Some remarks about vocabulary in this corpus

Language varieties can be distinguished along several dimensions, especially social and regional dialect, style and register

(Johansson, 1975: 1). According to Halliday et al (1964:88) "the crucial criteria of any given register are to be found in its grammar and its lexis. Probably lexical features are the most obvious… purely grammatical distinctions between the different registers are less striking, yet there can be considerable variation in grammar also." Scientific technical terms or technical vocabulary are the clearest signals of a particular register. Technical vocabulary has a very narrow range, that is, it is used within a specialised field.

Vocabulary, in general, is central to language and of critical importance to the typical language learner (Zimmerman, 1997: 5), and especially if the learner is trying to get proficiency in ESP. Nevertheless, the teaching and learning of vocabulary have been undervalued in the field of second language acquisition.

Taking into account the above considerations, we will show that a wide range of technical vocabulary can be obtained from the files that form our Corpus and that the design we did for the collection of the text samples serves its purpose. The subdivision of Biology into different sub-disciplines and the different sources employed (American Journals, British Journals, American Books and British books) provide us with vocabulary lists covering a wide range of technical words. In some instances, the words selected may be considered as not exactly technical but sub-technical, that is, they are words of general usage but with a especial meaning within the technical area (Inman, 1978). This type of vocabulary is often incorrectly used by students but rarely recognised as a problem (Martin, 1976). These lists can be used to make it easier for teachers and learners to present the type of vocabulary in the same way as high-frequency vocabulary – normally, by learning these items directly through vocabulary exercises or individual learning (Nation & Newton, 1997: 240).

2.1. Procedure

As a first step we took the frequency lists of all the sub-areas both from journals (US and UK) and books (US and UK). We had 36 files in total. We limited our study to the first 150 words occurring in each file.

2.2. Lexical items in the specialist corpus vs. lexical items of a general corpus

In order to illustrate the utility of a Corpus of Biology we compared the first twenty nouns appearing in a corpus of English (Lacell Corpus, 12.500.000 words in Lacell[2]) and those in our Corpus of Biology and we have obtained the following results:

General corpus	*Specialist corpus*
Education	Cells
School	Protein
Schools	Cell
Times	Gene
Time	Species
Pounds	DNA
Work	Genes
Children	Proteins
Supplement	Growth
People	Activity
University	Expression
Students	Different
Year	Sequence
Teachers	Results
Years	Analysis
Way	Plants
Features	Plant
Research	Data
Cent	Acid
Voice	Number

Table 3: The first twenty nouns appearing in a corpus of English and in a Corpus of Biology

[2] A Corpus of English based on the design of the Spanish Cumbre Corpus, so as to have two equivalent corpora.

As can be seen none of the top 20 nouns in the corpus of English occurs among the top 20 nouns of the specialised corpus. And even in the case there were items common to both the general and the specialist corpus, the items in this last one may have different uses, which will be corroborated through concordancing later in this paper.

2.3. Frequency as a criterion for course design

The novelty of our work with respect to some previous corpora of Biology lies on the very design of the corpus, that is, on the fact that the samples of texts have been taken from the different sub-areas that form Biology. This way we ensure the presence of a large number of different terms and a total coverage as far as the vocabulary of the area is concerned. The advantage that this design and the size of this Corpus presents is that the resulting material can be used for teaching purposes in any course independently of the level of the students and the time devoted to the syllabus.

To justify our approach we have collected the two most frequent terms related to science from the different areas and the results are as follows on Table 4.

As shown on Table 4 there are many terms that are unique to the sub-area even in the top frequency range and that would not appear in the whole Corpus if we had not kept this design.

As our interest was the scientific vocabulary present in the whole corpus, and in order to show the appropriateness of the design regarding the different sources from which texts have been taken, we selected only the words pertaining to technical vocabulary[3] in each file and we observed that as a rule, most of the technical words appeared in the second half of the list. Also, we found that the first words of the scientific

---

[3] Under the expression "technical vocabulary" we include highly technical words (words pertaining exclusively to the speciality) and sub-technical words (words which are not specific to a subject speciality and which occur regularly in one field of knowledge) (Kennedy & Bolitho, 1991; Dudley-Evans & St John, 1998).

register occurred between the first ten and twenty words in the books files, whereas in journals they are located after the twenty first words. This happens probably owing to the very nature of the content of journals and books. Journals offer new findings in science and a lot of wording is needed to present an experiment and report results of a large variety of topics (Joyce J. Repa and David J. Mangelsdorf: "The role of orphan nuclear receptors in the regulation of cholesterol homeostasis". *Rev. Cell. Dev. Biol.* 2000, **16:** 459-481). Books, on the contrary, usually deal with a single topic (*Extracellular Matrix*. Camper, W.D. (ed.) 1996 Amsterdam: Harwood Academic Publishers).

File	**USA Journals**	**UK Journals**	**USA Books**	**UK Books**
Genetics	Gene DNA	Gene Sequence	DNA Genes	Genes Embryos
Biochemistry	Protein Cells	Protein Activity	Copper Resistance	Peptide Virus
Microbiology	Proteins Cells	Species Strains	DNA Gene	Microbes Bacteria
Animal Physiology	Body Mass	Cells Activity	Animal Research	Water Oxygen
Citology	Cell Protein	Cell Protein	Cell Bacteria	Elastin Collagen
Ecology	Bottleneck Depression	Nuclear Climate	Species Biodiversity	Species Record
Plant Physiology	Plants Cells	Plant Gene	Cells Plants	Sucrose Gene
Zoology	Species Mass	Species Behaviour	Animals Research	Water Sodium
Botany	Species Pollen	Species Seed	Plants Water	Flooding Growth

Table 4: The two most common terms of the scientific register in each sub-area

Once the lists were obtained all the figures were summed up, and the results (listed in Table 5) showed that almost the same number of technical words resulted from each list, irrespective of the difference in length of the files under study.

Source	**Number of words**	**Number of technical words found**
USA Journals	1,470,000	490
UK Journals	630,000	454
USA Books	280,000	428
UK Books	120,000	488

Table 5: Source, number of words and number of technical words found in the lists

It is not immediately clear how a comparison at the level of vocabulary between large files should proceed (Ljung, 1991). In the present case, the problem increases due to the great discrepancy in length between the different files of our corpus. That is why we determined the relationship between technical words and total length of the files (lexical density) of journals and books from the USA and UK (Table 6).

Source	**Relationship between technical words and total length of the files**
USA Journals	0.033
UK Journals	0.077
USA Books	0.15
UK Books	0.40

Table 6: Relationship between the technical words encountred and the total length of the files under study

We found that the file with fewer words (UK Books) maintains the highest relationship encountered between the number of technical items and the number of words, followed by USA Books, UK Journals and USA Journals. Therefore, it seems that in comparative terms the larger the file, the shorter the list of new words, as previously reported by Sanchez and Cantos (1997).

However, to focus on this question we should look at the words in detail and not at the figures as a whole. Among the obvious approaches to clarify this point one is to determine how many of the words in each file are unique to that particular file. If we had individual words in our lists we should have to proceed with caution because the selection could be largely due to chance (Ljung, 1991). Apparently, that is not our case: the less repeated term in all the files analysed is *filament*, that appears 9 times in the file Zoology books UK. According to Ljung (1991), if words from the top frequency band in some files are missing from the others, that is a fairly strong indication that the two files differ in ways which are not due to chance, which again reinforces the idea that we can obtain most of the lexical terms from our Corpus.

We have made a global account of the relationship between size of the file and its lexical density. Now we will study in detail the behaviour of words in a sub-area.

2.4. Analysis of the vocabulary present in one sub-area

To illustrate our point of view we studied in detail the Genetics sub-area as a parameter to control the words appearing in all the files of the same sub-area. The results are presented in Tables 7 and 8.

USA Journals			UK Journals		
Type	**N of oc.**	**Frequency**	**Type**	**N of oc.**	**Frequency**
Acid	142	0.0877	Activity	67	0.1089
Activation	160	0.0989	Analysis	131	0.2130
Activity	190	0.1174	Assembly	47	0.0764
Amino	145	0.0896	Cell	89	0.1447
Analysis	234	0.1446	Cells	181	0.2943
Binding	253	0.1564	Cerevisiae	105	0.1707
Cell	249	0.1538	Chromatin	54	0.0878
Cells	356	0.2200	Chromosome	101	0.1642
Chromosome	248	0.1532	Complex	51	0.0829
Coli	250	0.1545	Control	61	0.0992
Data	165	0.1019	Data	109	0.1772
DNA	668	0.4319	DNA	228	0.3707
Drosophila	137	0.0846	Domain	44	0.0715
Effect	122	0.0754	Encodes	44	0.0715
Expression	293	0.1810	Encoding	60	0.0976
Fragment	204	0.1260	Expression	182	0.2959
Fragments	114	0.0704	Figure	99	0.1610
Function	154	0.0952	Function	70	0.1138
Gene	770	0.4758	Genes	298	0.4846
Genes	472	0.2916	Gene	457	0.7431
Genetic	306	0.1891	Genetic	71	0.1155
Genetics	292	0.1804	Genome	107	0.1740
Glucose	187	0.1155	Genomic	60	0.0976
Growth	233	0.1440	Growth	53	0.0862
Insertion	116	0.0717	Human	196	0.1724
Loci	149	0.0921	Identified	68	0.1106
Locus	143	0.0884	Level	48	0.0781
Medium	135	0.0834	Levels	48	0.0781
Ml	118	0.0729	Mitochondrial	115	0.1870
Mutant	352	0.2175	Mutant	118	0.1919

USA Journals (cont.)			UK Journals (cont.)		
Type	**N of oc.**	**Frequency**	**Type**	**N of oc.**	**Frequency**
Mutants	364	0.2249	Mutants	63	0.1024
Mutation	292	0.1804	Mutation	50	0.0813
Mutations	407	0.2515	Pombe	45	0.0732
Operon	139	0.0859	Protein	198	0.3220
Phage	125	0.0772	Proteins	73	0.1187
Phenotype	125	0.0772	Rcaf	55	0.0894
Plasmid	178	0.1100	Region	101	0.1642
Promoter	222	0.1372	Regions	50	0.0813
Protein	427	0.2638	Research	51	0.0829
Proteins	226	0.1396	Results	89	0.1447
Rbfa	117	0.0723	RNA	97	0.1577
Recombination	212	0.1310	Saccharomyces	77	0.1252
Region	330	0.2039	Sequence	246	0.4000
Regions	117	0.0723	Sequences	123	0.2000
Results	218	0.1347	Species	60	0.0976
Sequence	367	0.2268	Strain	56	0.0911
Sequences	194	0.1199	Strains	87	0.1415
Site	281	0.1736	Synexpression	45	0.0732
Sites	165	0.1019	Transcription	61	0.0992
Species	120	0.0741	Wild-type	47	0.0764
Strain	243	0.1501	Yeast	126	0.2049
Strains	334	0.2064			
Structure	118	0.0729			
Transcription	215	0.1328			
Transcriptional	127	0.0785			
Type	128	0.0791			
Wild-type	148	0.0914			

Table 7: Record of the technical words of the GENETICS sub-area from USA and UK journals

USA Books			**UK Books**		
Type	**N of oc.**	**Frequency**	**Type**	**N of oc.**	**Frequency**
Blood	37	0.0991	Abdominal	29	0.1337
Body	47	0.1259	Activity	41	0.1898
Cells	153	0.4099	Analysis	31	0.1429
Chromosome	86	0.2304	Bicoid	56	0.2581
Chromosomes	104	0.2786	Cell	71	0.3272
Clone	51	0.1366	Cells	104	0.4794
Cloning	69	0.1849	Cleavage	35	0.1613
Data	34	0.0911	Cycle	22	0.1014
Disease	89	0.2385	Cytoplasm	29	0.1337
Disorders	43	0.1152	Cytoplasmic	25	0.1152
DNA	375	1.0047	Development	57	0.2627
Gene	204	0.5466	DNA	28	0.1291
Genes	258	0.6913	Domain	26	0.1198
Genetic	223	0.5975	Dorsal	55	0.2535
Genome	108	0.2894	Egg	31	0.1429
Health	36	0.0965	Eggshell	22	0.1014
Human	155	0.4153	Embryo	71	0.3272
Map	69	0.1849	Embryonic	86	0.3964
Mapping	42	0.1125	Embryos	145	0.6683
Maps	38	0.018	Experiments	30	0.1383
Markers	37	0.0991	Expressed	48	0.2212
Physical	51	0.1366	Expression	134	0.6176
Protein	71	0.1902	Figure	46	0.2120
Research	62	0.1661	Gene	143	0.6591
Researchers	60	0.1608	Genes	228	1.0509
RNA	42	0.1125	Genetic	33	0.1521
Sequence	109	0.2920	Germ	48	0.2212
Sequences	50	0.1340	Granules	31	0.1429
Sequencing	44	0.1179	Homeotic	56	0.2581
Testing	42	0.1125	Interactions	32	0.1475

USA Books (cont.)			**UK Books (cont.)**		
Type	**N of oc.**	**Frequency**	**Type**	**N of oc.**	**Frequency**
Cell	90	0.2411	Lethal	48	0.2212
Disorder	61	0.1634	Map	22	0.1014
Fragments	53	0.1420	Maternal	44	0.2028
Test	51	0.1366	Maternal-effect	38	0.1751
Specific	46	0.1232	Maternally	28	0.1291
			Mutant	51	0.2351
			Mutants	64	0.2950
			Mutations	109	0.5024
			Oocyte	24	0.1106
			Pair-rule	26	0.1198
			Pattern	108	0.4978
			Patterns	35	0.1613
			Phenotype	29	0.1337
			Polarity	34	0.1567
			Pole	45	0.2074
			Product	50	0.2305
			Products	49	0.2258
			Protein	77	0.3549
			Region	36	0.1659
			Regulatory	23	0.1060
			Result	23	0.1060
			RNA	26	0.1198
			Segment	30	0.1383
			Stage	23	0.1060
			Stripes	25	0.1152
			Transcription	25	0.1152
			Ventral	28	0.1291
			Wild-type	44	0.2028
			Zygotic	34	0.1567

Table 8: Record of the technical words of the GENETICS sub-area from and from USA and UK books

A comparison between the 150 most frequent words of the sub-area Genetics in USA Journals, UK Journals, USA Books and UK Books reveals that, within the confines of the frequency band, the four files have only seven words in common: *Cell*, *Cells, Gene, Genes, Genetic, Protein* and *DNA.*

The highest occurrences and frequencies in the four files correspond to the shared words:

Gene: 770- 0.4758 (USA Journals)
Gene: 457- 0.7431 (UK Journals)
DNA: 375-1.0047 (USA Books)
Genes: 228-1.0509 (UK Books)

The lowest occurrences and frequencies correspond to the following words:

Insertion: 116-0.0717 (USA Journals)
Domain: 44-0.0715 (UK Journals)
Data: 34-0.0911 (USA Books)
Eggshell: 22-0.1429 (UK Books)

The words listed in the four columns of Tables 7 and 8 make a total of 196 with 76 different terms, UK Books being the file with the highest number of different or non-shared words (Table 9).

Source	**Number of non-shared words in each file**
USA Journals	4
UK Journals	16
USA Books	16
UK Books	37

Table 9: Source of the different files and number of non-shared words in each file

It is noteworthy that the files UK Journals and USA Books maintain the same number of non-shared words in each file despite the difference in length of both files: 630,000 words for the first one and 280,000 for the second. This fact could be understood on the basis that journals need more general usage terms to present experiments and report results, as previously stated.

2.5. Concordancing as a teaching tool

We have already reported the accuracy of the design of our corpus in order to obtain the widest range of vocabulary in use to be taught to students of Biology. Now we will compare the different senses and collocates of one term appearing in our Corpus with the form in which they appear in the texts of *English in Basic Medical Science (1996)* and *A Course in Intermediate Scientific English (1989)*. We have chosen one term that is emblematic in Biology: **body** with 1123 occurrences, respectively in the whole corpus since in order to study the behaviour of words in texts, we need to have available quite a large number of occurrences (Sinclair, 1991). The term, **body**, a semi-technical term, registers a wide variety of meanings as shown in Table 10.

If we compare these results with the forms in which both terms appear in the texts of English for Medical Science, English for Scientists, we find that the term **body** can be found only in the sense of "physical structure of a person or animal" (i.e. *The first compartment of the body consists of active tissue*). This as far as the meanings given in the textbooks.

On the other hand, not much more is offered in technical dictionaries if we search for the meaning of body. To illustrate our statement, we looked up the term **body** in *The Wordsworth Dictionary of Science and Technology* (1988) and we found two general entries, the first one followed by **(Build.)** and the second one by **(Print.).**

> **Body (Build.)**. (1) The degree of opacity possessed by a pigment. (2) The apparent viscosity of a paint or varnish. (3) The ability of a paint to give a good, uniform film over an irregular or porous surface.

Meaning	**Examples**
– The physical structure of a person or animal.	*If an animal is in caloric balance, its* **body** *weight, and thus volume, remains remarkably constant.*
– A set of something	*In contrast, the* **body** *of unitary organisms is a determinate structure consisting usually of a...*
– The main, central part of something	*In the* **body** *of the review, comparisons with analogous prokaryotic and higher eukaryotic..*
– Quantity	*...of species recognition, but there is a growing* **body** *of evidence for directional preferences based on sensory...*
– A part of a whole	*The spores are released as the fruit* **body** *deliquesces, turning the mushroom black.*
– Intracellular structure which is relevant in meiosis	*At about 0.6, the second polar* **body** *is extruded but does not separate from the...*
– Ribosome	*... von Willebrand disease, hemostasis, platelet adhesion, factor VIII, Weibel-Palade* **body**.
– Structures formed by some mysobacteria which contain spores under starvation conditions	*To further understand the molecular mechanisms involved in fruiting* **body** *formation,...*
– A large area of water (lake, reservoir)	*So by the 1950s, essentially every* **body** *of water receiving piped wastes was badly polluted with a...*

Table 10: Different meanings of the term "body" in our Corpus

> **Body (Print.).** (1) The measurement from top to bottom of a type, rule, etc. The unit is the **point**, 72 points amounting to (approx.)1 in. (2) The solid part of a piece of type below the printing surface or *face*. Also called *shank*, *stem*. (3) Body of a work, the text of a volume, distinguished from the preliminary matter, such as title and contents, and the end matter, such as appendices and index.

Consequently, none of them corresponds to Biology. However, after them, there are the following four definitions:

> **Body cavity (Zool.).** The perivisceral space, or cavity, in which the viscera lie; a vague term, sometimes used incorrectly to mean coelom.
>
> **Body cell (Bot.).** The cell that divides to give the two sperm cells in the gymnosperm pollen tube.
>
> **Body cell (Zool.).** Somatic cell.
>
> **Body wall (Zool.).** The wall of the perivisceral cavity, comprising the skin and muscle layers.

It can be observed that, on the one hand, the term **body** in itself has not an entry related to Biology and, on the other, when it is accompanied by cavity, cell or wall only the botanical or zoological meanings of the term are given. This observation reinforces the idea stated in these pages that considering Biology as formed by different sub-areas serves better even lexicographical purposes.

3. Conclusions

In the above discussion the general principles for the establishment of a Biology Corpus have been described. We have begun analysing the corpus and as part of preliminary results of the research we have focussed on its vocabulary application. We have generated some frequency lists of the different sub-areas and after extracting some technical and sub-technical terms we have shown that books have the highest lexical richness in the sub-areas explored.

As our corpus has been collected mainly with teaching purposes**,** and vocabulary has been shown to be one of the main ingredients in the learning of a language and the acquisition of a register, we consider that our corpus could be a reliable tool since it provides a wide range of scientific terms.

The positive results of the present investigation demonstrate the need to study register characteristics. The study of vocabulary has also served to reveal grammatical characteristics of Biology English. A deeper study of grammar should be carried out taking into account the different sources of our corpus.

Recebido em: 01/2002. Aceito em: 04/2002.

Acknowledgements: Special thanks are given to Dr. A. Sánchez, Professor of English Philology at the University of Murcia for his advice in the elaboration of this corpus.

References

BIBER, D. 1993 Representativiness in Corpus Design, *Literary and Linguistic Computing*, **8.4:** 243-257.

_____, S. CONRAD & R. REPPEN 1998 *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

CHAPLEN, F. 1989 *A Course in Intermediate Scientific English*. Thomas Nelson and Sons Limited.

DUDLEY-EVANS, T. & M.J. ST JOHN 1998 *Developments in English for Specific Purposes.* Cambridge University Press.

FLOWERDEW, J. 1993 Concordancing as a tool in course design, *System*, **21.2:** 231-244.

INMAN, M. 1978 *Foreign Languages, English as a second foreign language, and the U.S. multinational Corporation*. Arlington (Virginia): Center for Applied Linguistics.

HALLIDAY, M.A.K., P.D. STREVENS & A. MCINTOSH 1964 *The Linguistic Sciences and Language Teaching*. Longman.

JOHANSSON, S. 1975 *Some Aspects of the Vocabulary of Learned and Scientific English.* Goteborg: Acta Universitatis Gothoburgensis.

KENNEDY, C. & R. BOLITHO 1991 *English for Specific Purposes*. Macmillan Press Ltd.

LJUNG, M. 1991 Swedish teaching English as a foreign language meets reality. IN: S. JOHANSSON & A.B. STËNSTROM (eds) *English Computer Corpora: Selected Papers and Research Guide*. Mouton de Gruyter. 245-256.

MACLEAN, J. 1996 *English in Basic Medical Science*. Oxford University Press

MARTIN, A.V. 1976 Teaching Academic Vocabulary to Foreign Graduate Students. *TESOL Quarterly* **10:1**, March.

NATION, P. & J. NEWTON 1997 Teaching Vocabulary. IN: J. COADY & T. HUCKIN (eds) *Second Language Vocabulary Acquisition*. Cambridge University Press. 238-254.

SÁNCHEZ, A. (ed) 1995 *Cumbre, Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: Sgel.

_____ A. & P. CANTOS 1997 Predictability and Representativeness of Words, Word Forms and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMRE Corpus: An 8-Million Word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics*: **2.2:** 259-280.

SINCLAIR, J. 1991 *Corpus Concordance Collocation*. Oxford University Press.

ZIMMERMAN, C. 1997 Historial Trends in Second Language Vocabulary Instruction. IN: J. COADY & T. HUCKIN (eds) *Second Language Vocabulary Acquisition*. Cambridge University Press. 5-19.

Journal Citation Reports, 1998 Science edition published from the Institute for Scientific Information Data Base. This listing gives journals ranked by impact factor within subject categories.

LACELL Corpus 2000 Corpus compiled at the University of Murcia by the Research Group: Lingüística Aplicada Computacional, Enseñanza de lenguas y Lexicografía. Directed by A. Sánchez and P. Cantos. Murcia: University of Murcia.

The Wordsworth Dictionary of Science and Technology 1988 Ware: Wordsworth Editions Ltd.

Purificación Sánchez Hernández has been working as a translator for the CSIC (Spanish Research Council) for many years. She has a PhD in English Philology. Since 1997 she has been working at the English Department in the University of Murcia and teaches English Language and English Literature.